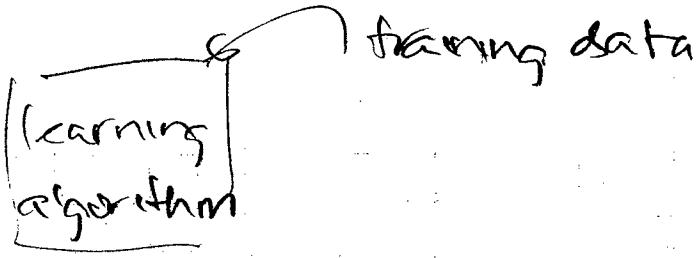
1.15.2009

Kinds of problems
- estimation
- classification (figuring out a class from which data comes)
  - e.g. classifying written digits
- regression (predicting a real number for data points)
  - e.g. predicting probability that a person gets in an accident based on age, income, # previous accidents

classification and regression are examples of supervised learning.



1. Train the algorithm using training data
2. Use the algorithm on real data.
   trained

raining	temp <32	wind >10mph	?	close windows
yes	no	no	no	yes
no	yes	no	yes	yes
yes	no	yes	no	no
		⋮		

It's supervised because training examples contain the answers (output classes).

We want the generalization (true) error to be small.
↑ error on the real data

Along the way, we will be minimizing the training error.

We want the training data and the real data to be related. We are going to assume that there is some distribution on the data and the training data and the real data are independent samples from the distribution

Classification

$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \cdots, (X_n, Y_n)$

The training data are independent (from some distribution, D)

After learning A using training data, what is the, e.g., expected error for (X, Y) from D?

The only difference between classification and regression is that Y is discrete for classification and continuous for regression.

Prof will not be providing rigorous proofs of error measures

Unsupervised Learning

Training data is $X_1, X_2, X_3, \cdots, X_n$

Goal is to cluster data into groups.

Basic Probability Theory (finite probability spaces first)

Ω is called the sample space.

For now, $|\Omega| < \infty$

The elements of Ω are called atomic events.
The subsets of Ω are called events.

$P: 2^{\Omega} \to \mathbb{R}_{\geq 0}$

(2^{Ω}: all subsets of Ω)

$P(\Omega) = 1$

If $A_1, \ldots, A_k$ are disjoint then $P(A_1 \cup \ldots \cup A_k) = \sum P(A_i)$

Of course, it is enough to know the probs. of the atomic events (i.e. $P(\{x\})$ for all $x \in \Omega$)

Then $P(A) = \sum_{x \in A} P(\{x\})$

Throw a die twice

$\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$

$P(x) = \frac{1}{36}$, e.g. $P((1,1)) = \frac{1}{36}$

Two events $A, B \subseteq \Omega$ are independent if

$P(A \cap B) = P(A)\,P(B)$

$A = \{(1, x),\ x \in \{1, \dots, 6\}\}$

$B = \{(x, 2),\ x \in \{1, \dots, 6\}\}$

$A \cap B = \{(1, 2)\}$

$P(A) = P(B) = \frac{1}{6}$, $P(A \cap B) = \frac{1}{36}$

$A = \{(x, y) : x + y = 7\}$

$B = \{(x, 2) ; x \in \{1, \dots, 6\}\}$

$P(A) = \frac{1}{6}$ $\quad A \cap B = \{(5, 2)\}$

$P(B) = \frac{1}{6}$ $\quad P(A \cap B) = \frac{1}{36}$

In the above case, A and B are independent
They wouldn't be if $A = \{\dots x + y = 3\}$

In the latter case, $P(A) = \frac{1}{18}$ and $P(A)P(B) \neq P(A \cap B)$

Conditional Probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{36}}{\frac{1}{18}} = \frac{1}{2}$$

Independent $\iff P(B|A) = P(B)$

Random Variables

A random variable is a function

$$X : \Omega \to \mathbb{R}$$

Throw two dice. If sum is 7, you pay \$7. Otherwise I pay \$1.
(e.g. $X((1,6)) = 7$, $X((1,2)) = -1$)

The expected value of a variable X is

$$E[X] = \sum_{w \in \Omega} P(w) X(w)$$

e.g. $7 \cdot \frac{1}{6} - 1 \cdot \frac{5}{6} = \frac{1}{3}$

Theorem (Weak Law of Large Numbers)

For any $\epsilon > 0$ and $X_1, \dots, X_n, \dots$ independent, identically distributed (i.i.d),

$$\lim_{n \to \infty} P\left(\left| \frac{X_1 + \dots + X_n}{n} - E[X] \right| > \epsilon \right) = 0$$

Entropy of a (discrete) random variable X is

$$\text{entropy}(X) = -\sum_{a} P(X=a) \cdot \log_2 P(X=a)$$

Random Variable $\rightarrow$ Events

X $\qquad$ $\{w \mid X(w) = a\}$

e.g. $\begin{cases} X=7, \{(1,6) \dots\} \\ X=-1, \Omega - \{(1,6), \dots\} \end{cases}$

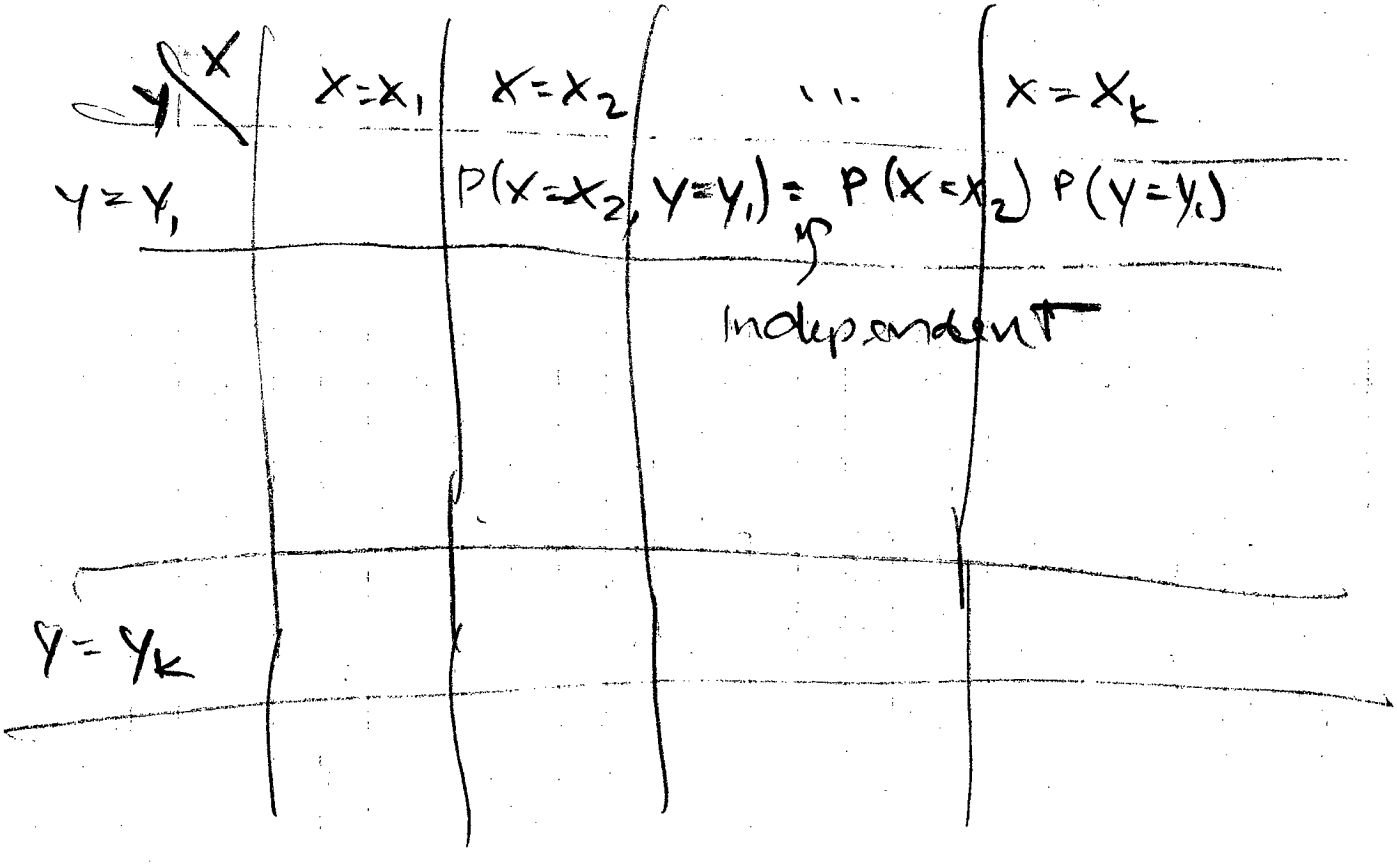
If $P(X=1) = 1/2$, $P(X=0) = 1/2$, then

entropy $(X) = 1$

For $P(Y=0) = P(Y=1) = P(Y=2) = P(Y=3) = 1/4$,
entropy $(Y) = 2$

If entropy $(X) = a$ and entropy $(Y) = b$,
X and Y independent

Two r.v. are independent if all events created from them are independent. For example, $(X \le c)$ and $(Y \le d)$ are independent for every c, d.



entropy $((X,Y)) =$

$$-\sum_{i,j} P(X = X_i \wedge Y = Y_j) \log_2 P(X = X_i \wedge Y = Y_j)$$

$p_i = P(X = x_i)$

$q_j = P(Y = y_j)$

$$-\sum_{i=1}^{k}\sum_{j=1}^{l} p_i q_j \log_2 p_i q_j$$

$$= -\sum_{i=1}^{k}\sum_{j=1}^{l} (p_i q_j \log_2 p_i + p_i q_j \log_2 q_j)$$

$$= -\sum_{i=1}^{k} p_i \log_2 p_i - \sum_{j=1}^{l} q_j \log_2 q_j = a + b$$

So, entropy$((X,Y))$ = entropy(X) + entropy(Y).

1.20.2009

Probability space

① set Ω (sample space)
$\mathcal{B}$ a collection of subsets of Ω

② function $P: \mathcal{B} \to [0,1]$ — $\mathcal{B} = 2^{\Omega}$ we took all subsets when $|\Omega| < \infty$

③ $P(\Omega) = 1$

$A_1, \dots, A_n, \dots$ disjoint sets from Ω

$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ ← before we did this for finite collections

$\mathcal{B}$ is a nice collection of subsets

if $\Omega = \mathbb{R}$

then
- $\mathcal{B}$ contains all intervals
- is closed under complement

$A \in \mathcal{B} \Rightarrow \mathbb{R} \setminus A \in \mathcal{B}$

- is closed under countable unions

$A_1, \dots, A_n, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$

σ-algebra
σ-field

a collection of sets that satisfy the 2 properties

DEF
A <u>random variable</u> is a measurable function $X: \Omega \to \mathbb{R}$

DEF
The <u>distribution function</u> of X is

$f: \mathbb{R} \to [0,1] \quad f(a) = P(X \leq a)$

Example:

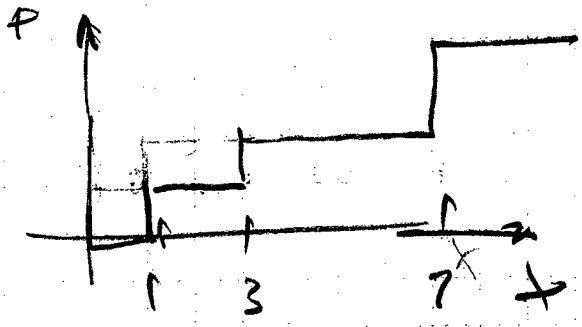
$\Omega = \{1, 2, 3\}$

$P(\{1\}) = 1/4$
$P(\{2\}) = 1/4$
$P(\{3\}) = 1/2$

$X(1) = 1$
$X(2) = 3$
$X(3) = 7$

What is the distribution function of X?

$P(X \leq 0) = 0$
$P(X \leq 1) = 1/4$
$P(X \leq 0.999) = 1/4 = 0$
$P(X \leq 5) = 1/2$

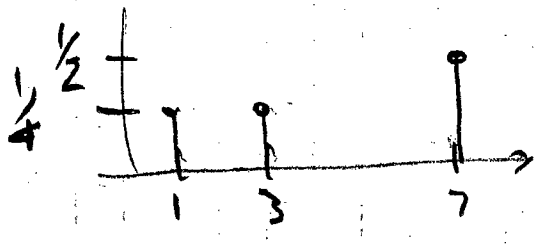


This is also called a cumulative density function if X is continuous

We will mainly encounter r.v. that are

① Discrete r.v. have countably many possible values

Probability mass function of X
$f(a) = P(X=a)$



② Continuous r.v. (e.g. random real number from $[0,1]$) with distr. function F

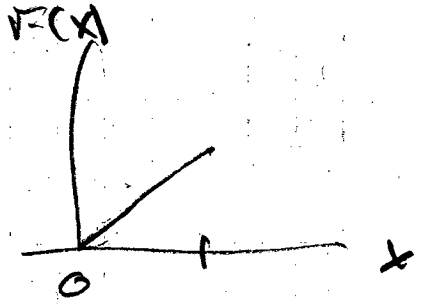
A random variable is continuous if there exists $f: \mathbb{R} \to \mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

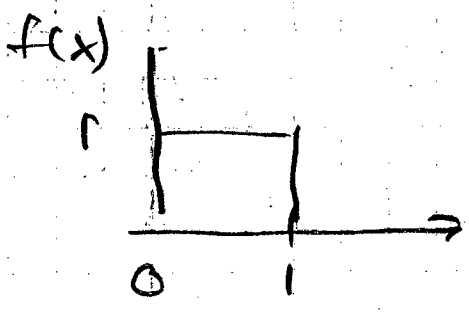
($-\infty$ to x; f = density of x)

$$= P(X \le x)$$

E.g. uniform
$P(X \le 1/2) = 1/2$
$P(X \le 1) = 1$

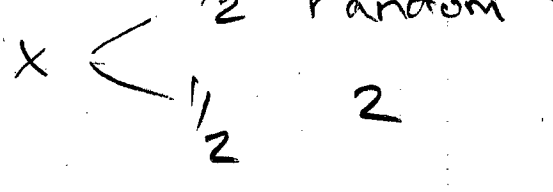


The density is

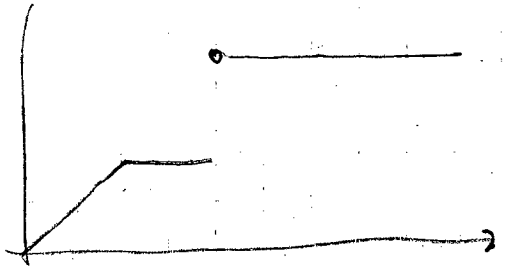


For a uniform r.v. $\in [0,2]$, the density would be $1/2$.

This is neither discrete nor continuous



because it's both



Basic Discrete Distributions

- Bernoulli Distribution (p)

$$P(X=1) = p$$
$$P(X=0) = 1-p$$

$$E[X] = p$$

- Binomial Distribution (n,p)

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$k \in \{0, \ldots, n\}$

Bernoulli is a special case of binomial.

If $X_1, \ldots, X_n$ are independent from Bernoulli (p), then $X = X_1 + \ldots + X_n$ is from Binomial (n,p)

$E[X+Y] = E[X] + E[Y]$

linearity of expectation

Observe Bernoulli. How many coin flips are needed to get 1 head?

- Geometric Distribution

$$P(X=k) = p(1-p)^{k-1}$$

Let $X_1 \dots X_n \dots$ be independent r.v. from Bernoulli (p).

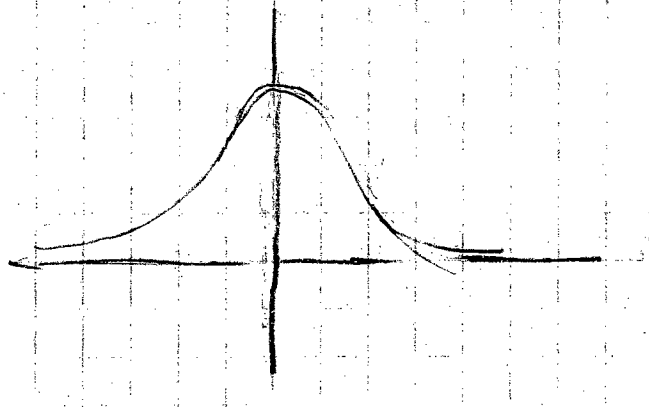
$$X = \min\{t \mid X_t = 1\}$$

$$E[X] = \frac{1}{p}$$

- Gaussian (Normal) Distribution (μ, σ)

density function =

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



Inequalities for R.Vs

- Markov's Inequality

Let X be a non-negative r.v. (i.e. $X \geq 0$)

$\forall t \quad P(X \geq t E[X]) \leq \frac{1}{t}$ $\quad (t > 0)$

Proof:

If $P(X \geq t E[X]) > \frac{1}{t}$, then:

$$E[X] = \int_0^\infty x f(x)\,dx \geq \int_{tE[X]}^\infty x f(x)\,dx$$

$$\geq \int_{tE[X]}^\infty f(x)\, tE[X]\,dx = tE[X] \int_{tE[X]}^\infty f(x)\,dx > E[X]$$

$$= P(X \geq tE[X]) > \frac{1}{t}$$

So, we have $E[X] > E[X]$, which is a contradiction.

DEF

Let X be a random variable. The variance is

$$V[X] = E\left[(X - E[X])^2\right]$$

For Bernoulli:

	X	$(X-E(X))^2$
p	1	$(1-p)^2$
$1-p$	0	p^2

$$V[X] = p(1-p)^2 + (1-p)p^2$$
$$= p(1-p)$$

If $X_1, \ldots, X_n$ are <u>independent</u>, then

$$V[X_1 + \ldots + X_n] = \sum_{i=1}^{n} V[X_i]$$

$$V[\alpha X] = \alpha^2 V[X]$$

- Chebyshev's Inequality

X, $E[X] < \infty$, $V(X) < \infty$

$$P(|X - E[X]| \geq t) \leq \frac{V(X)}{t^2}$$

Let $Y = (X - E[X])^2$ $\quad E(Y) = V[X]$

$$P(Y \geq t E[Y]) \leq \frac{1}{t}$$

$$P((X - E[X])^2 \geq t V[X]) \leq \frac{1}{t}$$

Let $t = \frac{t'^2}{V(X)}$

$$P((X - E[X])^2 \geq t'^2) \leq \frac{V(X)}{t'^2}$$

<u>Weak Law of Large Numbers</u>

$X_1, \ldots, X_n$ are independent, identically distributed ($V(X_i) < \infty$, $E(X_i) = \mu$)

$\forall \varepsilon > 0$

$$\lim_{n \to \infty} P\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0$$

Proof: Apply Chebyshev to random variable which is sum here. Variance goes to 0.

$$V\left(\frac{X_1+\cdots+X_n}{n}\right) = V(X_1)/n$$

Chernoff-Hoeffding

$X_1, \ldots, X_n$ are independent from Bernoulli(p)

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

Chebyshev would give us

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{\varepsilon^2 n}$$

Chernoff-Hoeffding gives us a stronger result.

We use these things for estimation. An example would be polling where Bernoulli samples are votes.

Given $X_1, \ldots, X_n$ from Bernoulli, the goal is to estimate the unknown p.

$$\text{Estimate} = \frac{X_1+\cdots+X_n}{n}$$

How good is this estimate?

Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. r.v. with mean μ and variance σ^2.

$$\frac{\sqrt{n}}{\sigma}\left(\frac{X_1+\cdots+X_n}{n} - \mu\right) = Z_n \quad \text{(let)}$$

$$\lim_{n\to\infty} \sup_a \left| P(Z_n \leq a) - \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right| = 0$$

1.22.2009

Parameter Estimation

$X_1, \dots, X_n$ i.i.d from $D(\alpha)$; α is unknown

The goal is to determine α.

$D(\alpha)$: distribution

ε precision
δ confidence

$f(X_1, \dots, X_n) \to G$ — r.v

We want $P(|G - \alpha| > \varepsilon) \leq \delta$

Maximum Likelihood

$P(X_1, \dots, X_n \mid \text{assuming the parameter is } \alpha)$ ← likelihood

Pick α for which expression is largest.

Example

Assume that the family of distributions considered is Bernoulli (α) $\alpha \in [0,1]$

$n = 6$ 0, 0, 1, 1, 0, 1

$P(X_1 = 0, X_2 = 0, \dots, X_5 = 1 \mid \alpha) = (1-\alpha)^3 \alpha^3$

The expression above is maximized for $\alpha = 1/2$.

Assume the family $N(\mu, \sigma)$ $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

Observed $X_1, \dots, X_n$ we want to pick μ, σ that maximize the likelihood.

The likelihood is

$$f(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$

Maximizing log is same as maximizing f.

$$\log f(\mu,\sigma) = \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} - n\ln\sigma - n\ln\sqrt{2\pi}$$

$$\frac{\partial}{\partial\mu}\log f(\mu,\sigma) = \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\sigma}\log f(\mu,\sigma) = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma}$$

$$\sigma_{ML}^2 = \frac{1}{n}\Sigma (x_i-\mu_{ML})^2$$

Assume that data actually come from $N(\mu,\sigma)$

What will μ_{ML}, σ_{ML} give us?

$$E[\mu_{ML} \mid x_1,\ldots,x_n \sim N(\mu,\sigma)] = \mu$$

$$E[\sigma_{ML}^2 \mid x_1,\ldots,x_n \sim N(\mu,\sigma)] =$$

(without loss of generality, we can assume $\mu = 0$)

$$E[\sigma_{ML}^2 \mid \cdots] = \frac{1}{n} E\left[\sum_{i=1}^{n} \left(x_i - \frac{\Sigma x_i}{n}\right)^2 \mid \cdots\right]$$

$$\frac{1}{n} E\left[\sum_i x_i^2 - 2\sum_i x_i \left(\frac{\sum_j x_j}{n}\right) + \frac{1}{n^2}\sum_i \left(\sum_j x_j\right)^2 \mid \cdots\right] =$$

$$V[x] = E[x^2] - E[x]^2 \qquad E[x_i x_j] = \rho \text{ for } i \neq j$$

$$\frac{1}{n}\left(n\sigma^2 - 2\sigma^2 + \sigma^2\right) = \frac{n-1}{n}\sigma^2$$

Since $\sigma_{ML} \neq \sigma$, we say that σ_{ML} is a <u>biased</u> estimator.

But it is assymptotically unbiased since

$$\lim_{n\to\infty} \sigma_{ML} = \sigma$$

<u>Bayes Theorem</u>

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

We will use Bayes theorem to calculate the parameters of a distribution.

