CSC 446 Notes: Lecture 2

1 Entropy

Entropy is:

$$\begin{aligned} H(X) &= \sum_x P(x) \log \frac{1}{P(x)} \\ &= \int P(x) \log \frac{1}{P(x)} \mathrm{d}x \end{aligned}$$

We can think of this as a measure of information content. An example of this idea of information content is seen in Huffman coding. High frequency letters have short encodings while rarer letters have longer encodings. This forms a binary tree where the letters are at the leaves and edges to the left are 0 bits and edges to the right are 1 bits. If the probablities for the letters are all equal then this tree is balanced.

In the case of entropy we notice that $\log \frac{1}{P(x)}$ is a non-integer, so it is like an expanded Huffman coding.

1.1 Bounds on Entropy for a Discrete Random Variable

If the variable is descrete $H(X)$ is maximized when the distribution is uniform since $P(x) = \frac{1}{K}$, we see:

$$H(X) = \sum_{i=1}^{K} \frac{1}{K} \log K = \log K$$

If K is 2^n then $H(X) = \log 2^n = n$. Part of Homework 2 will be to prove