A = distribution parameters

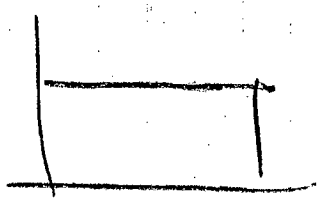
B = data

A prior is a belief on how likely p is.
We observe some data
We update the belief based on the data.
The updated belief is called the posterior.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$P(A|B)$: posterior
$P(B|A)$: likelihood
$P(A)$: prior
$P(B)$: normalizing factor

Example
Assume Bernoulli distribution.



Observe ϕ

posterior?

$$P(\text{parameter is } x \mid \text{observed } \phi) = \frac{(1-x)(1)}{?} = \frac{1-x}{1/2}$$

pick so that we get a probability distr.

$$P(B) = \int P(B|A)\cdot P(A)\,dA$$

$$P(\text{observing } 0 \mid \text{parameter is } x) = 1-x$$

$$P(\text{observing } 0) = \int (1-x)\cdot 1\,dx$$

Typically, we pick priors for mathematical convenience. We want to insure that posterior distribution is from same family as prior.

DEF

Let F be a family of likelihoods.

A family of distributions G is <u>conjugate prior</u> to F when

1. prior $\in F$, likelihood $\in G \Rightarrow$ posterior $\in F$

What would be a nice family of priors for Bernoulli?

$P(X_1 = a_1, \cdots, X_n = a_n \mid \text{parameter is } p)$

$a_1, \cdots, a_n \in \{0, 1\}$

$$P(\cdots) = (1-p)^{n-\Sigma a_i} p^{\Sigma a_i}$$

Prior from a distribution

$$P(z) \propto (1-z)^A z^B$$

Posterior is from

$$P(z \mid X_1 = a_1, \cdots, X_n = a_n) \propto (1-z)^A z^B (1-z)^{n-\Sigma a_i} z^{\Sigma a_i}$$

$$= (1-z)^{A'} z^{B'}$$

$A' = A + n - \Sigma a_i$

$B' = B + \Sigma a_i$

Beta (A, B) distribution is a distribution on $[0,1]$

$P(z) \propto z^{A-1}(1-z)^{B-1}$ continuous

The Beta distribution is the conjugate prior for the Bernoulli, discrete distribution.

Lets look at Gaussian distributions. Lets first look at μ, assuming σ is fixed.

$P(x_1, \ldots, x_n | \mu) =$

We are looking for priors over μ. The likelihood function is

$$P(x_1, \ldots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$

We need something that has a quadratic function as an exponent of e.

The prior on μ is $N(\mu_0, \sigma_0)$

1.27.2009

Bayesian "setting"

prior (distribution on the parameters)
↓ likelihood (from data)
posterior (different distribution on the data)

eg. $X_1, \ldots, X_n$ are from Bernoulli(p) (parameter)

We want prior * likelihood = posterior such that posterior is from same distribution as prior.

prior $p \sim \text{Beta}(a, b)$

$P(p = x) \propto x^{a-1}(1-x)^{b-1}$

observed $d_1, \ldots, d_n \in \{0, 1\}$

$P(0 \mid p = x) = 1 - x$

$P(1 \mid p = x) = x$

$P(d_1 \mid p = x) = x^{d_1}(1-x)^{1-d_1}$

$\text{likelihood} = x^{\sum d_i}(1-x)^{n - \sum d_i}$

posterior $\sim \text{Beta}(a + \sum d_i, b + n - \sum d_i)$

prior

↓ $data_1$

$posterior_1$ - use as prior for next batch of data.

↓ $data_2$

$posterior_2$

Is this the same

prior ⟶ posterior
data1 + data2

Is this the same as getting all of the data at once? Yes

This is called sequential approach to Bayesian.

Beta(a,b)

↓ $d_1, \ldots, d_n$

Beta$(a+\Sigma d_i, b+n-\Sigma d_i)$

↓ $d_1', \ldots, d_m'$

Beta(...)

Example 2

If our model of the world is Gaussian

$$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$$

We want to estimate $\lambda = \frac{1}{\sigma^2}$

In this case the conjugate prior is the gamma distribution

$$\text{prior} \sim \text{gamma}(a, b) \propto \lambda^{a-1} e^{-b\lambda}$$

---

prior $\sim$ Gamma(a, b)

↓ data $x_1, \ldots, x_n$

posterior $\sim$ Gamma(a', b')

$$a' = a + \frac{n}{2}$$

$$b' = b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In this case, processing data all at once yields different results than sequential. Consider adding one point at a time. b would not change in sequential case.

Parametric Framework

We always assume data is independent

MLE:

We have some parametric model (Θ = parameter)

likelihood (data | Θ) $= P(X_1 = d_1, \ldots, X_n = d_n \mid \Theta) = \prod P(X_i = d_i \mid \Theta)$

Maximum likelihood method says to pick Θ which maximizes likelihood.

The KL divergence of f, g is

$$D(f,g) = \int f \log \frac{f}{g}\, dx$$

- $D(f,g) \geq 0$
- $D(f,f) = 0$

Since log is monotonic, maximizing likelihood is same as maximizing log likelihood

Maximizing log likelihood

$$\frac{1}{n} \sum_{i=1}^{n} \log P(X_i = d_i \mid \Theta)$$

↑

what is this converging to?

$$E[\log P(X_1 = d_1 \mid \Theta] =$$

$$\int_{d_1 \sim \Theta^*} \quad (\text{"}d_1 \text{ comes from some distribution } \Theta^*\text{"})$$

$$= \int P(X_1 = d_1 \mid \Theta^*) \log P(X_1 = d_1 \mid \Theta)\, dx_1$$

not in lecture

$$= -KL(f_{\Theta^*} \mid f_\Theta) + \int P(X_1 = d_1 \mid \Theta^*) \log (X_1 = d_1 \mid \Theta^*)\, dx_1$$

We are trying to minimize the KL-divergence.

Regression

$(X_1, Y_1), \dots, (X_n, Y_n)$

$X_i \in \mathbb{R}^d$

$Y_i \in \mathbb{R}$

The goal is to fit a linear function to the data to predict Y_i from X_i.

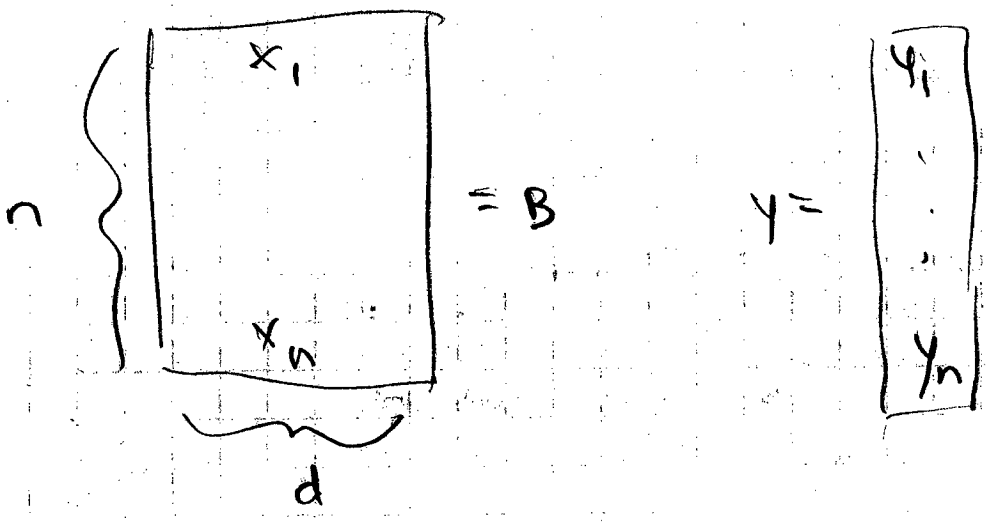
LSE (Least Squares Estimator)

$$\underline{X}_1 = (X_{11}, \dots, X_{1d})$$

$$W_1 X_{11} + \dots + W_d X_{1d} = W^T \underline{X}_1$$

For LSE we want to pick W to minimize

$$\sum_{i=1}^{n} (W^T \underline{X}_i - Y_i)^2$$



B W

$$BW = \begin{pmatrix} w_T x_1 \\ \vdots \\ w_T x_n \end{pmatrix}$$

Goal is to minimize $\|BW - y\|_2^2$

Given $B \in \mathbb{R}^{n \times d}$

$y \in \mathbb{R}^n$

Find w minimizing $\|BW - y\|_2^2$.

Note that $\|u\|_2^2 = u^T u$

Minimize:

$$y^T y + W^T B^T B W - 2 y^T B W$$

$\frac{\partial}{\partial w} a^T w = a^T$

$a_1 w_1 + a_2 w_2 + \cdots + a_d w_d$

For $w = (w_1, \ldots, w_d)$

$\frac{\partial}{\partial w} f = \left(\frac{\partial}{\partial w_1} f, \ldots, \frac{\partial}{\partial w_d} f \right)$

$$\frac{\partial}{\partial w} w^T G w = \frac{\partial}{\partial w} \left(\sum_{i=1}^{d} \sum_{j=1}^{d} w_i w_j G_{ij} \right) = w^T (G + G^T)$$

$G \in \mathbb{R}^{d \times d}$

$$\frac{\partial}{\partial w_k} (\quad) = \sum_{\substack{j=1, \\ j \neq k}}^{d} w_j G_{kj} + \sum_{\substack{i=1, \\ i \neq k}}^{d} w_i G_{ik} + 2 w_k G_{kk}$$

So, to minimize $W^T B^T B W - 2y^T B W - (\dots)$ (symmetric)

$$\frac{d}{dW}(\dots) = 2W^T B^T B - 2y^T B = \phi$$

$$W^T (B^T B) = y^T B$$

If $B^T B$ is invertible,

$$W^T = y^T B (B^T B)^{-1}$$

If $B^T B$ is not invertible, then ...

$A + \varepsilon I$ is singular iff $\det |A + \varepsilon I| = \phi$

polynomial in ε of degree $\leq d$

Let $B^\dagger = \lim_{\varepsilon \to 0} B (B^T B + \varepsilon I)^{-1}$

This is called the Moore Penrose pseudoinverse.

$W^T = y^T B^\dagger$ this is a "dagger", not a "plus"

Let A be an $m \times n$ matrix. The SVD of A is

$$A = U D V^T$$

where $U \in \mathbb{R}^{m \times m}$, $D \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$

U, V are orthogonal, that is

$$UU^T = I, \quad VV^T = I,$$

and D is diagonal.

How to compute A^+?

1. $A = UDV^T$
2. Let $A^+ = UD^+V^T$

For every $\neq 0$ entry take $1/$that

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}$$

Assume that $B = UDV^T$. Then $B^T = VD^TU^T$

$UDV^T(UDV^TVD^TU^T + \varepsilon I)^{-1} =$

$UDV^T(UDD^TU^T + \varepsilon UU^T)^{-1} =$

~~This someone said it should have been~~

$$W^T = y^T B(B^TB)^{-1}$$

$$B^+ = \lim_{\varepsilon \to 0} B(B^TB + \varepsilon I)^{-1}$$

$UDV^T(VD^TU^TUDV^T + \varepsilon I)^{-1} =$

$UDV^TV(D^TD + \varepsilon I)^{-1}V^T =$

$UD(D^TD + \varepsilon I)^{-1}V^T$

Identities

$(AB)^T = B^TA^T$

$(AB)^{-1} = B^{-1}A^{-1}$

Motivation

Assume the model is $y = W^Tx + \varepsilon$

$W \in R^d$, $\varepsilon \in N(0, \sigma)$.

Then MLE = least squares.

1.29.2009

Classification

regression → continuous output
classification → discrete "

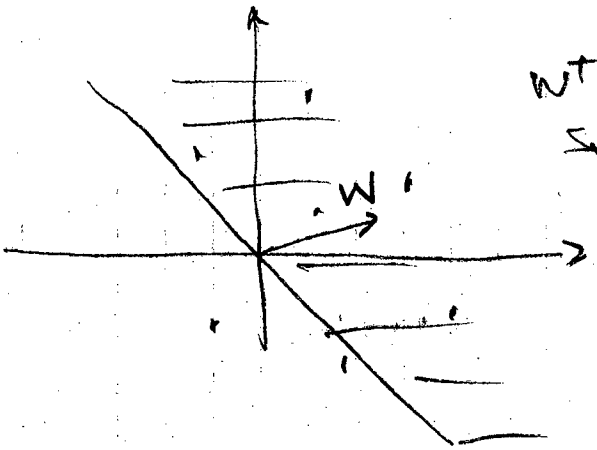
Simplest classifier is the:

Perceptron

Perceptron is a linear classifier.

$x = \begin{bmatrix} \\ \\ \end{bmatrix}$ $\quad t \in \{+,-\}$ or $t \in \{1,-1\}$

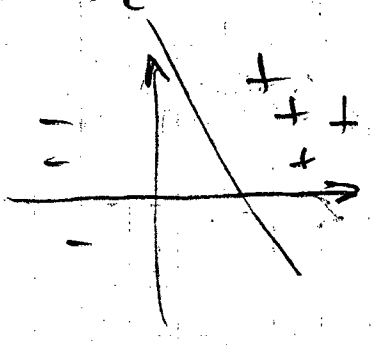
We get binary classification from $\text{sign}(w^T x)$



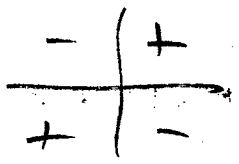
$w^T x$ is + on this side of boundary

For a boundary not passing through origin append a 1 to x, add new w_0

$x = \begin{bmatrix} \\ \\ 1 \end{bmatrix}$ $\quad w = \begin{bmatrix} \\ \\ b \end{bmatrix}$

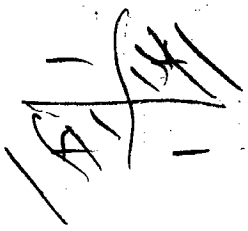


Perceptron doesn't work for

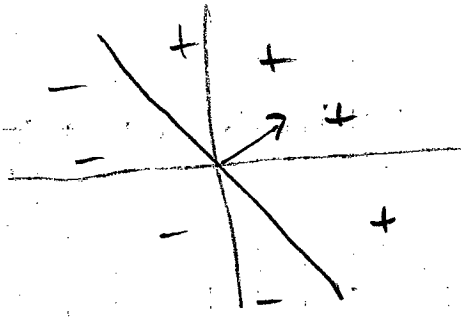


Perceptron only works if data is linearly separable.

We need to transform data



How do we pick w?



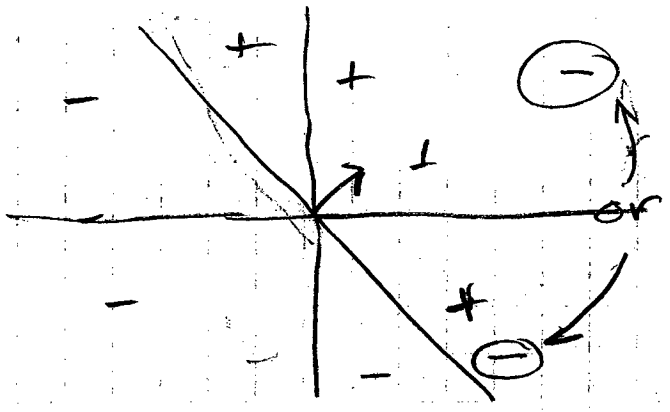
while some prints misclassified
  for i
    if $\text{sign}(w^T x_i) \neq t_i$
      $w \mathrel{+}= \eta t_i x_i$

(eta is the learning rate

The above is called the perceptron update rule. It will converge if data is linearly separable.

What if your data is not linearly separable?

How much do we want to penalize a weight vector for a difficult training example?



We will have an error function that tells us how bad a point is.

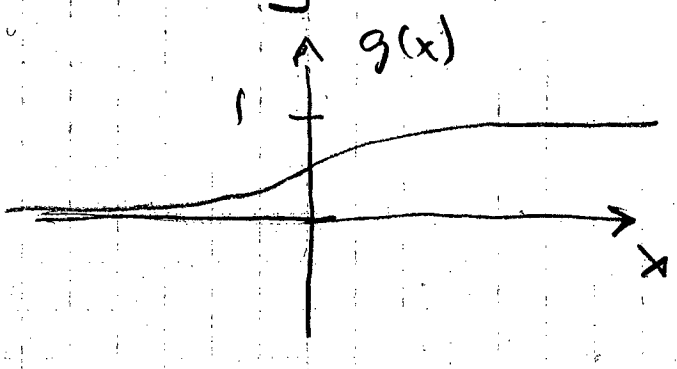
$$E(y,t) = \frac{1}{2}(t-y)^2 \quad \text{for } y = w^T x$$

This is not a good fit for classification because we don't want to respond too much for really bad data points.

Define an activation function g

$$y = g(w^T x)$$

Try $g(x) = \dfrac{e^x}{e^{-x} + e^x}$



Lets use $t \in \{0,1\}$

We will use gradient descent to minimize the error

while not satisfied
for i 1 to N

$$w += \eta \frac{\partial E(y_i, t_i)}{\partial w}$$

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial y} \frac{\partial y}{\partial w}$$

$$= \frac{\partial E}{\partial y} g'(w^T x) x$$

This ends up being a little more sophisticated than previous case.

"while not satisfied" $\Rightarrow$ while $|w - w_{old}| > \varepsilon$

Maximum Entropy = Logistic Regression

This is good for problems with > 2 labels.

You are learning the distribution

$$P(y|x)$$

and you want it to have a high entropy. You want to be "uncertain about what you can't see from the data".

$f_i(x, y)$

$f = [\ ]$

If x any y are both discrete, for example,

$$f_i = \begin{cases} 1 & \text{if } x = \text{tion}, y = N \\ 0 & \text{otherwise} \end{cases}$$

$$\max_p H_p(y|x) \quad \text{s.t.} \quad E_{\tilde{p}}[f_i] = E_p[f_i] \quad \forall i$$

$$\forall_x \sum_y p(y|x) = 1$$

$\tilde{p}$ = empirical distribution = sample distribution

$$\tilde{p}(x,y) = \frac{n(x,y)}{N}$$

All of our features should have the same expectation in the real distribution as in the empirical distribution.

How do we solve this?

$$L = \left[\sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} + \sum_i \lambda_i \left(p(x,y) f_i(x,y) - \tilde{p}(x,y) f_i(x,y) \right) \right] + \sum_x \mu_x \left(\sum_y p(y|x) - 1 \right)$$

Lagrangian

Whenever you have a problem in the form of

max $f(x)$ subject to the constraint $g(x) = 0$,

you can solve by

$$0 = f'(x) + \lambda g'(x)$$
$$0 = g(x)$$

$$H(y|x) = \sum_{x,y} p(x)\, p(y|x) \log \frac{1}{p(y|x)}$$

We assume $p(x,y) = \tilde{p}(x)\, p(y|x)$

$$\frac{\partial L}{\partial p(y|x)} = \left(\log \frac{1}{p(y|x)} - 1\right) \tilde{p}(x) + \sum_i \lambda_i \tilde{p}(x) f_i(x,y) + \mu_x = 0$$

$$\log p(y|x) = \sum_i \lambda_i \tilde{p}(x) f_i(x,y) + c$$

$$p(y|x) = \frac{1}{Z(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

$Z(x)$ is the normalization factor.

$$Z(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$

$$\sum_{x,y} \tilde{p}(x)\, p(y|x) \left[-\sum_i \lambda_i f_i(x,y) + \log Z(x) \right] + \sum_i \lambda_i \tilde{p}(x)\, p(y|x) \cdot f_i(x,y)$$

$$\sum_{x,y} \tilde{p}(x)\, p(y|x) \log Z(x)$$

$$\max_p E_p \log Z(x)$$

This is a simplification of the original equations.

There is no closed form for λ.

This is another way to penalize pts on wrong side of decision surface

2.3.2009

Tutorial W6-7

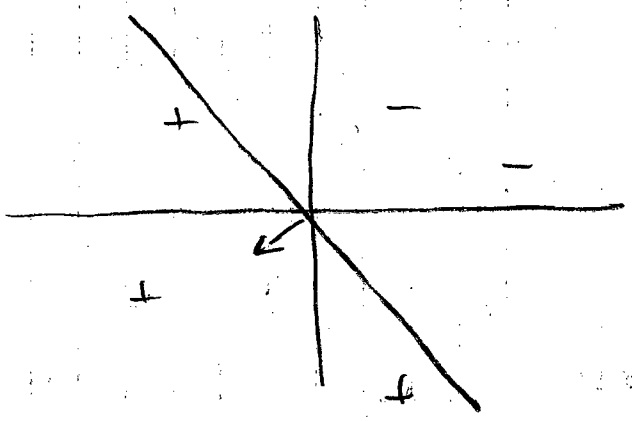
Perceptron Algorithm

Input is x_1, y_1 ; x_n, y_n

$x_i \in \mathbb{R}^d$

$y_i \in \{\pm 1\}$

Goal is to learn classifier $\text{sign}(w^T x)$

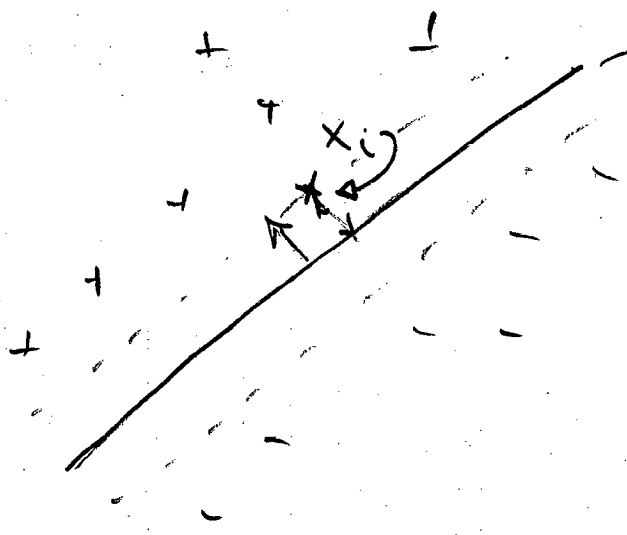


$w \leftarrow (0, \ldots, 0)$

for i from 1 to n
  if $\text{sign}(w^T x) \neq y_i$ then
    $w \leftarrow w + y_i x_i / \|x_i\|_2$

Assume that the negative examples can be separated from the positive examples by a hyperplane (w).

Assume that no point can get "too close" to the hyperplane.



distance from hyperplane $= \frac{w^T x_i}{\|w\|_2}$

We assume the quantity above is "large" for every x_i.

Even stronger assumptions ($\exists w^*$) a "good w"

- separable
- large margin $\sigma = \min_i \frac{x_i^T w^*}{\|w^*\|_2}$ (The closest pt to hyperplane is "far" from it)

One more assumption: $\|x_i\|_2 \leq R$

The perceptron algorithm makes, at most, $O(R^2/\sigma^2)$ updates to w and finds the separating hyperplane w.

In the algorithm, lets assume that $\|x_i\|_2 = 1$. Then, nothing changes.

Lets also assume, WLOG, that $\|w^*\|_2 = 1$

all x_i are normalized to length 1.

What happens after update?

$w' \leftarrow w + y_i x_i$

$w'^T w^* = w^T w^* + y_i x_i w^*$

$\geq w^T w^* + \delta/R$

This R is here because we normalized x_i.

This indicates that w' is getting closer to w^*.

$\|w'\|_2 = w^T w + 2 y_i x_i w + x_i^T x_i$

≤ 0

Because of normalization $= 1$

$\therefore \|w'\|_2^2 \leq \|w\|_2^2 + 1$

After φ updates,

$w^T w^* \geq \varphi \cdot \frac{\delta}{R}$

$w^T w^* \leq \|w\| \|w^*\| \leq \sqrt{\varphi}$

So, $\varphi \leq \frac{R^2}{\delta^2}$

$\|w\|_2^2 \leq \varphi$ because increasing by 1 each update

$\|w^*\|_2^2 = 1$ by assumption

Now, we are going to make perceptron algorithm work in a higher dimensional space.

Project $x \longmapsto f(x) \in \mathbb{R}^{d'}$

Find the **separator** in $\mathbb{R}^{d'}$ (instead of $\mathbb{R}^d$)

We assume that d' is big. Our goal is to run the perceptron algorithm in $\mathbb{R}^{d'}$ without computing f.

eg. $(x_1, x_2) \to (x_1, x_2, x_1^2, x_1x_2, x_2^2)$

For the perceptron algorithm

$$w = \sum_{i=1}^{n} \alpha_i x_i$$

on input x, y, compute

$$\text{sign}(w^T x) \stackrel{?}{=} y$$

$$\text{sign}\left(\sum_{i=1}^{n} \alpha_i (x^T x)\right) \stackrel{?}{=} y$$

We only need to look at inner products $(x_i^T x)$ of data points to update α_i.

In the projected space, these become

$$f(x_i)^T f(x)$$

But we want to avoid calculating f. So, instead, we have

$$K(x_i, x_j) = f(x_i)^T f(x_j)$$

DEF $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a <u>kernel</u> if

$\exists f: \mathbb{R}^d \to \mathbb{R}^\infty$ such that

$$K(x_i, x_j) = f(x_i)^T f(x_j)$$

Examples of Kernels

(We drop i, j subscripts)

$K(x, z) = x^T z$ $\quad f$ is identity

$K(x, z) = (x^T z)^2$

In 2 dimensions

$$K(x,z) = (x^T z)^2 = (x_1 z_1 + x_2 z_2)^2$$
$$= x_1^2 z_1^2 + 2 x_1 x_2 z_1 z_2 + x_2^2 z_2^2$$

So, $f(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$

$$K(x,z) = e^{-\|x - z\|_2^2}$$

↑ in this case, f would go to ∞ dimensional space.

Which functions are kernels?

Mercer's theorm states that K is a kernel iff

$$(\forall n)(\forall x_1, \ldots, x_n \in \mathbb{R}^d)$$

the matrix $A = (K(x_i, x_j))$ is positive semi-definite.

A is positive semi-definite

$$A \geq 0 \iff (\forall x)\ x^T A x \geq 0$$

$$(1, -1) \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = -4$$

So ↗ is not positive semi-definite

$$\left.\begin{array}{l} A \succeq 0 \\ B \succeq 0 \end{array}\right\} \Rightarrow A + B \succeq 0$$

Corollary: If K_1, K_2 are kernels, then their sum $K_1 + K_2$ is a kernel.

$$\left.\begin{array}{l} A \succeq 0 \\ B \succeq 0 \end{array}\right\} \Rightarrow \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{n1} & \cdots & a_{nn}b_{nn} \end{pmatrix} \succeq 0$$

$K(x,y) = (x^T y)^2$ is a kernel

Note that if $A \succeq 0$, then $A^T A \succeq 0$.

$x^T A^T A x = \|Ax\|_2^2 \geq 0$

If $A \succeq 0$, then

$$\begin{pmatrix} e^{a_{11}} & \cdots & e^{a_{1n}} \\ \vdots & \ddots & \vdots \\ e^{a_{n1}} & \cdots & e^{a_{nn}} \end{pmatrix} \quad ?$$

Note that $e^x = 1 + x + \frac{x^2}{2!} + \cdots$

$$x^T \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} x = x^T 1 1^T x \geq 0$$

2.4.2009

Tutorial

Andrew Ng web page

From (2.14)

$$F[y(x)+\varepsilon\eta(x)] = F[y(x)] + \varepsilon\int \frac{\delta F}{\delta y(x)}\eta(x)\,dx + O(\varepsilon^2)$$

$$\int [p(x)+\varepsilon\eta(x)]\ln[p(x)+\varepsilon\eta(x)]\,dx = \int p(x)\ln p(x)\,dx + \ldots$$

$$\ldots + \varepsilon\int \frac{\delta}{\delta p(x)} p(x)\ln p(x)\,dx$$

We are not taking derivative w.r.t. $p(x)$. We are interested in the derivative above being $\emptyset$ everywhere

2.5.2009

Goal: Get a good classifier which is a hyperplane (maybe in a higher dimensional space)

If the data is separable with a large margin, the perceptron algorithm "works".

For today, the goal is to find the hyperplane with the largest possible margin.

$x^{(i)}, y^{(i)} \quad i \in \{1, \dots, n\}$

$y^{(i)} \in \{\pm 1\}$

$x^{(i)} \in \mathbb{R}^d$

$$(w^T x^{(i)} + b) y^{(i)} \geq \gamma \quad , \max \gamma, \|w\|_2^2 = 1$$

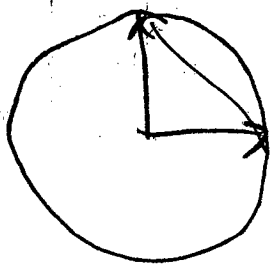
The variables are w, b, γ

We want to make this a convex optimization problem.

Let $b' = \frac{b}{\gamma}$, $w' = \frac{w}{\gamma}$, $\|w'\|_2 = \frac{1}{\gamma}$

$\min \|w'\|_2$, $(w'^T x^{(i)} + b') y^{(i)} \geq 1$

Constraint is not convex, because you can draw a line between 2 w's that satisfy $\|w\|_2 = 1$ but some points on the line have length < 1



The variables are now w', b'.

Assume w_1, b_1 and w_2, b_2 satisfy constraints.

Does $\frac{w_1+w_2}{2}, \frac{b_1+b_2}{2}$ satisfy the constraint?

Answer is <u>yes</u>.

$$\left(\frac{w_1^T+w_2^T}{2}x^{(i)} + \frac{b_1+b_2}{2}\right) y^{(i)} \geq 1 \; ?$$

$$\frac{1}{2}\underbrace{(w_1^T x^{(i)} + b_1) y^{(i)}}_{\geq 1} + \frac{1}{2}\underbrace{(w_2^T x^{(i)} + b_2) y^{(i)}}_{\geq 1} \geq 1$$

- Convex feasible set
- Objective is a convex function (minimization)

If you have these two, optimization is always possible

$$\left\| \frac{w_1+w_2}{2} \right\|_2^2 \leq \frac{\|w_1\|_2^2 + \|w_2\|_2^2}{2}$$

$$(w_1+w_2)^T(w_1+w_2) \leq 2(w_1^T w_1 + w_2^T w_2)$$

$$0 \leq w_1^T w_1 + w_2^T w_2 - w_1^T w_2 - w_2^T w_1 =$$

$$\|w_1 - w_2\|_2^2 \geq 0$$

Constrained optimization

$$\min f(w)$$
$$h_i(w) = \phi \quad i = 1, \dots, l$$
$$g_i(w) \le \phi \quad i = 1, \dots, k$$

Primal Problem

p^* = optimal solution

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

α_i: coefficients must be ≥ 0

β_i: coefficients unconstrained

DEF

$$\Theta_p(w) = \max_{\substack{\alpha_i \ge 0 \\ \beta_i}} L(w, \alpha, \beta)$$

If w is feasible (it satisfies all the constraints) g_i are all negative and h_i are all ϕ. So we take $\alpha_i = 0$.

If w is not feasible, we take it to infinity through α's and β's

$$\Theta_p(w) = \begin{cases} f(w) & w \text{ is feasible} \\ \infty & \text{otherwise} \end{cases}$$

$$\min_{w} \max_{\alpha_i \geq 0} L(w, \alpha, \beta)$$

Bc

To solve above pick w first, to minimize and then solve max.

We must pick a feasible w.

$\min_w \theta_p(w)$ = optimal value of the primal problem $= p^*$

$$\max_{\alpha_i \geq 0} \min_{w} L(w, \alpha, \beta) = \max_{\alpha_i \geq 0} \theta_D(\alpha, \beta)$$

bi c Bc

DEF

$$\theta_D(\alpha, \beta) = \min_w L(w, \alpha, \beta)$$

d^* = optimal solution

We can show that $d^* \leq p^*$

$$\max_x \min_y \varphi(x, y) \leq \min_y \max_x \varphi(x, y)$$

$$\min_y \varphi(x^*, y) \leq \varphi(x^*, y^*) \leq \max_x \varphi(x, y^*)$$

Thm.

If f, g_i are convex, g_i strictly feasible, h_i are affine, then

$$p^* = d^*$$

$$h_i(w) = \theta_i^T w + z$$

Strictly feasible means that there is a solution where $g_i(w) < 0$

$\Updownarrow$ strictly $<$

In fact, there exist w^*, α^*, β^* such that

$$p^* = d^* = L(w^*, \alpha^*, \beta^*)$$

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0$$

$$\frac{\partial}{\partial p_i} L(w^*, \alpha^*, \beta^*) = 0$$

$$\boxed{\alpha_i g_i(w^*) = 0} \quad \forall i$$

$$g_i(w^*) \leq 0$$

$$\alpha_i^* \geq 0$$

$$h_i(w^*) = 0$$

We go back to original Lagrangian, we change $\min \|w\|$ to $\min \frac{1}{2}\|w\|$

for convenience

$$L(w,b,\alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x^{(i)} + b) y^{(i)}\right)$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i x^{(i)} y^{(i)} = \phi$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y^{(i)} = \phi$$

Immediate solution is

$$\boxed{w = \sum_{i=1}^{n} \alpha_i x^{(i)} y^{(i)}}$$

We plug this w back into the Lagrangian. Because of the nature of $\frac{\partial L}{\partial b} = 0$, b will not figure into the equation

$$\Theta_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \underline{x^{(i)T} x^{(j)}}$$

$\alpha_i \geq \phi$

This inner product can be replaced by a kernel.

We can compute soln of the dual (even use kernels).

Now we have $\alpha_1, \ldots, \alpha_n$. What is the classifier?

$$x \rightarrow \text{sign}\left(w^T x + b\right)$$

$$\text{sign}\left(\sum_{i=1}^{n} \alpha_i y^{(i)} (x^T x^{(i)}) + b\right)$$

$$b = \left(\max_{i \,:\, y^{(i)} = -1} w^T x^{(i)} - \min_{i \,:\, y^{(i)} = +1} w^T x^{(i)} \right) / 2$$

We may allow for an error, i.e.

$$(w'^T x^{(i)} + b) y^{(i)} \geq 1 - \xi_i \quad i \in \{1, \ldots, n\}$$

$$\min \frac{1}{2} \| w' \|_2^2 + \sum \xi$$

2.10.2009

Last Class

We were searching for maximum-margin classifiers in form for which we can use kernel trick.

$$\min \frac{1}{2}\|w\|_2^2$$
$$y^{(i)}(w^T x^{(i)} + b) \geq 1$$

Dual:

$$\max \sum \alpha_i - \frac{1}{2}\sum y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T (x^{(j)})$$

↑ a constraint for each $x^{(i)}, y^{(i)}$ data point

$$\alpha_i \geq 0$$
$$\sum \alpha_i y^{(i)} = 0$$

We want to add constraints to allow violations

Modification: (Add slack variables)

$$\min \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \qquad \xi_i \geq 0$$
$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$$

Dual:

Same as above with $0 \leq \alpha_i \leq C$

For the project, find the optimal solution for the dual problem above.

Sequential Minimal Optimization (SMO)

Pick two indices i, j

Try to modify α_i, α_j to
- increase objective
- stay in the feasible region

$$\alpha_i = z$$

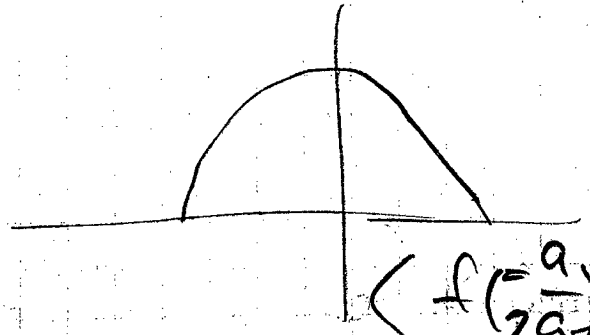
$$\alpha_i y^{(i)} + \alpha_j y^{(j)} = -\sum_{k \neq i,j} \alpha_k y^{(k)} = A$$

$$\alpha_j = (A - y^{(i)} z) y^{(j)}$$

In the end, we are optimizing a quadratic function on an interval.

$\max a_2 z^2 + a_1 z + a_0$ on $[c_1, c_2]$

If $a_2 < 0$

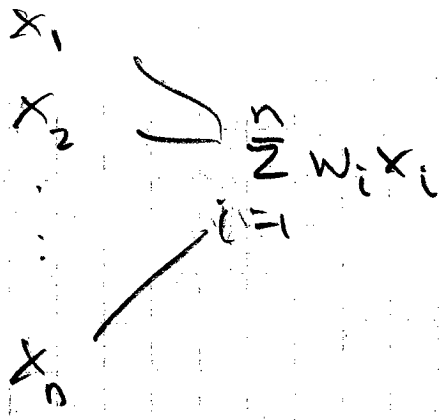


$$z z a_2 + a_1 = 0 \rightarrow z = \frac{-a_1}{2a_2}$$

output is $\begin{cases} f\left(\frac{-a_1}{2a_2}\right), \text{if } \left(\frac{-a_1}{2a_2}\right) \text{ is in } [c_1, c_2] \\ \max\{f(c_1), f(c_2)\}, \text{otherwise} \end{cases}$

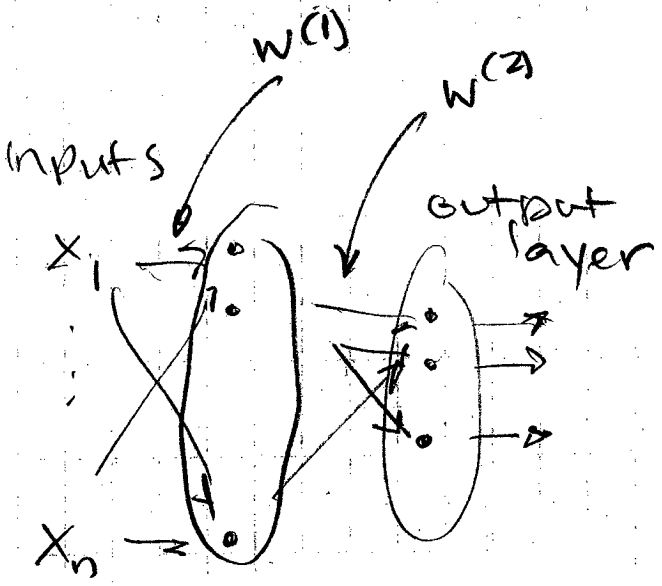
Suggested initial value of α_i is 0, but "it shouldn't matter".

Neural Networks



perceptron

A neural network is a "network of perceptrons"



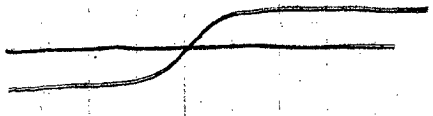
hidden layer

M nodes

We have to add non-linearity to get something interesting.

We use $h()$.

For example $h(x) = \dfrac{-e^{-x} + e^{x}}{e^{-x} + e^{x}}$

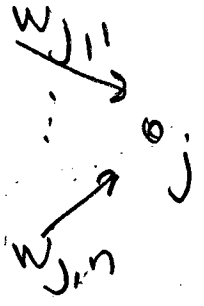


Lets use $h()$ for the hidden layer and $\sigma()$ for the output layer.

If the perceptrons have weights

$w_{j,i}^{(1)}$ for the first layer

$w_{k_j}^{(2)}$ for the second layer



The function computed is

$$y_k = \sigma\left(\sum_j w_{kj}^{(2)} h\left(\sum_i w_{ji}^{(1)} x_i\right)\right)$$

The goal is to find good w using $\bar{x}^{(1)}, \ldots, \bar{x}^{(D)}$ and $\bar{t}^{(1)}, \ldots, \bar{t}^{(D)}$.

Find w's such that

$$\sum_{\ell=1}^{D} \sum_{k=1}^{K} \left(y_k^{(\ell)} - t_k^{(\ell)}\right)^2 \text{ is minimized.}$$

$$y_k^{(\ell)} = \sigma\left(\sum w_{kj}^{(2)} h\left(w_{ji}^{(1)} w_i^{(\ell)}\right)\right)$$

To find w, we use <u>gradient descent (back propagation)</u>

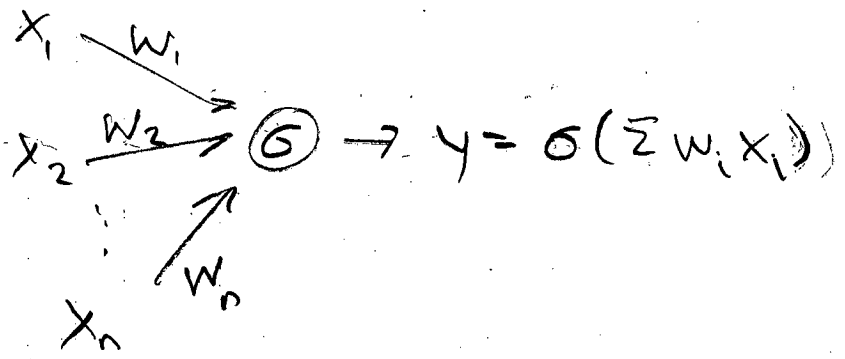
$E(w, X, T)$

$$\nabla w = \frac{\partial E}{\partial w}$$

learning rate

Let $w = w - \eta \cdot \nabla w$

Lets start by solving for 1 perceptron.



$$E(w, x, T) = (y - T)^2$$

What is $\frac{\partial E}{\partial w} = \left(\frac{\partial E}{\partial w_1}, \cdots, \frac{\partial E}{\partial w_n}\right)$

$$\frac{\partial E}{\partial w_\ell} = 2(y - T)\,\sigma'(\Sigma w_i x_i) \cdot x_\ell$$

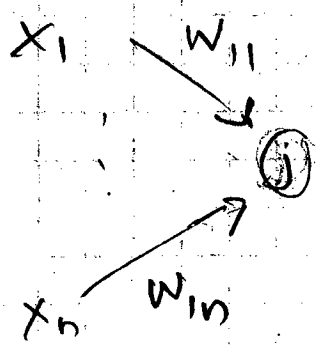
Chain Rule

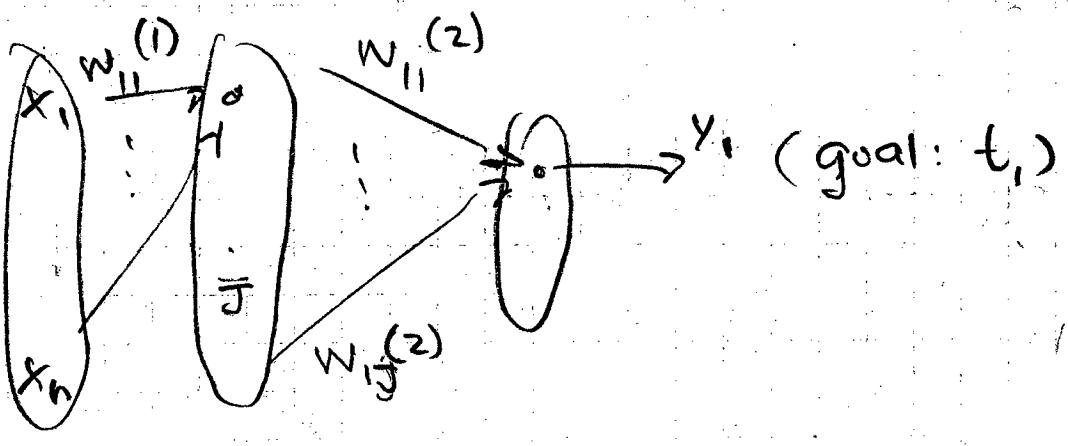
$f(x_1, \cdots, x_n)$, x_i is a function of t

$$\frac{df}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \frac{dx_i}{dt}$$

Let the activation of unit j be

$$a_j = \Sigma w_{ji} x_i$$





$$E = \sum_{k=1}^{K} (y_k - t_k)^2$$

$$\frac{\partial E}{\partial w_{11}^{(2)}} = 2(y_1 - t_1)\frac{\partial y_1}{\partial w_{11}^{(2)}}$$

$$y_1 = \sigma\left(\sum_{j=1}^{J} w_{1j}^{(2)} z_j\right)$$

$$\frac{\partial y_1}{\partial w_{11}^{(2)}} = \sigma'(\dots) \cdot z_1$$

For the first layer

$$E = \sum_{k=1}^{K} (y_k - t_k)^2, \quad y_k = \sigma(\Sigma w^{(2)} h(\Sigma w^{(1)} \dots))$$

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \sum \frac{\partial E}{\partial a_k} \cdot \frac{\partial a_k}{\partial w_{11}^{(1)}}$$

2.11.2009

Tutorial

(1.31) $H(x) + H(y) \geq H(x,y)$

$$H(x) = -\int p(x) \log p(x)\, dx = -\iint p(x,y) \log p(x)\, dx\, dy$$

$$H(x) + H(y) - H(x,y) =$$

$$-\iint p(x,y) \log \frac{p(x)\,p(y)}{p(x,y)}\, dx\, dy =$$

$$E_{x,y}\left(-\log \frac{p(x)\,p(y)}{p(x,y)}\right) \geq -\log E_{x,y} \frac{p(x)\,p(y)}{p(x,y)} = 0$$

$\forall_{x,y}$ $\dfrac{p(x)\,p(y)}{p(x,y)}$ must be 1 for $=$

(1.33) $H[y|x] = 0 \Rightarrow \forall x\ p(x) > 0\ \exists!\, y$ s.t. $p(y|x) = 1$

$$H(y|x) = -\sum_{x,y} p(x,y) \log p(y|x)$$

$$= -\sum_{\substack{x \\ p(x)>0}} p(x) \sum_{y} p(y|x) \log p(y|x) = 0$$

always > 0 ; always ≤ 0

Proof: Assume $\exists x, y$ s.t. $p(x) > 0$ and $0 < p(y|x) < 1$

But then there is a term

$$p(y|x) \log p(y|x) < 0$$

and $H[y|x] > 0$ contradiction

(1.32)

$y = Ax$, A nonsingular

$$H[Y] = H[X] + \ln|A|$$

$$p_Y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

$$\left|\frac{dy}{dx}\right| = |A|$$

$$p_Y(y) = \frac{p_x(x)}{|A|}$$

$$H[Y] = -\int p_Y(y) \ln p_Y(y)\, dy = -\int \frac{p_x(x)}{|A|} \ln \frac{p_x(x)}{|A|} |A|\, dx$$

Jacobian determinant

$$= -\int p_x(x) \ln p_x(x)\, dx + \int p_x(x) \ln|A|\, dx$$

$$= H[X] + \ln|A|$$

2.12.2009

Graphical Models

We want to describe a distribution of $X_1, \dots, X_n$ in a compact form.

Example

$$X_i \in \{0,1\}$$

To describe $X_1, \dots, X_n$ requires 2^n numbers

Perhaps X_1, X_2 and X_3 are generated by some process.

Example

$$X_1 \to X_2 \to X_3$$

$f(X_1)$: $P(X_1 = 0)$, $P(X_1 = 1)$

$f(X_2 \mid X_1)$: $P(X_2 = 0 \mid X_1 = 0)$, $P(X_2 = 1 \mid X_1 = 1)$

$f(X_3 \mid X_2)$: $P(X_3 = 0 \mid X_2 = 0)$, $P(X_3 = 1 \mid X_3 = 1)$

Directed Graphical Model

- directed graph G (acyclic)
- vertices correspond to $X_1, \dots, X_n$ (each vertex one variable)
- we have $f(X_v \mid X_{\pi_v})$ for each vertex v — distribution of X_v given parents

Lets use notation $X_{\{2,3\}} = (X_2, X_3)$

The distribution of $X_1, \ldots, X_n$ described by the model is

$$P(X_1 = a_1, X_2 = a_2, \ldots X_n = a_n) = \prod_{v \in V} f(X_v = a_v \mid X_{\pi_v} = a_{\pi_v})$$

Lets go back to the example.

$X_1 \to X_2 \to X_3$

$f_1(X_1 = 0) = 1/2$ $\qquad$ $f_3(X_3 = 0 \mid X_2 = 0) = 2/3$

$f_2(X_2 = 0 \mid X_1 = 0) = 2/3$ $\qquad$ $f_3(X_3 = 0 \mid X_2 = 1) = 1/3$

$f_2(X_2 = 0 \mid X_1 = 1) = 1/3$

Then, $P(0,1,0) = P(X_1 = 0, X_2 = 1, X_3 = 0) =$

$1/2 \cdot 1/3 \cdot 1/3$

Note that we are using

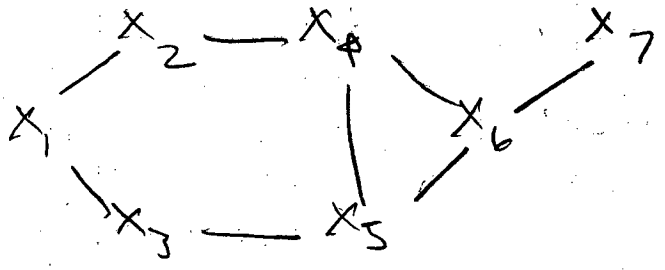
(*) $\sum_{a_v} f(X_v = a_v \mid X_{\pi_v} = a_{\pi_v}) = 1$ for any a_{π_v}

If f's above satisfy (*), then P is a probability.

Fact 1:

Fact 2: $P(X_v \mid X_{\pi_v}) = f(X_v \mid X_{\pi_v})$

Undirected Graphical Model



- undirected graph
- $f(X_c)$ for every maximal clique

In the drawing above, there are 6 maximal cliques

A Maximal clique is a clique that is not a subset of another clique.

$$P(X_1, \cdots, X_n) \propto \prod_c f(X_c)$$
$$= \frac{1}{Z} \prod_c f(X_c)$$

$$P(X_1 = a_1, \cdots, X_7 = a_7) = \frac{1}{Z} f_{13}(a_1, a_3) \cdot f_{456}(a_4, a_5, a_6) \cdot \cdots$$

6 terms

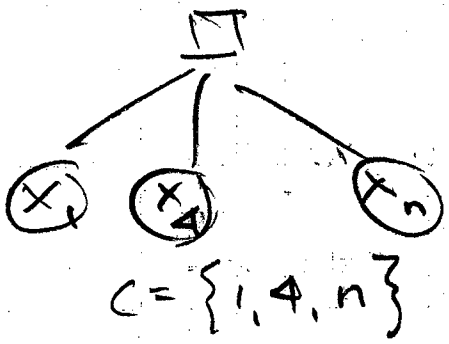
The above holds for the case when we are specifying values for all X_i.

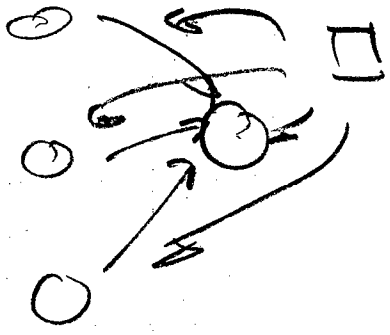
Factor Graphs

collection of subsets $C \subseteq \{1, \dots, n\}$

f_c for each of your chosen c

$P(x_1, \dots, x_n) \propto \prod f_c(x_c)$





$f_{\Pi_C} = f(x_v \mid x_{\pi_v})$

To go from ~~factor graph to undirected graph~~:

- Take G.
- For each v, add undirected edges between the parents of v.
- Forget the direction of the original edges. They become undirected

Returning to undirected graphs....

$P(X_3 = 0) = \frac{1}{2}$

0 0 0
0 1 0
1 0 0
1 1 0
⋮

How do we calculate probabilities of subsets of the variables?

$$P(X_1 = a_1) = \sum_{a_2} \sum_{a_3} \cdots \sum_{a_7} P(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n)$$

$$= \frac{1}{Z} \sum_{a_2} \cdots \sum_{a_7} f_{12}(a_1, a_2) f_{24}(a_2, a_4) \cdots$$

$$f_{13}(a_1, a_3) f_{35}(a_3, a_5) f_{456}(a_4, a_5, a_6) \cdot f_{67}(a_6, a_7)$$

The above would take D^6 time

What if we precomputed

$$\boxed{m_{6,7}(a_6) = \sum_{a_7} f_{6,7}(a_6, a_7)}$$

for all a_6, it would take D^2 time.

We have gotten rid of one of summations

$$\frac{1}{Z} \sum_{a_2} \sum_{a_3} \sum_{a_4} \sum_{a_5} \sum_{a_6} \rightsquigarrow m_{6,7}(a_6)$$

Our summation now takes D^5 time.

Lets try again

$$m_{4,5,6}(a_4, a_5) = \sum_{a_6} f_{4,5,6}(a_4, a_5, a_6) m_{6,7}(a_6)$$

Computing $M_{4,5,6}$ for all a_4, a_5 requires D^3 time.

Now, we have

$$P(\ ) = \frac{1}{Z} \sum_{a_2} \sum_{a_3} \sum_{a_4} \sum_{a_5} f_{1,2}(a_1,a_2) f_{1,3}(a_1,a_3) f_{2,4}(a_2,a_4) \cdot f_{3,5}(a_3,a_5) \cdot M_{4,5,6}(\ )$$

This requires D^4 time.

Now, what if we tried to eliminate a_3?

$$\sum_{a_3} f_{1,3}(a_1,a_3) f_{3,5}(a_3,a_5) = m_{1,5,3}(a_1,a_5)$$

$\}$ D^3 time

This leaves

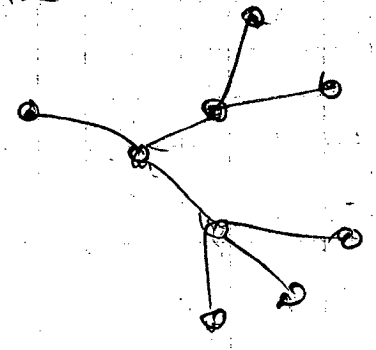
$$\frac{1}{Z} \sum_{a_2} \sum_{a_4} \sum_{a_5} f_{1,2}(a_1,a_2) f_{2,4}(a_2,a_4) m_{1,5,3}(a_1,a_5) \cdot M_{4,5,6}(a_4,a_5)$$

$\uparrow$ D^4 time

$$Z = \sum_{a_1} \cdots \sum_{a_7} \cdots$$

What we have been doing is called the elimination algorithm.

Given a graph, G, what order should we choose for elimination? What is the running time.



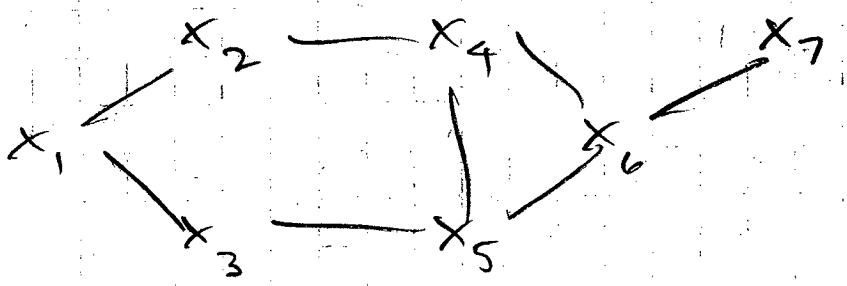
If G is a tree, we use D^2 time to eliminate a leaf node.

If G is a tree, always process a leaf. Running time $\sim nD^2$

In general, the running time is $D^{\text{treewidth}(G)}$

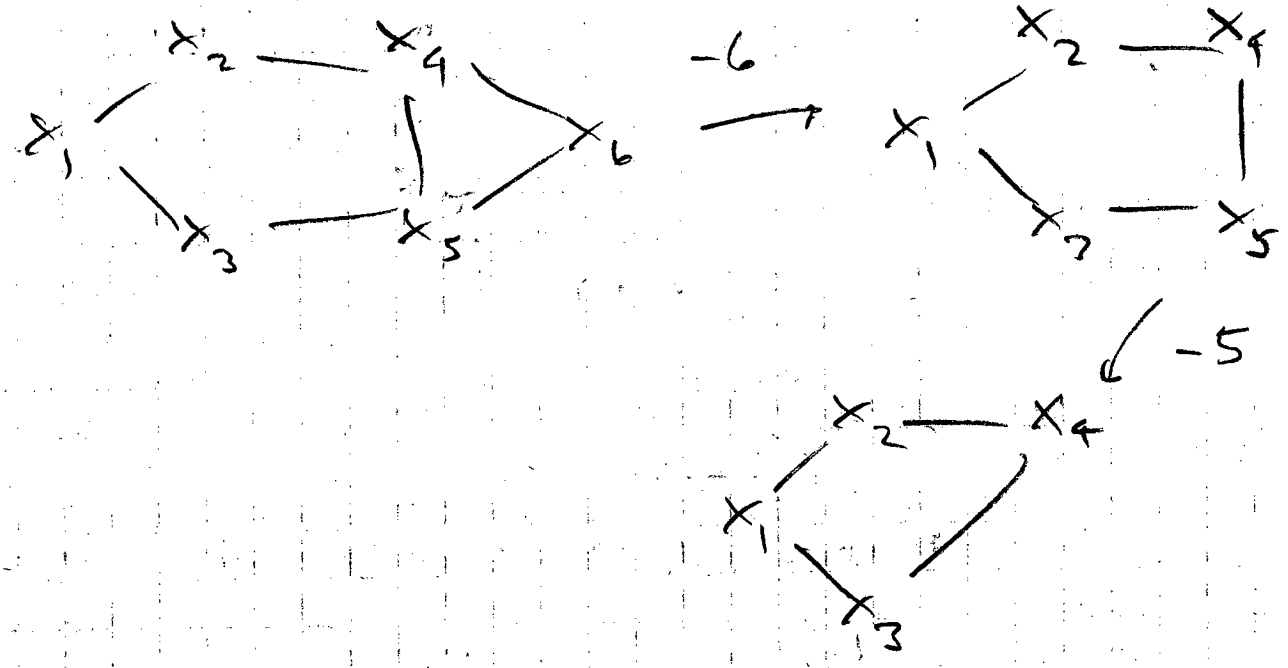
Treewidth of G is the minimum, over all possible orderings of the vertices of G, of the maximum degree of an eliminated vertex in the following process:

To eliminate a vertex, put edges between all of its neighbors!...



7, 6, 5, 4, 2, 3, 1

Eliminate 7

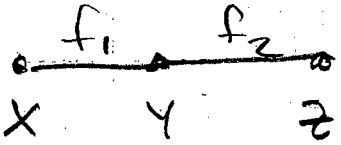


tw = 2 + 1

tree width of a tree is 2

2.17.2009

Conditional Independence in Graphical Models



Are X, Z independent?

not necessarily

$f_1(0,0) = 1$

$f_1(1,1) = 1$

$f_1(0,1) = 0$

$f_1(1,0) = 0$

$f_2 = f_1$

$P(X=0, Y=1, Z=1) = 0$

$P(X=Y=Z=1) = 1/2$

$P(X=Y=Z=0) = 1/2$

X, Z are not independent

Are X, Z independent given Y?

U, W are independent $\Leftrightarrow$

$\forall u,w \quad P(U=u \wedge W=w) = P(U=u)\,P(W=w)$

U, W are independent given $\varphi \Leftrightarrow$

$\forall u,v,q \quad P(U=u \wedge W=w \mid \varphi=q) =$

$P(U=v \mid \varphi=q)\; P(W=w \mid \varphi=q)$

In the concrete example,

$P(X=1 \wedge Z=0 \mid Y=0)$

Check all 8 cases to show that X, Z are independent given Y for this example

In general

$$P(X,Z \mid Y) \stackrel{?}{=} P(X \mid Y)\, P(Z \mid Y)$$

$$\frac{P(X,Y,Z)}{P(Y)} = \frac{f_1(X,Y)\, f_2(Y,Z)}{\sum_{X'} \sum_{Z'} f_1(X',Y)\, f_2(Y,Z')}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \quad \frac{\sum_{Z'} f_1(X,Y) f_2(Y,Z')}{P(Y)} =$$

$$\frac{f_1(X,Y) \sum_{Z'} f_2(Y,Z')}{\sum_{X'} \sum_{Z'} f_1(X',Y) f_2(Y,Z')}$$

$$P(X \mid Y)\, P(Z \mid Y) = \frac{f_1(X,Y) \sum_{Z'} f_2(Y,Z') \cdot f_2(Y,Z) \cdot \sum_{X'} f_1(X',Y)}{\left(\sum_{X'} \sum_{Z'} f_1(X',Y) f_2(Y,Z')\right)^2}$$

$$\frac{P(X,Y,Z)}{P(Y)} = \frac{f_1(X,Y)\, f_2(Y,Z)}{\left(\sum_{X'} f_1(X',Y)\right)\left(\sum_{Z'} f_2(Y,Z')\right)}$$

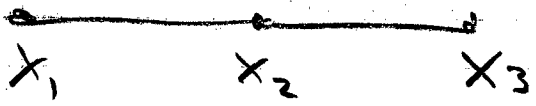
equal

So, its true that $X \perp\!\!\!\perp Z \mid Y$ (X and Z are independent given Y)

For $S \subseteq V$ ~~~ vertices $= V_1, \ldots, V_k$

Let $X_S = (X_{V_1}, \ldots, X_{V_k})$

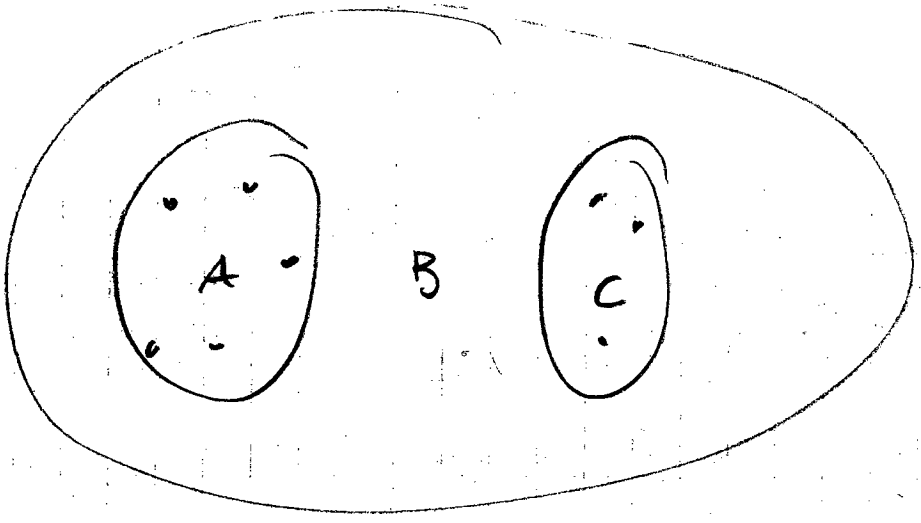
For example



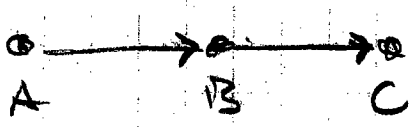
$X_{\{1,3\}} = (X_1, X_3)$

$A, B, C \subseteq V$

In general $X_A \perp\!\!\!\perp X_C \mid X_B$ if B separates A from C



Lets look at directed graphs.



Is $X_A \perp\!\!\!\perp X_C \mid X_B$? YES

Is $X_A \perp\!\!\!\perp X_C$? NO

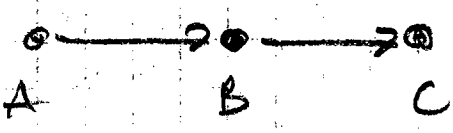
$P(A,B,C) = P(A)\,P(B|A)\,P(C|B)$

We want to show $P(A,C|B) = P(A|B)\,P(C|B)$

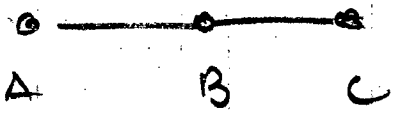
$$P(A,C|B) = \frac{P(A,B,C)}{P(B)} = \frac{P(A)\,P(B|A)\,P(C|B)}{P(B)} = P(A|B)\,P(C|B)$$

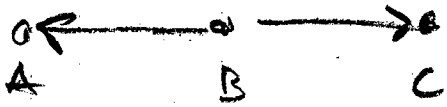
$P(A|B)$
By Bayes

Note that

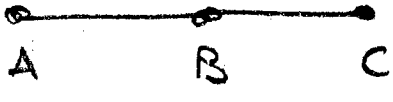


reduces to

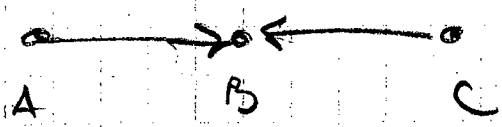




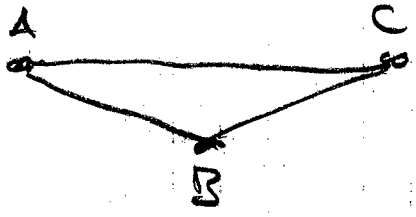
reduces to



what about



reduces to



$X_A \perp\!\!\!\perp X_C \mid X_B$? NO

$X_A \perp\!\!\!\perp X_C$? YES

$P(A,B,C) = P(A)\,P(C)\,P(B|A,C)$

$$\frac{P(A,B,C)}{P(B|A,C)} = P(A,C) = P(A)\,P(C)$$

$X_A \in_R \{0,1\}$

$X_C \in_R \{0,1\}$

$X_B = X_A + X_C \mod 2$

Let $X_B = 0$

Then $P(X_A = 0 \wedge X_C = 1 \mid X_B = 0) \stackrel{?}{=}$

$\stackrel{\swarrow}{0}$ $\qquad P(X_A = 0 \mid X_B = 0) \cdot P(X_C = 1 \mid X_B = 0)$

$\qquad\qquad \stackrel{\swarrow}{1/2} \qquad\qquad \stackrel{\swarrow}{1/2}$

So $X_A \not\perp X_C \mid X_B$

To prove independence must exist for all edge functions

To disprove $\exists$ edge function s.t. not independent.

In undirected, conditional independence implies that there is no path from A to C that does not go through B.

For directed,

We say that A, C are d-connected w.r.t. B if there exists an undirected path from $a \in A$ to $c \in C$ such that every vertex $v \in P$ for

1. if v is non-blocking, v is not in B
2. if v is blocking, v or its descendent is in B

Blocking $\rightarrow \circ \leftarrow$ (V)

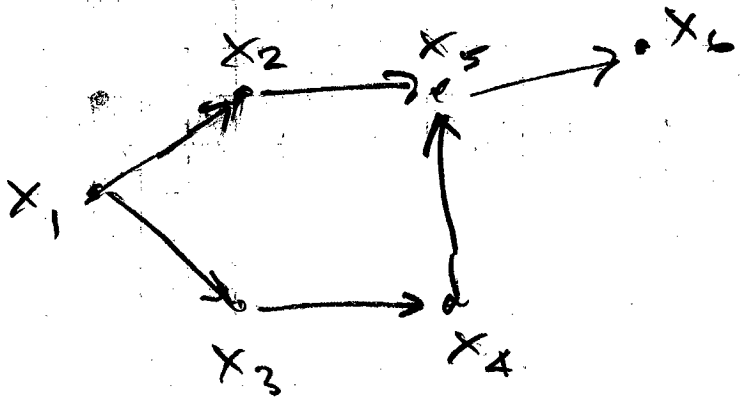
DEF

A, C are d-separated by B if they are not d-connected w.r.t. B

THM

If A, C are d-separated by B, then

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

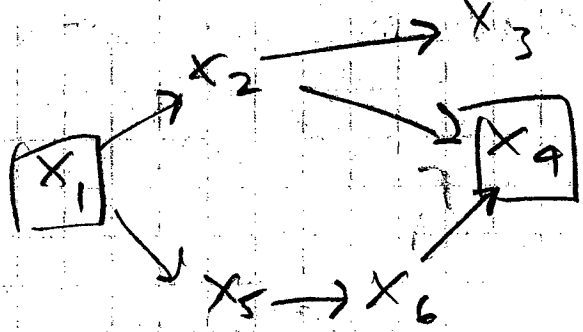


$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$? YES

Look at paths

X_1, X_2, X_5 separated

X_1, X_3, X_4, X_5 separated



$X_2 \perp\!\!\!\perp X_5 \mid X_1, X_4$?

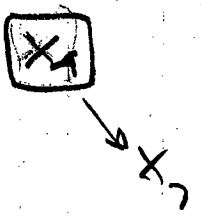
X_5, X_1, X_2 is not a d-connecting path

So, X_2, X_5 is d-separated by X_1.

X_5, X_6, X_4, X_2 is a d-connecting path.

$\therefore X_2 \perp\!\!\!\perp X_5 \mid X_1, X_4$ is <u>FALSE</u>

If you added X_7 to the graph s.t.

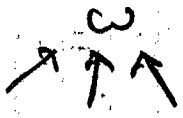


then $X_2 \perp\!\!\!\perp X_5 \mid X_1, X_7$ is also FALSE. This shows the descendent case.

<u>Proof of THM</u> (G is a directed acyclic graph (DAG))

Induction on the number of vertices in G.
Skip the base, single vertex, case

Let w be a sink of G



Case : $W \notin A \cup B \cup C$

A and C are d-separated by B in G

$\Downarrow$

A and C are d-separated by B in G-w

$\Downarrow$

$X_A \perp\!\!\!\perp X_C \mid X_B$

We choose a sink w because there will always be a sink in a directed graph.

2.19.2009

Continuation of D-Separation Proof...

Directed graphical model $G=(V,E)$

$A, B, C \subseteq V$ (disjoint)

DEF

A is d-connected to C w.r.t. B if
$\exists$ directed path $P = q_1, \ldots, c$

$q \in A$
$c \in C$

for every $v \in P$
- if v is collider then v or an ancestor of v is in B
- if v is not a collider, then $v \notin B$

A is d-separated from C by B if A is not d-connected to C w.r.t. B

Theorem

If A is d-separated from C by B, then

$$X_A \perp\!\!\!\perp X_C \mid X_B$$

Proof

Induction on $|V|$

- Let w be a vertex of out-degree ϕ in G (i.e. w is a sink)
- Let $V' = V - w$, $G' = (V', E')$

Fact:

$X_{V'}$ factorizes over G'

Case 1: $W \notin A \cup B \cup C$

Clearly A is d-separated from C by B in G'

By IH (inductive hypothesis)

$X_A \perp\!\!\!\perp X_C \mid X_B$ in G'

$\Downarrow$

$X_A \perp\!\!\!\perp X_C \mid X_B$ in G

Case: $W \in A$

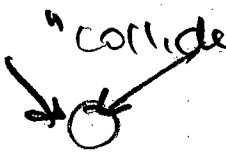
- Let $A' = A - W$
- Let P be the set of parents of W which are not in B

Claim 1:
A' is d-separated from C by B in G'
Claim 2: P is d-separated from C by B in G'

Claim 1 is satisfied by assumption of theorem.

Proof of Claim 2:
Suppose that the claim is false, i.e.
$\exists$ good path from $c \in C$ to $p \in$ Parents$(W) - B$

Then the path extended by (p, w) is good, too



If this was true, then A is d-connected to C w.r.t. B. This is a <u>contradiction</u>.

By IH,

$X_{A' \cup P} \perp\!\!\!\perp X_C \mid X_B$

$X_P \perp\!\!\!\perp X_C \mid X_B$

Fact

$X_W \perp\!\!\!\perp X_{any} \mid X_P$

anything that doesn't include W or P

$A \perp\!\!\!\perp B \mid C \cup D \;,\; A \perp\!\!\!\perp D \mid C \Rightarrow A \perp\!\!\!\perp B \cup D \mid C$ (2)

$U \perp\!\!\!\perp V \mid W \overset{\text{def}}{\Leftrightarrow} P(U \mid V, W) = P(U \mid W)$

Given the above

$P(A \mid B, C, D) = P(A \mid C, D)$

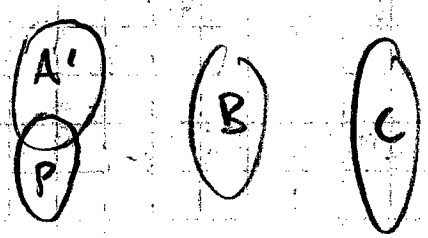
$P(A \mid C, D) = P(A \mid C)$

$P(A \mid B, C, D) = P(A \mid C)$

$A \perp\!\!\!\perp B \cup D \mid C$

Going back to our fact

$X_W \perp\!\!\!\perp X_C \mid X_{A' \cup P \cup B}$



We want to conclude that

$$X_{A' \cup \{W\}} \perp\!\!\!\perp X_C \mid X_B$$

Repeating, ...

By the IH

$$X_{A' \cup P} \perp\!\!\!\perp X_C \mid X_B \qquad (2)$$

$$X_W \perp\!\!\!\perp X_C \mid X_{B \cup A' \cup P} \qquad (1)$$

Conclusion

$$X_{A' \cup W \cup P} \perp\!\!\!\perp X_C \mid X_B \Rightarrow X_{A' \cup W} \perp\!\!\!\perp X_C \mid X_B$$

Map to laws on previous page

A	X_C
B	X_W
C	X_B
D	$X_{A' \cup P}$

$A \perp\!\!\!\perp B \mid C \Leftrightarrow P(A \mid B, C) = P(A \mid C)$ D1

$B \perp\!\!\!\perp A \mid C \Leftrightarrow P(B \mid A, C) = P(B \mid C)$

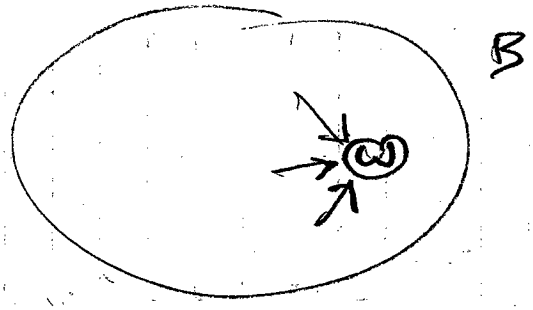
$A \perp\!\!\!\perp B \mid C \Leftrightarrow P(A \mid C)\, P(B \mid C) = P(A, B \mid C)$ D2

$P(A \mid C)\, P(B \mid C) = P(A, B \mid C) = P(A \mid B, C)\, P(B \mid C)$

$\therefore$ Independence is symmetric

Case: $W \in B$

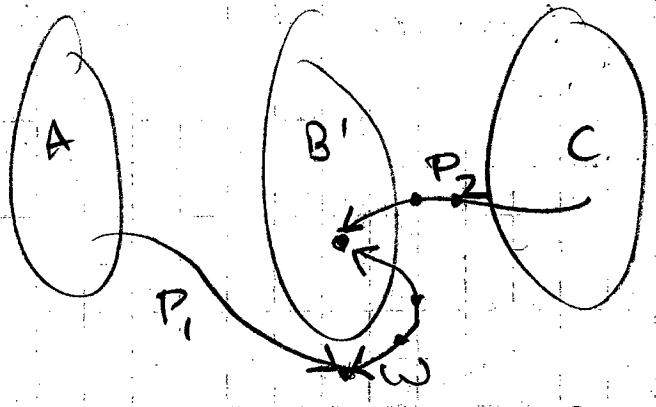
Claim
A is separated from C by $B' = B - W$



Proof

Let P be a path d-connecting A to C w.r.t. B'. Then $W \notin P$ and W is a path d-connecting A to C w.r.t B. This is a contradiction.

Claim
W is d-separated from A or C by B'



Suppose the claim is false. Then $P_1 + P_2$ is a path d-connecting A, C w.r.t. B'.

Say, w.o.l.o.g, that W is d-separated from C by B'. Then

$A \cup W$ is d-separated from C by B'

By IH

$X_{A \cup W} \perp\!\!\!\perp X_C \mid X_{B'}$

$A \perp\!\!\!\perp (B \cup D) \mid C \Rightarrow A \perp\!\!\!\perp B \mid C \cup D$ Rule

---

using the above

$X_A \perp\!\!\!\perp X_C \mid X_{B' \cup W} \Rightarrow X_A \perp\!\!\!\perp X_C \mid X_B$

Sum-Product Algorithm

- Assume an undirected graphical model where G is a tree
- We have $f_{ij}(X_i, X_j)$ for every edge $(i,j) \in G$

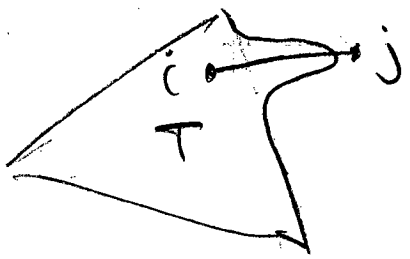
$$P(X_1, \cdots, X_n) \propto \prod_{i,j} f_{ij}(X_i, X_j)$$

Goal: Compute $P(X_i)$

M_{ij} is a message from i to j

$X_1, \cdots, X_n \in \{1, \cdots, D\}$

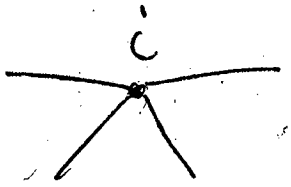
$m_{ij}(X_j)$ = "weight of the subtree rooted at edge i,j if value of X_j is given"



$$m_{ij}(x_j) = \sum_{x_i} f_{ij}(x_i, x_j) \prod_{k} m_{ki}(x_i)$$

k: neighbors of i excluding j

Messages start at the leaves and proceed toward the root.



$$P(x_i) = \frac{1}{Z} \prod_{k} m_{ki}(x_i)$$

k: neighbors of i

$$Z = \sum_{x_i} \prod_{k} m_{ki}(x_i)$$

2.24.2009

EM Algorithm

How do we implement ML?

Lets use an example model of mixtures of Gaussians. In one dimension

$$f(x) = \sum_{i=1}^{k} \pi_i N(x | \mu_i, \Sigma_i) \quad , \quad \Sigma_i = \frac{1}{\sigma_i^2}$$

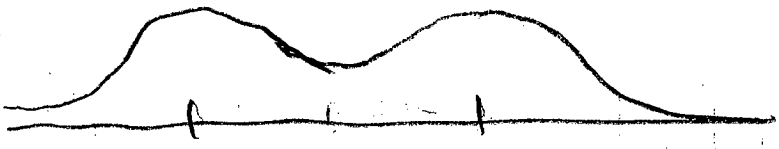
(π_i: mixture coefficients $\sum \pi_i = 1$)

Input

datapoints $x_1, \ldots, x_n \in \mathbb{R}$

Output

$$\frac{1}{2} N(x | -1, 1) + \frac{1}{2} N(x | +1, 1)$$



↓ setting of $\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k$ maximizing the likelihood of the input data $(x_1, \ldots, x_n)$.

What we are maximizing is

$$P(x_1, \ldots, x_n \mid \pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$$

Since we don't like products, we take the logarithm of the likelihood!

$$\log P(\dots) = \sum_{i=1}^{n} \log P(x_i | \dots) =$$

$$= \sum_{i=1}^{n} \log \sum_{j=1}^{k} \pi_j \frac{e^{-(x_i-\mu_j)^2/2\sigma_j^2}}{\sqrt{2\pi}\,\sigma_j} = \varphi$$

Lets think of the mixture model as a model with a hidden variable ("that tells us which Gaussian to use")

The hidden variables are

$$z_1, \dots, z_k \;,\; \sum z_i = 1 \;,\; z_i \in \{0,1\}$$

$z_i \in \{0,1\}$ ← binary variable

$$P(z_i = 1) = \pi_i$$

$$P(x | z_i = 1) = N(x | \mu_i, \sigma_i)$$

This gives the same distribution as f. For each x, only one of the z_i is 1.

Suppose we knew the parameters π_i, μ_i, and σ_i, which component did x come from?

$$P(z_j = 1 | x) = \frac{P(x | z_j = 1)\, P(z_j = 1)}{P(x)}$$

$$= \frac{\pi_j N(x | \mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l N(x | \mu_l, \Sigma_l)}$$

$$\frac{\partial \varphi}{\partial \mu_j} = \sum_{i=1}^{n} \frac{\partial \varphi_i}{\partial \mu_j}$$

Let $\varphi_i = \log \sum_{j=1}^{k} \pi_j N(x_i | \mu_j, \Sigma_j)$

Since $(\log f)' = \frac{f'}{f}$,

$$\frac{\partial \varphi_i}{\partial \mu_j} = \frac{\frac{\partial N(x_i | \mu_j, \Sigma_j)}{\partial \mu_j} \pi_j}{\sum_{l=1}^{k} \pi_l N(x_i | \mu_l, \Sigma_l)}$$

$$\frac{\partial N(x | \mu_j, \Sigma_j)}{\partial \mu_j} = N(x | \mu_j, \Sigma_j) \Sigma_j (x - \mu_j)$$

So,

$$\frac{\partial \varphi_i}{\partial \mu_j} = \frac{N(x_i | \mu_j, \Sigma_j) \cdot \Sigma_j (x_i - \mu_j) \pi_j}{\sum_{l=1}^{k} \pi_l N(x_i | \mu_l, \Sigma_l)}$$

and

$$\frac{\partial \varphi}{\partial \mu_j} = \sum_{i=1}^{n} \frac{\pi_j (x_i - \mu_j) \Sigma_j N(x_i | \mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l N(x_i | \mu_l, \Sigma_l)} = \emptyset$$

$$= \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j) \Sigma_j = \emptyset$$

$$\gamma_{ij} = \frac{\pi_j N(x_i | \mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l N(x_i | \mu_l, \Sigma_l)} = P(z_j = 1 | x_i)$$

Now, we can solve for all μ_j's.

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{n} \gamma_{ij} x_i \,, \quad N_j = \sum_{i=1}^{n} \gamma_{ij} \qquad (1)$$

If we know γ_{ij}'s, then we can compute μ_j

We will not derive the expression for Σ_j, π_j

$$\Sigma_j = \frac{1}{N_j} \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)^2 \,, \quad \text{where the } \mu_j \text{ are calculated by previous formula (1)} \qquad (2)$$

$$\pi_j = \frac{N_j}{\sum_{l=1}^{k} N_l} \qquad (3)$$

EM Algorithm $\quad \pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}$

- Initialize π_j's, μ_j's, and Σ_j's somehow
- E-step: Compute $\gamma_{ij}^{(t+1)}$ using $\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}$
- M-step: Compute $\pi^{(t+1)}, \mu^{(t+1)}, \Sigma^{(t+1)}$ using (1), (2), (3), and $\gamma_{ij}^{(t+1)}$

EM Algorithm (In general)

Model: X observable $\rightarrow P(X, Z \mid \Theta)$

Z hidden (latent) variables

Θ parameters

Goal: Find settings of Θ maximizing $P(X \mid \Theta)$

$$P(X|\theta) = \sum_z P(X, z|\theta)$$

Goal:

$$\arg\max_\theta \log \sum_z P(X, z|\theta)$$

$$L(q, \theta) := \sum_z q(z|X) \log \frac{P(X, z|\theta)}{q(z|X)}$$

EM Algorithm

- Set $\theta^{(0)}$ somehow
- E-step: $q^{(t+1)} = \arg\max_q L(q, \theta^{(t)})$
- M-step $\theta^{(t+1)} = \arg\max_\theta L(q^{(t+1)}, \theta)$

The q's should correspond to the r_{ij} in the Gaussian case

What should q be in the E-step?

2.26.2009

EM Algorithm for models with hidden variables.

$\underset{\theta}{\operatorname{argmax}}\ p(x|\theta)$ ← GOAL

What we have is $p(x, z|\theta)$.

So, the goal is really a marginal distribution

$$\underset{\theta}{\operatorname{argmax}} \sum_z p(x, z|\theta)$$

$$L(q, x, \theta) = \sum q(z|x) \log \frac{p(z, x|\theta)}{q(z|x)}$$

Let $l(x, \theta) = \log p(x|\theta)$

EM Algorithm Steps

0. First set θ to something (i.e. $\theta^{(0)}$)
1. $q^{(t+1)} = \underset{q}{\operatorname{argmax}}\ L(q, x, \theta^{(t)})$
2. $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\ L(q^{(t+1)}, x, \theta)$

"q is a distribution on the hidden variable"

The EM algorithm is an unsupervised learning method.

For the mixture of Gaussians, we had

1. calculate $p(z_i|x_j) = \gamma_{ij}$
2. Use the above to estimate the parameters of the Gaussians (i.e. μ_i's, σ_i's, α_i's)

In general, it is true that

$$q^{(t+1)} = p(z \mid x, \theta^{(t)})$$

Lets look at

$$l(x;\theta) - L(q, x, \theta) =$$

$$\sum_z q(z|x)\left(\log p(x|\theta) - \log \frac{p(z,x|\theta)}{q(z|x)}\right) =$$

$\log p(x|\theta)$: $l(x,\theta)$

$$\sum_z q(z|x)\left(\log \frac{q(z|x)}{p(z|x,\theta)}\right) =$$ ← This is the KL difference between q and $p(z|x,\theta)$

$$KL(q(z|x) \parallel p(z|x,\theta))$$

$$\frac{p(z,x|\theta)}{p(x|\theta)} = p(z|x,\theta)$$

$KL(\,)$ is always positive. A minimum occurs when 2 distributions are equal.

$$l(x,\theta) \geq L(q,x,\theta) \quad \text{since } KL(\cdots) \geq 0$$

Fact 1 ↑

Fact 2 ↓

$$\underset{q}{\text{argmax}}\, L(q,x,\theta) = p(z|x,\theta)$$

Fact 2 is true because maximizing $L(q,x,\theta)$ is the same as minimizing $l(x,\theta) - L(q,x,\theta) = KL(\cdots$

Is $\ell(x, \theta^{(t+1)}) \geq \ell(x, \theta^{(t)})$?

In step 2, $L(q^{(t+1)}, x, \theta)$ will be at least

$L(q^{(t+1)}, x, \theta) \geq \ell(x, \theta^{(t)})$

$\therefore \ell(x, \theta^{(t+1)}) \geq \ell(x, \theta^{(t)})$

After the E-step

$L(q^{(t+1)}, x, \theta^{(t)}) = \ell(x, \theta^{(t)})$

After the M-step

$L(q^{(t+1)}, x, \theta^{(t+1)}) \geq L(q^{(t+1)}, x, \theta^{(t)})$

and

$\ell(x, \theta^{(t+1)}) \geq L(q^{(t+1)}, x, \theta^{(t+1)})$

$\therefore \ell(x, \theta^{(t+1)}) \geq \ell(x, \theta^{(t)})$

Sampling

We have a density function p? How do we sample x with density p?

Approach #1:

Sample $x \sim$ Uniform $[0,1]$

Output $z = f(x)$

$P(Z \leq z) = P(f(x) \leq z) = P(x \leq g(z)) = g(z)$

g is the inverse of f

because of uniform distribution

$$g(x) = \inf \{ u ; f(u) \geq x \}$$

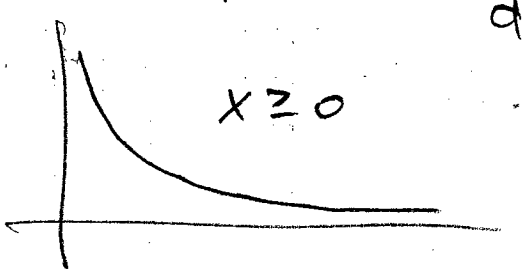
Given density p, compute the cumulative distribution function

$$g(x) = \int_{-\infty}^{x} p(t)\,dt$$

Take f to be the inverse of g.

Example ($\mathrm{Exp}(\lambda)$) – Sample from exponential distribution

$$p(x) = \lambda e^{-\lambda x}$$



$$g(x) = \int_0^x \lambda e^{-\lambda t}\,dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}$$

$$f(1 - e^{-\lambda x}) = x$$

$$u \mapsto \frac{\ln(1-u)}{\lambda}$$

How would we sample from the normal distribution?

$X \sim \mathrm{Uniform}([0,1])$

$Y \sim \mathrm{Exp}(1)$

$Z_1 = \sqrt{2Y}\sin 2\pi X$

$Z_2 = \sqrt{2Y}\cos 2\pi X$

Then Z_1, Z_2 are independent samples from $N(0,1)$

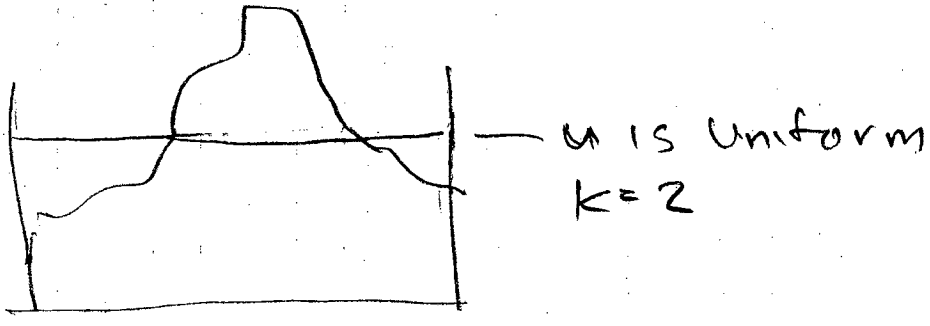
$X \sim N(\mu, \sigma)$ ⟶ $\frac{X-\mu}{\sigma} \sim N(0,1)$

$X \sim N(0,1)$ ⟶ $\sigma X + \mu \sim N(\mu, \sigma)$

Rejection Sampling

We have a density function p that we want to sample from

Find q such that
- we know how to sample from q
- $q(x) \geq \frac{1}{k} p(x)$ for reasonably small k



1. Sample x from q
2. Accept x with probability $\frac{p(x)}{k q(x)}$

   otherwise

What is the probability of accepting x on the first try?

$$q(x) \cdot \frac{p(x)}{q(x) \cdot k} = \frac{p(x)}{k}$$

$P(\text{outputting } x \mid \text{accepting on first try}) = p(x)$

$$P(\text{accepting some } x) = \sum_x q(x) \cdot \frac{p(x)}{q(x)k} = \frac{1}{k}$$

$$p(x) = (2+\sin x)e^{-x} \cdot \frac{1}{z}$$

z is chosen so that $\int p(x)\,dx = 1$

Chose $q(x) = e^{-x}$

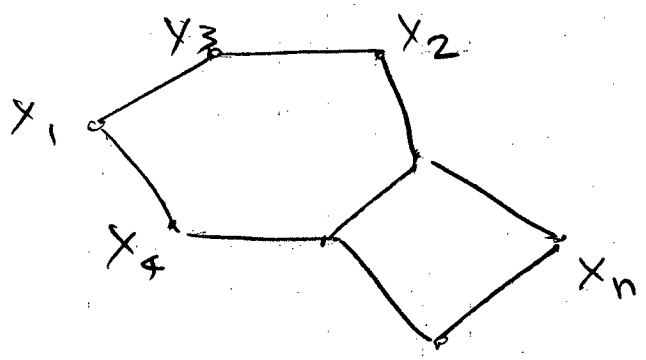
$$k = \max_x \frac{p(x)}{q(x)} = \max_x (2+\sin x) \cdot \frac{1}{z} \leq 3$$

Then, sample $\sim \text{Exp}(1)$.
Accept with probability $\frac{2+\sin x}{3z}$

Markov Chain Monte Carlo

Example:
Sampling from a distribution with an undirected graphical model.

Gibbs sampler



Pick $i \in \{1, \cdots, n\}$

Change

$$X_i \leftarrow p(X_i \mid X_1, \cdots X_{i-1}, X_{(+1)}, \cdots X_n)$$

3.3.2009

We want to sample from an "ugly" distribution. We design a process..."

A Markov chain is a sequence of random variables

$X_1, \dots, X_n, \dots$ ← Lets assume these are discrete.

such that

$$P(X_{n+1} | X_n) = P(X_{n+1} | X_1, \dots, X_n)$$

This is time homogeneous MC if

$$P(X_{n+1} = a \mid X_n = b) \left[= P(X_{n+1} = a \mid X_n = b, X_1, \dots, X_{n-1}) \right]$$

$$\| \\ P_{a,b}$$

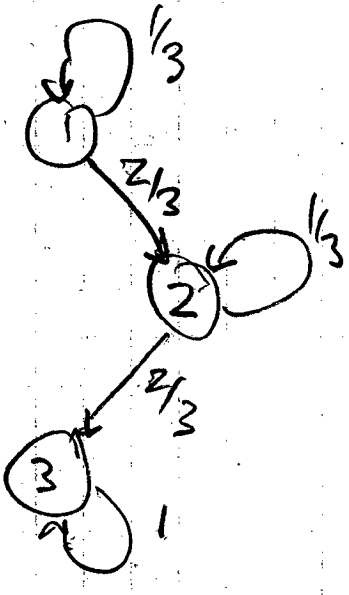
← transition probability from a to b (not a function of n)

Lets assume $X_i \in \{1, \dots, n\}$

$\{P_{a,b}\}_{a=1,b=1}^{n,n}$ ← transition matrix of the Markov chain

P^4

$$P = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 1/3 & 2/3 \\ 0 & 0 & 1 \end{pmatrix}$$



$\xrightarrow{(1,0,0)}$

If $X_1 = 1$, $X_2 = \begin{cases} 2 & P = 2/3 \\ 1 & P = 1/3 \end{cases}$

If X_n has distribution ν, then X_{n+1} has distribution νP

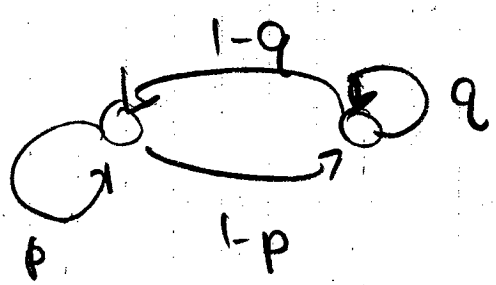
DEF

The stationary probability of a chain with transition matrix P is a probability distribution π such that

$$\pi P = \pi$$

For the example $\pi = (0, 0, 1)$

$$(0,0,1)\begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 1/3 & 2/3 \\ 0 & 0 & 1 \end{pmatrix} = (0,0,1)$$



$$(x, 1-x)\begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix} = (x, 1-x)$$

$$xp + (1-x)(1-q) = x$$

So, VP gives us the next state.

$$P\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \checkmark$$

The all ones vector, $\vec{1}$, is a right-eigenvector of P (with eigenvalue 1)

In $\pi P = \pi$, π is a left-eigenvector corresponding to eigenvalue 1.

Frobenius-Perron

Let P be a matrix with non-negative entries. Then P has a non-negative eigenvalue, λ, with non-negative eigenvector V.

For all other eigenvalues μ of P,

$$|\mu| \leq \lambda .$$

If we replace "non-negative" with "positive" everywhere in the above, we get

$$|\mu| < \lambda .$$

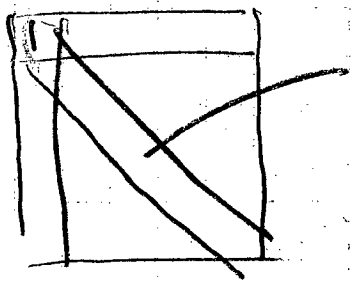
If we have a matrix P and we are computing P^m and can write $P = SDS^{-1}$, D a diagonal matrix, then

$$P^m = (SDS^{-1})^m = SD^mS^{-1}$$

Note that D is a diagonal matrix whose entries are the eigenvalues of P

If we can't diagonalize P, then we can use Jordan Normal Form. We will not talk about this further.

If P has all entries positive (and we assume it's diagonalizable), how does D^M look?

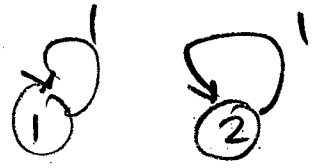


All of these are <1 in absolute value. Therefore, the M^{th} powers go to 0 as $M \to \infty$.

This speaks to the stationary distribution.

The conclusion is that all entries in P positive implies there is a unique stationary distribution

The follow is not unique:



$X_1 = 1, X_2 = 1, \ldots$ or $X_1 = 2, X_2 = 2, \ldots$

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}$$

Note that not all entries in P are positive.

$x + y = 1, \quad x, y \geq 0$

It is sufficient to have z such that all entries in P^z are positive.

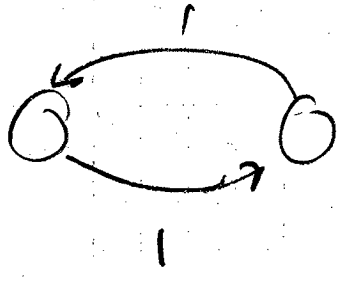
How do we achieve this? What are the obstacles

To have all entries in P^2, we need

1. connected (for every pair of states a, b

$$\exists \{ c_1, \ldots, c_q \;,\; P_{c_i, c_{i+1}} > 0 \}$$
$$c_1 = a \qquad c_q = b$$

What about



0	1
1	0

$$(1/2, 1/2)\ P = (1/2, 1/2)$$

All entries in P positive

$\Downarrow$

There is a unique stationary distribution π

and

$$\lim_{m \to \infty} \| V P^m - \pi \|_1 = 0 \quad \text{for any } V$$

So, we need

2. aperiodic

We avoid periodicity if $P_{v,v} > 0$ for all v.

Ergodic means connected and aperiodic.

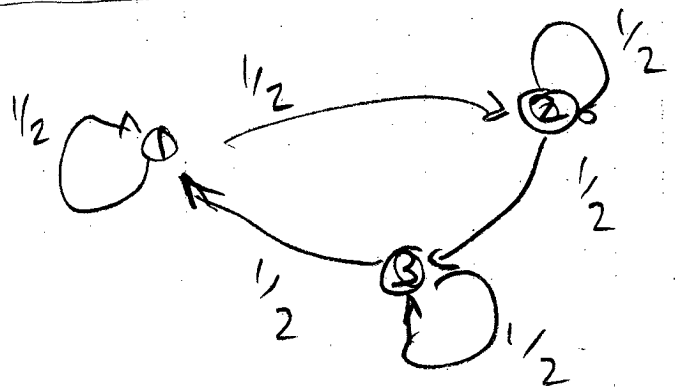
DEF

A Markov chain is reversible (let P be the transition matrix of an ergodic MC) if

$$\pi_i P_{ij} = \pi_j P_{ji} \quad (*)$$

↖ detailed balance condition

If P is ergodic and π satisfies the detailed balance condition, then π is a stationary distribution.



aperiodic, connected $\Rightarrow$ ergodic $\Rightarrow$ unique stationary distribution

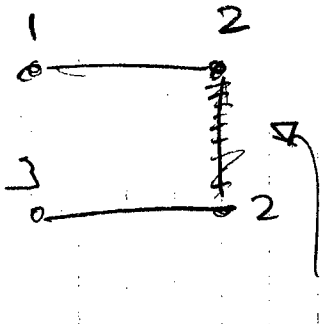
$\pi = (1/3, 1/3, 1/3)$

$\pi_1 = \pi_2 = 1/3$

$P_{12} = 1/2$, $P_{21} = 0$

$\therefore$ not reversible

Sample labelings of vertices of a graph by 3 colors



$P(\text{labeling}) \propto \lambda^{\#\text{ monochromatic edges}}$

$= \frac{1}{z} \lambda^{\#\text{ mono}\cdots}$

MC

Current state $X_n = C$

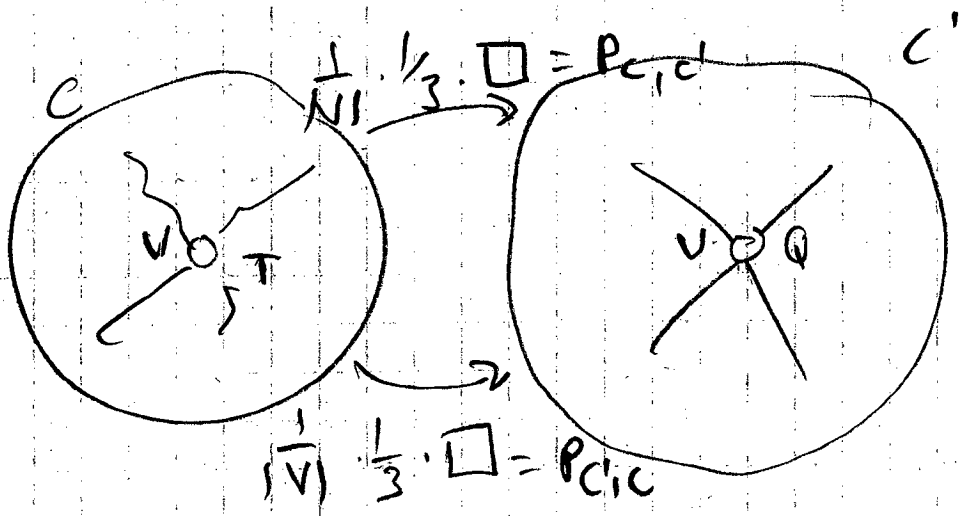
Pick a random vertex v of G. Pick a random color $\in \{1, 2, 3\}$.

Let C' be coloring C with color of v changed to Q.

With probability $\square$ let $X_{n+1} = C'$, otherwise

let $X_{n+1} = C$

Transitions in MC are differences in coloring of a single vertex, we want to satisfy the detailed balance condition



Goal $\pi_C P_{C,C'} = \pi_{C'} P_{C',C}$

The goal is

$$\frac{\pi_c}{\pi_{c'}} = \frac{p_{c'|c}}{p_{c|c'}}$$

$$\| \qquad \| \qquad \|$$

$$\lambda^{\#\text{mono in } c - \#\text{mono in } c'} = \frac{\square}{\square}$$

$1 \leq \lambda$

If c' has less monochromatic edges, then $\square = 1$, otherwise

$$\square = \lambda^{\#\text{mono in } c - \#\text{mono in } c'}$$

Read about Metropolis filter in the book. Don't need to read about extended version of Metropolis.

$$\square = \begin{cases} 1 & \text{if } \pi_{c'} > \pi_c \\ \dfrac{\pi_{c'}}{\pi_c} & \text{otherwise} \end{cases}$$

3.6.2009

Midterm Review

ML for multinomial distribution

$p_k = p(x=k)$, $k = 1, \cdots, K$

$$= \prod_{n=1}^{N} p(x_n|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} p_k^{I\{x_n=k\}}$$ (Bishop p. 75)

↑ indicator function is 1 if $x_n = k$, 0 otherwise

Task is to find p_k that maximizes probability given observations x_n

$$\ln L = \sum_{n=1}^{N} \sum_{k=1}^{K} I\{x_n=k\} \ln p_k - \lambda\left(\sum_{k=1}^{K} p_k - 1\right)$$

↑ Remember the constraint that $\sum p_k = 1$

$$\frac{\partial \ln L}{\partial p_k} = \sum_{n=1}^{K} \frac{I\{x_n=k\}}{p_k} - \lambda = 0$$

$$\sum_{n=1}^{N} I\{x_n=k\} = \lambda p_k$$

Sum both sides over all k yields

$$N = \lambda$$

$$p_k = \frac{\sum_{n=1}^{N} I\{x_n=k\}}{N} = \frac{N_k}{N}$$

Steps
1. Create an expression for seeing $x_1, \cdots, x_n$
2. Eliminate π by taking log
3. Add constraints
4. Optimize by setting $\frac{\partial}{\partial \theta}$ to 0

Markov Chain

$p^T = (p_1, \dots, p_N)$

$A \in \mathbb{R}^{N \times N}$

$p^T(t+1) = p^T(t) \cdot A$

Solve for $p^T = p^T A$ to find the stationary distribution.

EM Algorithm

Z hidden
X observed
Θ parameters

We want to find $\operatorname{argmax}_\Theta p(X|\Theta)$.

If this is too difficult, we look at $p(Z, X|\Theta)$.

E: $p(Z|X, \Theta^{old})$ ← this is the form of q

M: $\Theta^* = \operatorname{argmax}_\Theta \sum_Z p(Z|X, \Theta^{old}) \ln p(Z, X|\Theta)$

$$L(q, \Theta) = \sum_Z q(Z) \ln \frac{p(X, Z|\Theta)}{q(Z)}$$

Bernoulli EM

$$p(X) = \sum_{k=1}^{K} \pi_k p(X|\mu_k) = \prod_{k=1}^{K} \left[\pi_k p(X|\mu_k)\right]^{z_k}$$

$z_k = 1$ iff X is from the k^{th} Mixture, otherwise 0

$$p(x|\mu_k) = \mu_k^{x}(1-\mu_k)^{1-x}$$

(E:) $$\gamma(z_{nk}) = \frac{\pi_k\, p(x_n|\mu_k)}{\sum_j \pi_j\, p(x_n|\mu_j)}$$

n^{th} sample came from k^{th} class

(M:)

$$E_z[\ln p(z, x|\theta)] = E_z\left[\ln \prod_{n=1}^{N}\prod_{k=1}^{K} (\pi_k \cdot \mu_k^{x_n}\cdot(1-\mu_k)^{1-x_n})^{z_{nk}}\right]$$

$$= \sum_n \sum_k \gamma(z_{nk})\left(\ln \pi_k + x_n \ln \mu_k + (1-x_n)\ln(1-\mu_k)\right)$$

$$-\lambda\left(\sum \pi_k - 1\right)$$

$$\frac{\partial E_z}{\partial \pi_k} = 0 \quad \searrow$$

$$\sum_n \gamma(z_{nk}) = \lambda \pi_k \qquad \sum_k \sum_n \gamma(z_{nk}) = \lambda = N$$

$$\pi_k = \frac{N_k}{N}$$

For μ_k, we take

$$\frac{\partial E_z}{\partial \mu_k} = 0$$

we end up with $$\sum_n \frac{\gamma(z_{nk})\, x_n}{\mu_k} = \sum_n \frac{\gamma(z_{nk})(1-x_n)}{1-\mu_k}$$

$$\mu_k = \frac{\sum_n \gamma(z_{nk})\, x_n}{\sum_n \gamma(z_{nk})}$$

4.2.2009

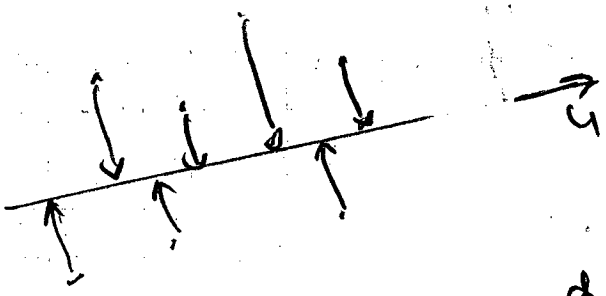
PCA

We have data points $x_1, \dots, x_n \in \mathbb{R}^d$ (d is "big")

We want to find points $y_1, \dots, y_n \in \mathbb{R}^M$ (M is "small") capturing as much variance of x_i as possible

No labels, unsupervised

Case $M=1$



We want to find $u \in \mathbb{R}^d$, $\|u\|_2 = 1$

$y_i = u^T x_i$

Maximize variance of $u^T x_1, \dots u^T x_n$, which is

$$\frac{1}{n} \sum_{i=1}^{n} (u^T x_i - u^T \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$V = \frac{1}{n} \sum_{i=1}^{n} u^T (x_i - \bar{x})(x_i - \bar{x})^T u$$

$$= u^T S u, \quad S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \in \mathbb{R}^{d \times d}$$

We want to maximize $u^T S u$ subject to $\|u\|_2 = 1$.
We use Lagrange:

$$L(\lambda, u) = u^T S u - \lambda (u^T u - 1)$$

$$\frac{\partial L}{\partial u} = 2Su - 2\lambda u = \phi$$

$$\boxed{Su = \lambda u}$$

u is an eigenvector of S

The value of $u^T S u = u^T \lambda u = \lambda$

The optimal choice of u is the eigenvector of S corresponding to the largest eigenvalue.

So for case $M=1$, we take $x_1, \dots, x_n$ and compute S

$$Su = \lambda u$$

$M=2$

u_1 : $Su_1 = \lambda_1 u_1$

u_2 : $Su_2 = \lambda_2 u_2$

the eigenvalues of S satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$

SVD

Let A be an $m \times n$ matrix. Then A can be written

$$A = U \Sigma V$$

where $U \in \mathbb{R}^{m \times m}$, $U^T U = I$

$V \in \mathbb{R}^{n \times n}$ $V^T V = I$ (U, V are orthogonal)

$\Sigma \in \mathbb{R}^{m \times n}$ diagonal matrix with ≥ 0 entries

The entries in Σ are called singular values of A.

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & 0 \\ & & \ddots & \\ & 0 & & \sigma_m \end{bmatrix}, \quad \sigma_1 \geq \sigma_2 > \cdots \geq \sigma_m$$

$$\Sigma_t = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & 0 & \ddots \\ & & & 0 \end{bmatrix}, \quad t < m$$

PCA

$x_1, \ldots, x_n \in \mathbb{R}^d$

Project to $\mathbb{R}^M$ while maximizing variance of the projection.

Facts About SVD

We have A and want to approximate it with a matrix of rank k

$$\min_{B,\, rk(B) \leq k} \|A - B\|_2$$

$$\min_{B,\, rk(B) \leq k} \|A - B\|_F$$

largest eigenvalue

$$\|A\|_2 := \max_{\|x\|_2 = 1} \|Ax\|_2 = \left(\rho(A^T A)\right)^{1/2} = \sigma_1(A)$$

$$\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2}$$

$$A = U \Sigma V$$

$$A - B := A - U \Sigma_k V$$

For $rk(A) = a$ and $rk(B) = b$, then

$$rk(AB) \leq \min\{a, b\}$$

$$X = \begin{bmatrix} X_1^T - \bar{X}^T \\ X_2^T - \bar{X}^T \\ \vdots \\ X_n^T - \bar{X}^T \end{bmatrix}$$

this is a row vector, making X a matrix $\in \mathbb{R}^{n \times d}$

↓

SVD of X

$$X = U \Sigma V$$

$$X' = U \Sigma_M V$$

$$S = \frac{1}{2} X^T X$$

$M = 1$ → largest eigenvalue of S and the corresponding eigenvector is the first column of V, $V \in \mathbb{R}^{d \times d}$

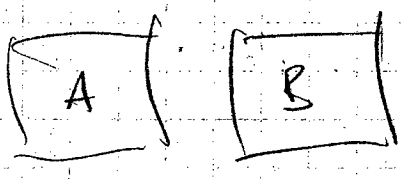
$$S = \frac{1}{n}(X^T X) = \frac{1}{n}(V^T \Sigma^T U^T U \Sigma V)$$

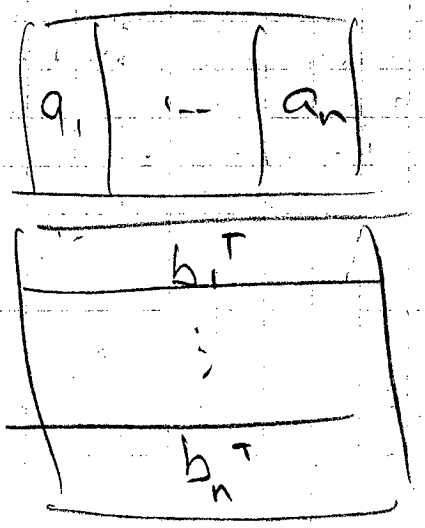
$$= \frac{1}{n}(V^T \Sigma^T \Sigma V)$$

$$\Sigma^T \Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_k^2 \end{bmatrix}$$

$X = U \Sigma V$

Use the first M columns of V for the projection.





$(a_i b_i)^T_{u,v} =$

$(a_i)_u (b_i)_v$

$$\sum_{i=1}^{n} (a_i b_i^T)_{u,v}$$

$$= \sum_{i=1}^{n} (a_i)_u (b_i)_v =$$

$A_{u,i} \; B_{i,v}$

$$X = \begin{pmatrix} x_1^T - \bar{x}^T \\ \vdots \\ x_n - \bar{x}^T \end{pmatrix}$$

$n \times d$ matrix

goal SVD of X

- eigenvalues / eigenvectors of XX^T $n \times n$
- eigenvalues / eigenvectors of $X^T X$ $d \times d$

These two matrices have the same set of eigenvalues.

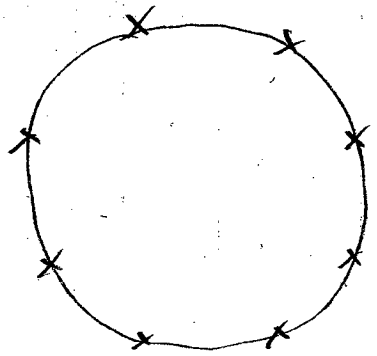
4.7.2009

Manifolds

Goal: Find a NON-LINEAR low dimensional structure in data

(e.g. PCA was extracting a LINEAR subspace of dim M which contained most "information")

"We want to figure out low dimensional structures which are not necessarily linear."



data is lying on a circle

Imagine data lying on a bowl in 3-space or on the surface of a rolled-up piece of paper.

Assumption: The data comes from a low-dimensional manifold (embedded in a high-dimensional space)

Isomap

(1) Look at data points $X_1, \dots, X_n \in \mathbb{R}^d$

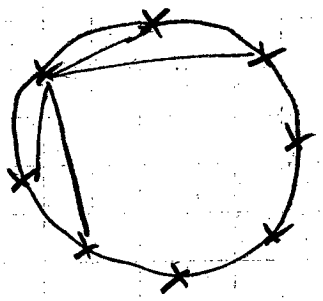
Create a graph. Vertices correspond to data points

Vertices i and j will be connected if

$$\| X_i - X_j \|_2 \leq \varepsilon$$

Alternatively, connect i to $j_1, \dots, j_k$ where $X_{j_1}, \dots X_{j_k}$ are k-NN of X_i

e.g.



We have a graph. Lets add edge weights

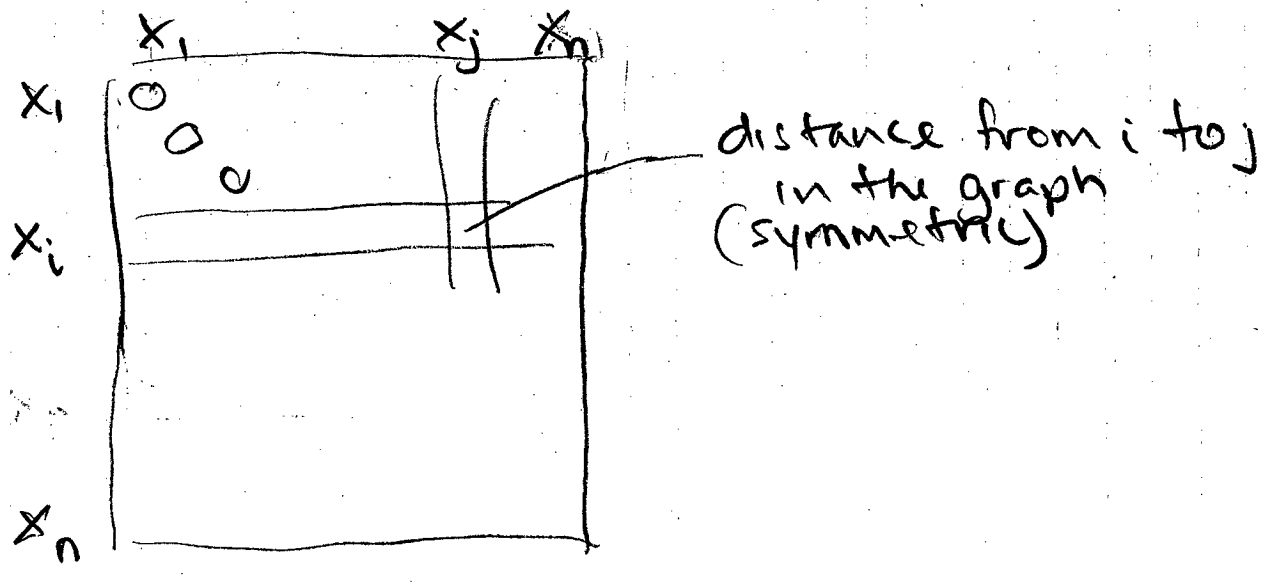
$$w_{ij} = \|x_i - x_j\|_2$$

The algorithm works by finding geodesics on the manifold.

(2) "Recover" approximate <u>geodesics</u>" (shortest curve between points)

Compute shortest path between every pair of vertices.

(3) Use the distances to embed the data into the low-dimensional Euclidean space



We more or less preserve the distances.

MDS (Multidimensional Scaling)

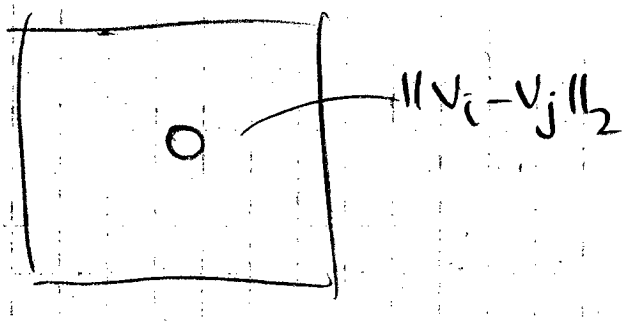
Given distance matrix, find a good embedding in $\mathbb{R}^M$

Given distance matrix
↓
Find matrix of inner products
↓
SVD, Take top M eigenVectors
↓
Find good embedding in $\mathbb{R}^M$

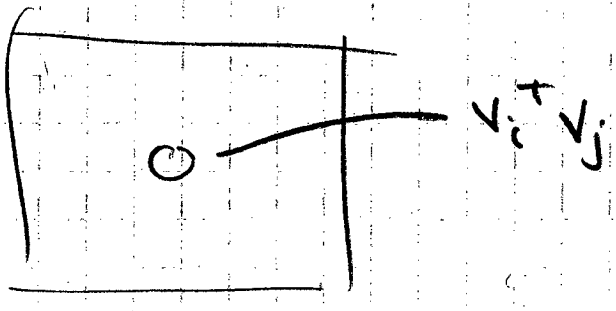
There are unknown $v_1, \ldots, v_n$ in $\mathbb{R}^n$

$dist(v_i, v_j) = \|v_i - v_j\|_2$

We are given



We want



A

$$\begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & 0 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 0 \end{bmatrix}$$

→

C

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Given matrix of distances, we want to calculate matrix of inner products assuming

$$\sum_{i=1}^{n} v_i = 0$$

i.e. The points are centered around a mean.

A. Given $\|v_i - v_j\|_2 \ (\forall i,j)$

B. Find $v_i^T v_j \ (\forall i,j)$

$$B_{ij} = A_{ij}^2 = \|v_i - v_j\|_2^2$$

Given C, $B_{ij} = c_{ii} + c_{jj} - 2c_{ij} = (v_i - v_j)^T (v_i - v_j)$

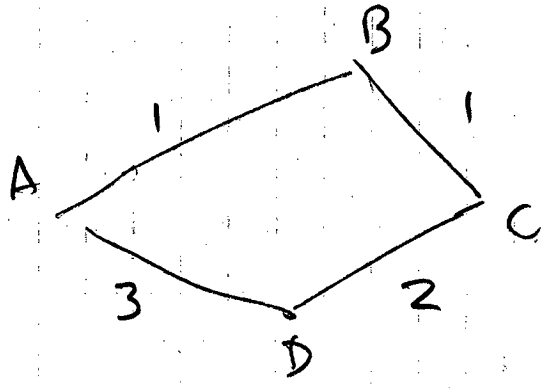
How do we compute c from A?

$$C = -\frac{1}{2} HBH \quad \text{where} \quad H = I - \frac{1}{n} J, \quad J = 1\,1^T$$

$$J = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Laplacian Eigenmaps

DEF
Given a graph $G = (V, E)$ with weighted edges (NON-NEGATIVE)
(Let W_{ij} = weight of edge $\{i,j\}$, 0 if not an edge)



$$W = \begin{array}{c|cccc} & A & B & C & D \\ A & 0 & 1 & 0 & 3 \\ B & 1 & 0 & 1 & 0 \\ C & 0 & 1 & 0 & 2 \\ D & 3 & 0 & 2 & 0 \end{array} \qquad D = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

Let $D = \text{diagonal}\left(\sum_{j=1}^{n} W_{ij}\right)$

Laplacian matrix of G is $L = D - W$

$$L = \begin{bmatrix} 4 & -1 & 0 & -3 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 3 & -2 \\ -3 & 0 & -2 & 5 \end{bmatrix}$$

Fact

$$0 \leq \sum_{\{i,j\} \in E} W_{ij}(y_i - y_j)^2 = 2y^T L y \Rightarrow$$ L is positive semidefinite

$$\sum_{i,j} w_{ij}(y_i - y_j)^2 = \sum_{i,j} w_{ij} y_i^2 + \sum_{i,j} w_{ij} y_j^2 - 2\sum_{i,j} w_{ij} y_i y_j$$

$$= \sum_i D_i y_i^2 + \sum_j D_j y_j^2 - 2\sum_{i,j} w_{ij} y_i y_j$$

$$2y^T L y = 2\sum_{i,j} y_i y_j L_{ij}$$

$$= -2\sum_{(\neq j)} y_i y_j w_{ij} + 2\sum_{(=j)} y_i^2 D_i$$

---

take data points $x_1, \ldots, x_n$.

(1) Create graph (either k-NN or $\leq \varepsilon$ distance)

$w_{ij} = 1$ if i,j is connected
ϕ otherwise

or

$w_{ij} = e^{-\|x_i - x_j\|_2^2/t}$ if i,j is connected
ϕ otherwise

(2) Compute L, the Laplacian of G

(3) This will be ϕ

~~$Lv_1 = \lambda_1 D v_1$~~

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{M+1}$$

$M+1$ smallest eigenval. of L

$$Lv_{M+1} = \lambda_{M+1} D v_{M+1}$$

$$x \mapsto (x^T v_2, \ldots, x^T v_{M+1})$$

$$0 \leq \sum_{\{i,j\} \in E} w_{ij} (y_i - y_j)^2 = 2 y^T L y$$

We want to enforce $y^T 1 = 0$

$y^T D y = 1$

Hidden Markov Models (HMMs)

Markov Chain (time homogeneous)

Sequence of random variables $X_1, \cdots, X_n, \cdots$

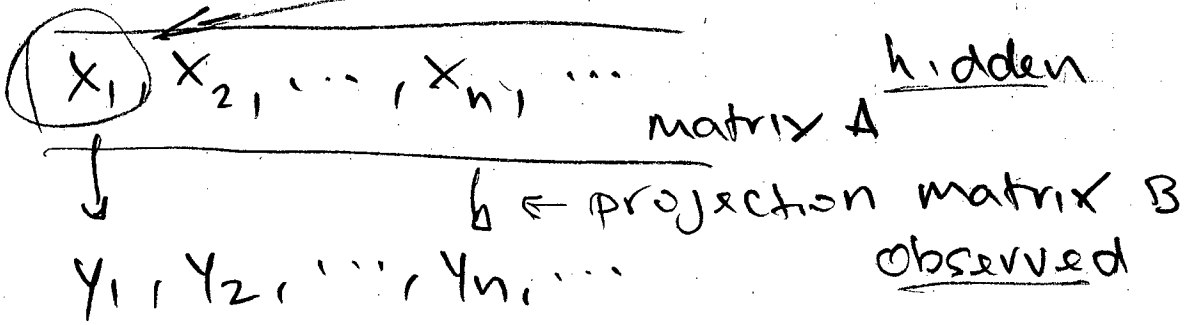
$P(X_{n+1} = a \mid X_n = b, X_{n-1} \cdots, X_1) = P(X_{n+1} = a \mid X_n = b) = A_{\cdot}$

Time homogeneous means that it is not dependent on n

$X_i \in \Omega$

$A \in \mathbb{R}^{|\Omega| \times |\Omega|}$

Need to know
π = distribution of X_1



$P(Y_i = c \mid X_i = d) = B_{c,d} = P(Y_i = c \mid X_i = d, X_1, \cdots, X_{i-1}$

An HMM is given by (M hidden states, N observable states

π distribution on $[M]$

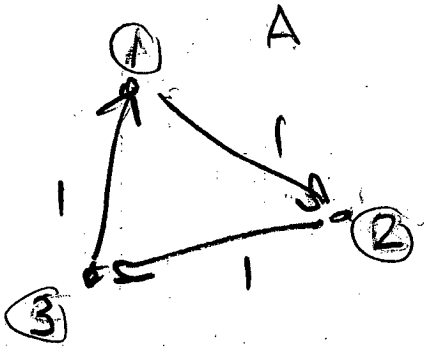
$A \in \mathbb{R}^{M \times M}$

$B \in \mathbb{R}^{M \times N}$

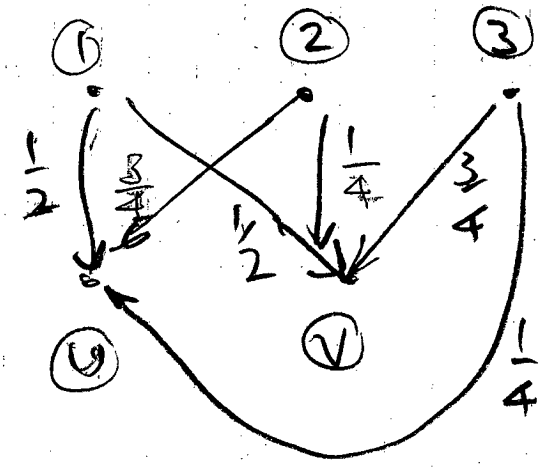
Ex.

$\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

A



B



hidden

visible

You see a sequence of U and V, you want to discover the sequence of hidden states

---

DECODING (ML)

1) Given A, B, π, and $Y_1, \cdots, Y_T$

find $X_1, \cdots, X_T$ which maximizes the likelihood.

DECODING (MAP)

2) Given $A, B, \pi, Y_1, \cdots, Y_T$

Find posterior for X_i

EVALUATION

3. Given $A, B, \pi, Y_1, \ldots, Y_T$

what is the probability of this sequence?

LEARNING

4. Given $Y_1^{(1)}, \ldots, Y_T^{(1)}$

$\vdots$

$Y_1^{(k)}, \ldots, Y_T^{(k)}$

Find A, B, π which maximize the likelihood.

Decoding

1) Viterbi's Algorithm

- dynamic programming

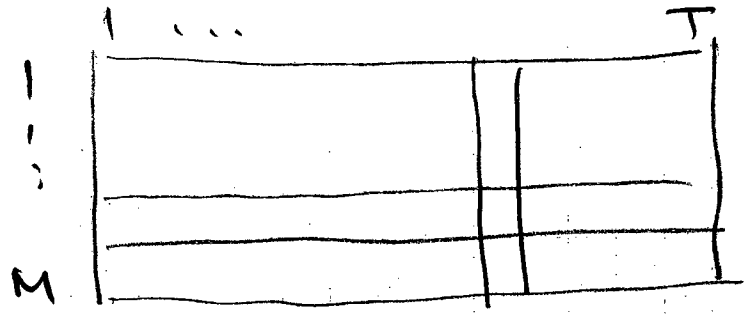


Table D

D_{ij} = maximum probability of a sequence $X_1, \ldots, X_j$ with $X_j = i$

Take all sequences of internal states of length j which end with i. Take the one with the highest probability. D_{ij} is its probability.

$$\text{Prob}(X_1, \cdots, X_T \mid y_1, \cdots, y_T) =$$

$$\pi(X_1)\left(\prod_{i=1}^{T-1} A_{X_i, X_{i+1}}\right)\left(\prod_{i=1}^{T} B_{X_i, y_i}\right)$$

We calculate D_{ij} by first determining

$$D_{k,j-1} \quad \text{for } k \in [M]$$

$$D_{iN} = \max_{k \in [M]} D_{k,j-1} \cdot A_{k,i} \cdot B_{i, y_j}$$

The first column of the table is initialized with

$$\pi_i B_{i, y_1}$$

The running time is $(MT)\, M = M^2 T$.

(MT: # of entries; M: time per entry)

Evaluation

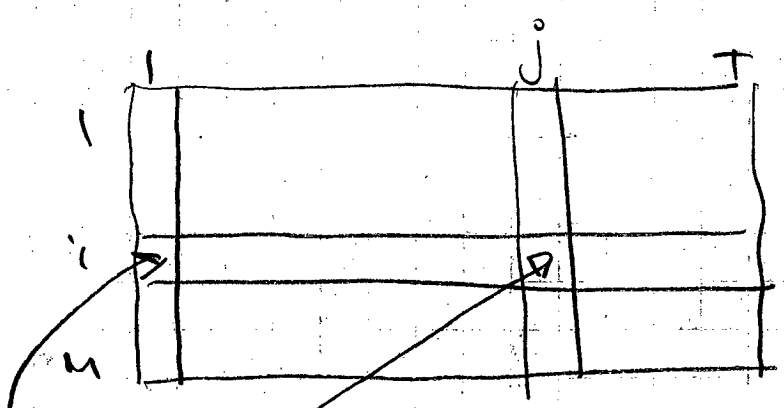
3) Dynamic programming

We want to compute

$$P(y_1, \cdots, y_T \mid A, B, \pi) =$$

$$\sum_{X_1, \cdots, X_T = 1}^{M} P(y_1, \cdots, y_T, X_1, \cdots, X_T \mid A, B, \pi)$$

This is actually T sums.



$$D_{i,j} = P(Y_1, \cdots, Y_j, X_j = i \mid A, B, \pi)$$

$$D_{i,1} = \pi_i B_{i,Y_1} =$$

The recurrence relation is

$$\boxed{D_{i,j} = \sum_{k=1}^{M} D_{k,j-1} \cdot A_{k,i} \cdot B_{i,Y_j}}$$

(Note that k is the value to which we condition X_{j-1}.)

$$P(Y_1, \cdots, Y_j \mid X_j = i) \stackrel{?}{=} \sum_{k=1}^{M} P(Y_1, \cdots, Y_{j-1} \mid X_j = k) A_{k,i} B_{i,}$$

↗ is this true

$$P(Y_1, \cdots, Y_j \mid X_j = i) = \frac{P(Y_1, \cdots, Y_j, X_j = i)}{P(X_j = i)}$$

$$\sum_{k=1}^{M} (\text{from above}) = \sum_{k=1}^{M} \frac{P(Y_1, \cdots, Y_{j-1}, X_{j-1} = k)}{P(X_{j-1} = k)} \cdot A_{k,i} \cdot B_{i,}$$

We got <u>confused</u>, so we changed to

$$D_{i,j} = P(Y_1, \cdots, Y_j, X_j = i \mid A, B, \pi)$$

$$D_{i,1} = \pi_i B_{i,Y_1}$$

$$P(Y_1, \cdots, Y_j, X_j = i) = \sum_{k=1}^{M} P(Y_1, \cdots, Y_{j-1}, X_{j-1} = k) A_{k,i} B_{i,Y_j}$$

The answer is the sum of the last (T) column:

$$P(Y_1, \cdots, Y_T \mid A, B, \pi) = \sum_{i=1}^{M} D_{i,T}$$

---

2)

$$P(X_i \overset{=q}{|} Y_1, \cdots, Y_T, A, B, \pi) = ?$$

$$P(X_i = q \mid Y_1, \cdots, Y_t, A, B, \pi)$$

$$P(Y_{t+1}, \cdots, Y_T \mid X_i = q, A, B, \pi)$$

$$= \frac{P(Y_1, \cdots, Y_t \mid X_i = q)\ P(X_i = q)}{P(Y_1, \cdots, Y_T)}$$

leaving A, B, π out for brevity

4.14.2009

HMM

transition matrix $A \in \mathbb{R}^{M \times M}$
initial probability $\pi \in \mathbb{R}^M$
emission matrix $B \in \mathbb{R}^{M \times N}$

$X_i \sim \pi$ $\quad P(X_{i+1} = a \mid X_i = b, \cdots, X_1) = A_{a,b}$

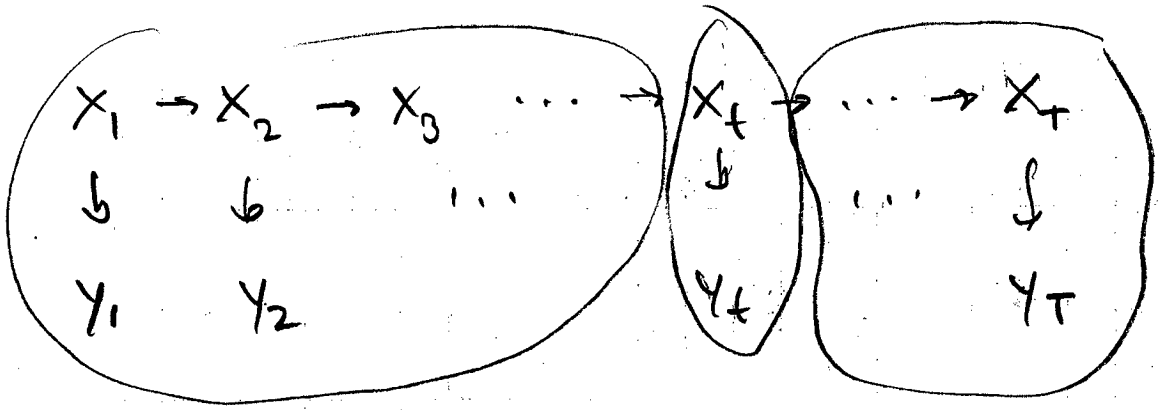
$P(Y_i = c \mid X_i = d) = B_{c,d}$

p2 Given $Y_1, \cdots, Y_T$, what is the likelihood of X_t

$P(X_t \mid Y_1, \cdots, Y_T)$

We know A, B, π.

$$P(X_t \mid Y_1, \cdots, Y_T) = \frac{P(Y_1, \cdots, Y_T \mid X_t) P(X_t)}{P(Y_1, \cdots, Y_T)} = *$$



$$P(Y_1, \cdots, Y_T \mid X_t) = P(Y_1, \cdots, Y_t \mid X_t) P(Y_{t+1}, \cdots, Y_T \mid X_t)$$

$$* = \frac{P(Y_1, \cdots, Y_t, X_t) P(Y_{t+1}, \cdots, Y_T \mid X_t)}{P(Y_1, \cdots, Y_T)}$$

$P(Y_1, \cdots, Y_t, X_t)$: $\alpha_t(a)$ $\quad$ $P(Y_{t+1}, \cdots, Y_T \mid X_t)$: $\beta_t(a)$

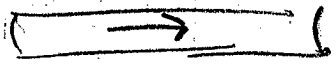
Let $X_t = a$

$$P(Y_1, \cdots, Y_T) = \sum_{a \in [M]} \alpha_t(a)\, \beta_t(a)$$

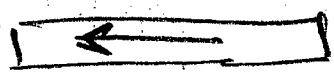
$\alpha_t(a)$ = probability of seeing $Y_1, \cdots, Y_t$ and ending up in $X_t = a$

$\beta_t(a)$ = probability of seeing $Y_{t+1}, \cdots, Y_T$ given that $X_t = a$

$$\alpha_{t+1}(a') = \sum_{a \in [M]} \alpha_t(a)\, A_{a,a'}\, B_{a', Y_{t+1}}$$



$$\beta_t(a) = \sum_{a' \in [M]} \beta_{t+1}(a')\, A_{a,a'}\, B_{a, Y_t}$$



$$\alpha_1(a) = \pi_a\, B_{a, Y_1}$$

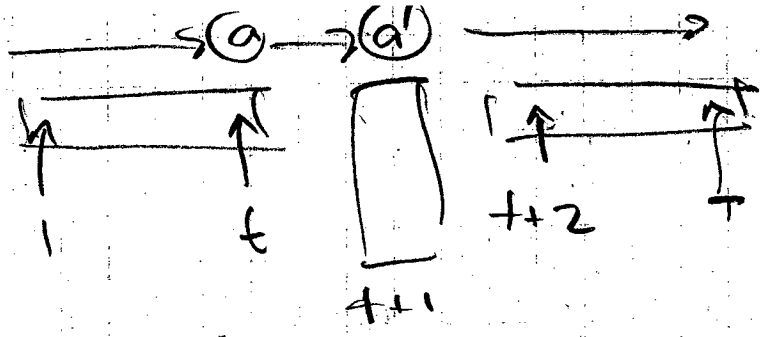
$$\beta_{T-1}(a) = B_{a, Y_T} \qquad \beta_T(a) = 1$$

use recurrence, we get (as a check)

$$\beta_{T-1}(a) = \sum_{a' \in [M]} \beta_T(a')\, A_{a,a'}\, B_{a, Y_T} = B_{a, Y_T}$$

$$P(X_t = a, X_{t+1} = a' \mid Y_1, \cdots, Y_T) = \alpha_t(a)\, \beta_{t+1}(a')$$

$$\alpha_t(a)\, \beta_{t+1}(a')\, A_{a,a'}\, B_{a', Y_{t+1}}$$



Q4 Learning

We get $Y_1^{(1)}, \dots, Y_T^{(1)}$

$Y_1^{(K)}, \dots, Y_T^{(K)}$

We want to learn A, B, π.

We are going to use the EM algorithm.

Initialize A, B, π.
Update A, B, π using "inferred internal state"

Compute

$$P(X_t = a, X_{t+1} = a' \mid Y_1^{(k)}, \dots, Y_T^{(k)}, A, B, \pi), \quad k = 1, \dots$$

$$\sum_{k=1}^{K} \sum_{t=1}^{T-1} P(\dots) = A'_{a,a'}$$

Then normalize rows of $A'_{a,a'}$ so that they sum to one.

$$\sum_{\lambda=1}^{K} \sum_{t;Y_t=b} P(X_t = a \mid Y_1^{(k)}, \dots, Y_T^{(k)}, A, B, \pi) = B'_{a,b}$$

Then normalize rows of $B'_{a,b}$ to sum to one

; means "such that"

$$\pi'(a) \propto \sum_{l=1}^{k} P(x_1 = a \mid y_1^{(l)}, \ldots, y_{T_l}^{(l)}, A, B, \pi)$$

What does the likelihood function for this EM problem look like?

$$P(x_1, \ldots, x_T, y_1, \ldots, y_T) =$$

$$\pi_{x_1} \prod_{t=1}^{T-1} A_{x_t, x_{t+1}} \prod_{t=1}^{T} B_{x_t, y_t}$$

Markov Decision Process (MDP) $= (S, A, P_a, R, \gamma)$

S = set of states
A = actions
$P_a(s, s')$ = probability of going from s to s' (with action a)
$R: S \to \mathbb{R}$ = reward
γ = discount factor (inflation), $\gamma < 1$

Given S, A, P_a, R, γ, the goal is to figure out what to do?

Determine a policy $\pi: S \to A$

good

Initial state is $z \in S$

$x_1 = z, x_2, \ldots, x_t, \ldots$

$P(x_{t+1} = s' \mid x_t = s) = P_{\pi_s}(s, s')$

Reward in instantiation $X_1, \ldots X_{t+1}, \ldots$ is

$$R(X_1) + \gamma R(X_2) + \gamma^2 R(X_3) + \ldots = \sum_{t=0}^{\infty} R(X_{t+1}) \gamma^t$$

We want to maximize

$$E\left[\sum_{t=0}^{\infty} R(X_{t+1}) \gamma^t\right] \quad \text{by picking } \pi$$

4.15.2009

From HW # 4.3

If x and y are random variables, and $z = x + y$ then the distribution of z is the convolution of the distribution of x by the distribution of y.

$$p(z=a) = \sum_k p(x=k)\, p(y=a-k)$$

this is a convolution

VC-dimension of a circle is 3. Prove by setting 2 most distant points to '+', others to '−'.

Prove that $\exists d \ \Pi_c(m) = O(m^d) \Rightarrow VC(C) < \infty$.

If $VC(C) = \infty$ then $\forall m \ \Pi_c(m) = 2^m \Rightarrow \not\exists d \ \forall m \ 2^m = O(m^d)$
(Proof by contradiction)

---

1. Prove that, given a set of n i.i.d. observations $x_1, \ldots, x_n$ a distribution Q that minimizes the KL-divergence with empirical distribution P, also maximizes the probability of the data $Q(x_1, \ldots, x_n)$

2. Derive (w/o using VC-dim) a PAC-learning bound for the concept class of axis-aligned SQUARES.

4.16.2009

Markov Rewards Process

$S, \cancel{A}, R, P, \gamma$

no actions

Start in state s, make transition according to P, repeat, in each state get discounted reward.

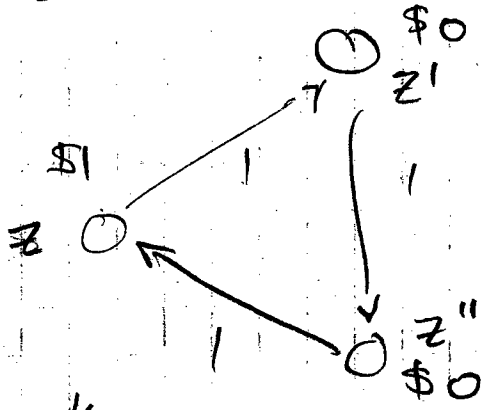
$X_0, X_1, \ldots, X_n, \ldots$

$X_0 = z$

$P(X_{i+1} = a \mid X_i = b, \ldots, X_0) = P_{a,b}$

$E\left[\sum_{i=0}^{\infty} \gamma^i R(X_i)\right]$ = expected total reward

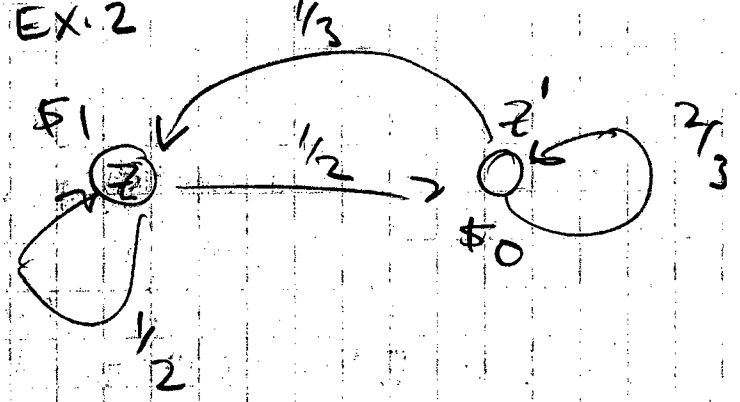
EX. 1



$\gamma = 0.99$

$\frac{1}{1-(0.99)^3}$

EX. 2



V_s = expected reward at s

$= \left(\gamma \sum_{s' \in S} P(s,s') V_{s'}\right) + R(s)$

$(\forall s)$

For previous Ex. 1:

$$\left.\begin{aligned} V_z &= 1 + \gamma V_{z'} \\ V_{z'} &= \gamma V_{z''} \\ V_{z''} &= \gamma V_z \end{aligned}\right\} \quad V_z = 1 + \gamma^3 V_z$$

For Ex. 2:

$$V_z = 1 + \left(\tfrac{1}{2} V_z + \tfrac{1}{2} V_{z'}\right)\gamma$$

$$V_{z'} = \left(\tfrac{2}{3} V_{z'} + \tfrac{1}{3} V_z\right)\gamma$$

$(V_1, \dots, V_n)$ = vector of expected utilities $= V^T$

The matrix form for the equation for V_s is

$$V^T = \gamma V^T P + R^T$$

where $R^T = (R(1), \dots, R(n))$.

$$V^T(1 - \gamma P) = R^T$$

$$V^T = \underbrace{(1 - \gamma P)^{-1}}_{\text{is this always invertible}} R^T$$

P satisfies $P_{ij} \geq 0$, $\sum_{j=1}^{n} P_{ij} = 1$. P is a stochastic matr

$S(P) = 1$ (absolute value of)

spectral radius = largest eigenvalue of P

Largest eigenvalue of $\gamma P = \gamma < 1$

$AV_1 = \lambda_1 V_1$

$(I - A)V_1 = V_1 - \lambda V_1 = (1-\lambda)V_1$

A eigenvalues $\lambda_1, \ldots, \lambda_n$

$I - A$ eigenvalues $1-\lambda_1, \ldots, 1-\lambda_n$

Since $0 < \lambda_i < 1$, $0 < 1-\lambda_i < 1$.
Therefore A is not singular

Markov Decision Process

$S, A, \{P_a, a \in A\}, R, \gamma$

We study policies (functions $S \to A$).

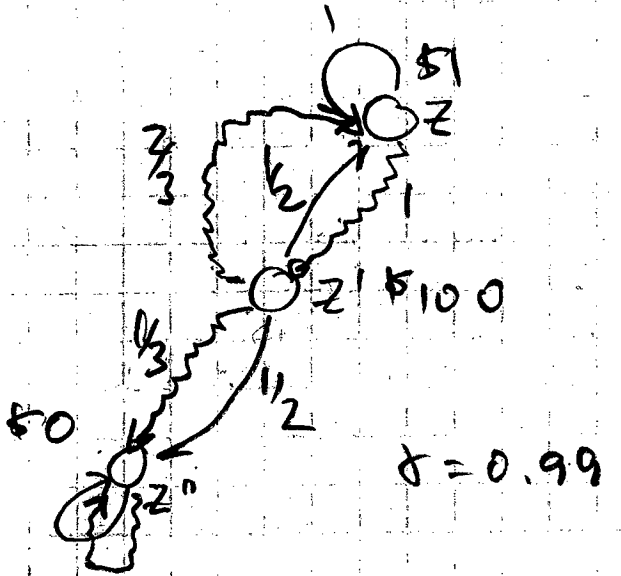
Given a policy π, the utility is

$$E\left[\sum_{i=0}^{\infty} \gamma^i R(X_i)\right]$$

where

$X_0, X_1, \ldots, X_n, \ldots$

$$P(X_{i+1} = s' \mid X_i = s, \ldots) = P_{\pi(s)}(s, s')$$



policy a ——

policy b ~~~

There always exists an optimal policy.

How do we find optimal policies?

What is the utility of state s under optimal policy π^*?

$$V_s = R(s) + \gamma \left(\max_a \sum_{s' \in S} P_a(s,s') V_{s'} \right)$$

(Bellman's equations)

thm Bellman's equations have a unique solution.

Value Iteration

Start with some V's, e.g. $V^{(0)}(s) = 0 \quad (\forall s \in S)$

Repeat

$$V^{(t+1)}(s) = R(s) + \gamma \max_a \sum_{s' \in S} P_a(s,s') V^{(t)}(s')$$

We hope this converges.

V_1 $\quad$ V_2

$\downarrow$ value iteration $\quad$ $\downarrow$ value iteration

V_1' $\quad$ V_2'

$$\|V_1 - V_2\|_\infty \leq \Delta$$

i.e.

$$\forall s \quad |V_1(s) - V_2(s)| \leq \Delta$$

$$\|V_1' - V_2'\|_\infty \leq ? \quad \Delta\gamma$$

$$V_1'(s) - V_2'(s) = \gamma \max_a \sum_{s' \in S} P_a(s,s') V_1(s') - \gamma \max_{a'} \sum_{s' \in S} P_{a'}(s,s') V_2(s')$$

$$\leq \gamma \max_a \sum_{s' \in S} P_a(s,s') \left(V_1(s') - V_2(s') \right) \leq \Delta\gamma$$

We get this by choosing suboptimal $a' = a$

Banach's Contraction Theorem

$F: B \to B$

$\|f(x) - f(y)\| \leq \gamma \|x - y\| \Rightarrow$ unique fixed point

If a mapping is contracting, it has a unique fixed point.

V^* = fixpoint (utilities for optimal policy)

$\| V^{(0)} - V^* \|_\infty \leq ?$

$\| V^{(T)} - V^* \|_\infty \leq 2\gamma^T$

$$V_s = R(s) + \gamma \max_{a \in A} \sum_{s'} P_a(s, s') V_{s'}$$

Find the solution of the above equation.

$$V_s \geq R(s) + \gamma \sum_{s'} P_a(s, s') V_{s'} \quad (\text{for all } a \in A, s \in S)$$

How do we achieve equality? We add an objective to push V_s as low as possible

Linear Program

$$\min \sum_{s \in S} V_s$$

We claim that the optimal solution of the linear program is the unique solution to Bellman's equation

Policy Iteration

Start with policy $\pi^{(0)}$

Take the Markov Rewards Process for $\pi^{(t)}$
Compute its utilities $V^{\pi^{(t)}}$
(solving a system of equations)

Greedily update the policy.

$$\pi^{(t+1)}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P_a(s,s') V^{\pi^{(t)}}(s')$$

Take action a from state s and then follow $\pi^{(t)}$

You could put $R(s)$ and γ in here, but it doesn't effect the result.

Thm 1) Policy improvement theorem

$$V^{\pi^{(t+1)}}(s) \geq V^{\pi^{(t)}}(s)$$

Thm 2) One shot improvement theorem

If $\pi^{(t)}$ is not optimal, then

$$\pi^{(t+1)} \neq \pi^{(t)}$$

4.21.2009

MDP

$S, A, P(s'|s,a), R, \gamma$

policy $\pi : S \to A$

$X_0, X_1, \dots, X_t \qquad X_0 = s$

$P(X_{t+1} = s' \mid X_t = s) = P(s'|s, \pi(s))$

We want to maximize $E\left[\sum_{t=0}^{\infty} \gamma^t R(X_t)\right]$

Last time, we did value iteration:

$V_0 \xrightarrow{T} V_1 \xrightarrow{T} \dots \xrightarrow{T} V^*$

$$(TV)(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|a,s)\, V(s')$$

T is a contraction.

$\|TV_1 - TV_2\|_\infty \le \gamma \|V_1 - V_2\|_\infty$

take a policy π. Define $T_\pi : \mathbb{R}^S \to \mathbb{R}^S$

$$(T_\pi V)(s) = R(s) + \gamma \sum_{s' \in S} P(s'|s, \pi(s))\, V(s')$$

Is T_π a contraction?

$\|V_1 - V_2\|_\infty \le \Delta$

$$TV_1(s) - TV_2(s) = \gamma \sum_{s'} P(s'|s,\pi(s))\left(V_1(s) - V_2(s)\right)$$
$$\le \gamma \sum_{s'} P(s'|s,\pi(s))\,\Delta = \Delta\gamma$$

Yes. It is a contraction. What is the fixed point?

$V_0 \xrightarrow{T_\pi} V_1 \xrightarrow{T_\pi} \cdots \xrightarrow{T_\pi} V^\pi$, $V_\pi(s) = E\left[\begin{matrix}\text{reward}\\ \text{under}\\ \text{policy } \pi\end{matrix} \mid X_0 = s\right]$

Markov Rewards Process is the same as MDP with a fixed policy. So, above is iterative solution to linear equations from MRP

DEF: π is optimal if $V^\pi = V^*$
Lemma: π is optimal if $T_\pi V^* = V^*$

Policy iteration

$\pi_0 \to \pi_1 \to \cdots \to$

↓ compute V^{π_0} (fixed point of T_{π_0})

$\pi_1(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s,a) V^{\pi_0}(s')$

Lemma

$V^{\pi_{t+1}} \geq V^{\pi_t}$ (i.e. $(\forall s \in S)\ V^{\pi_{t+1}}(s) \geq V^{\pi_t}(s)$)

For value iteration, running time is dependent on γ. For policy iteration, running time is not dependent on γ.

$V^{\pi_{t+1}}, \ldots \pi_{t+1}, \pi_t, \ldots, \pi_1, \ldots$

⇓

$V^{k \times \pi_{t+1} + \infty \times \pi_t}$

Apply π_{t+1} for first k steps, then π_t forever.

Proof by induction $V^{k \times \pi_{t+1}, \pi_t} \geq V^{\pi_t}$ (t is fixed)

Base case: $k = 0$ $V^{\pi_t} \geq V^{\pi_t}$

$$V^{\pi_{t+1},\pi_t}(s) = \max_{a \in A} \sum_{s'} P(s'|a,s) V^{\pi_t}(s')$$

$$\geq \sum_{s'} P(s'|\pi_t(s),s) V^{\pi_t}(s')$$

$$V^{\pi_{t+2},\pi_{t+1},\pi_t}(s) =$$

$$\sum_{s'} P(s'|\pi_{t+1}(s),s) V^{\pi_{t+1},\pi_t}(s') \geq$$

$$\sum_{s'} P(s'|\pi_{t+1}(s),s) V^{\pi_t}(s') =$$

$$\max_a \sum_{s'} P(s'|a,s) V^{\pi_t}(s') \geq \sum_{s'} P(s'|\pi_t(s),s) V^{\pi_t}(s')$$

$$= V^{\pi_t}(s)$$

For induction, replace 2 by k+1 ...

So, now, we know that

$$V_{\pi_0} \leq V_{\pi_1} \leq \cdots \leq V_{\pi_t} \leq V_{\pi_{t+1}} \leq \cdots$$

There are at most $|S|^{|A|}$ different policies (different valuation vectors)

There exists $t \leq |S|^{|A|}$, such that $V_{\pi_t} = V_{\pi_{t+1}}$

At this point, policy iteration stops and declares that π_t is the optimal policy.

How do we prove this?

$T_{\pi_{t+1}} V_{\pi_t} = T V_{\pi_t}$ — This is the definition of $T_{\pi_{t+1}}$

(picks the optimal greedy action for any t

$$T_{\pi_{t+1}} V_{\pi_t} = TV_{\pi_t}$$

$$\| \qquad\qquad \|$$

$$T_{\pi_{t+1}} V_{\pi_{t+1}} \qquad TV_{\pi_{t+1}}$$

$$\|$$

$$V_{\pi_{t+1}} \quad \text{(since } V_{\pi_{t+1}} \text{ is the fixed point of } T_{\pi_{t+1}}\text{)}$$

$$V_{\pi_{t+1}} = TV_{\pi_{t+1}} \Rightarrow V_{\pi_{t+1}} = V^*$$

What do you do if you don't know what the rewards are in advance?

We don't know $P(s|s',a)$, $R(s)$.

There is an underlying MDP with S, A, P, R, γ. We know S, A, γ.

Q-Learning

$$Q(s,a) = R(s) + \gamma \sum_{s'} P(s'|s,a)\, V(s')$$

If we knew Q, what is the optimal π?

$$V(s) = \max_a Q(s,a)$$

$$\pi(s) = \arg\max_a Q(s,a)$$

α_t decreases with time (as in simulated annealing)

Goal: Figure out Q.

$$\hat{Q}(s,a) \leftarrow (1-\alpha_t)\hat{Q}(s,a) + \alpha_t\left(R(s) + \gamma \max_{a'} Q(s',a')\right)$$

In state s, we pick action a and end up in state s'

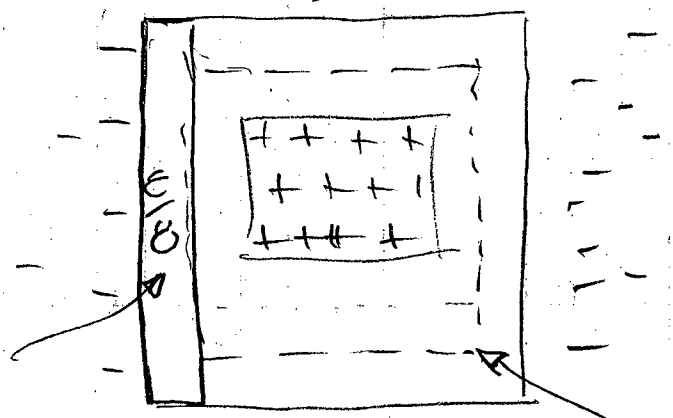
Thm. If $\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$, and every state visited infinitely often, then $\hat{Q} \to Q$

ε random action

1−ε action $a = \text{argmax}\ q(s,a)$

4.22.2009

PAC learning for squares from 4.15
Must use negatives as well as positives



(stripes)
Eight regions with ε/8 error

error for this stripe

--- actual square

KL-divergence from 4.15

$$D(p\|q) = \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log p(x) - \sum p(x) \log q(x)$$

$$p(x) = \frac{n_x}{N} \quad / \quad \sum_x n_x = N$$

$$= \frac{\sum_i \mathbb{I}\{x_i = x\}}{N}$$

$$D(p\|q) = \text{const} - \sum \frac{n_x}{N} \log q(x)$$

$$\prod_i q(x_i) = \prod_x q^{n_x}(x)$$

$$\underset{q}{\text{argmax}} \prod_i q(x_i) = \underset{q}{\text{argmax}} \prod_x q^{n_x}(x) = \underset{q}{\text{argmax}} \sum \ln q^{n_x}(x)$$

$$= \underset{q}{\text{argmax}} \sum n_x \ln q(x)$$

4.23.2009

EM Algorithm (from Midterm problem)

E-step: "empirical assignment of data points to clusters"

$P(X_1 | z=1) \propto \frac{1}{10}$

$P(X_1 | z=2) \propto \frac{1}{11}$ ...

$P(X_1 | z=3) \propto \frac{1}{12}$

M-step: "setting α_i and β_i to maximize 'z'"

$$L = \sum_{X,z} P(z=z | X=x, \theta) \log P(z=z, X=x | \theta^*)$$

(E-step over X, z; M-step: Find θ^* maximizing L)

$$\log P(z, X | \theta^*) = \begin{cases} -\infty & \text{if } \alpha_z < x \\ \log \frac{\beta_z}{\alpha_z} & \text{if } \alpha_z \geq x \end{cases}$$

$L = \sum_{z,x}$ "non-zero".

"It had better not be $-\infty$!"

$\max \sum$ "non-zero" $\cdot \log \frac{\beta_z}{\alpha_z}$ → choose α_z as small as possible but $\geq x_i$

(constant)

The first part of L, $P(z=z | X=x, \theta)$ is a constant w.r.t. θ^*, which we are maximizing. Note that these always have constraints (e.g. $\sum \beta_i = 1$), so we end up with Lagrangians.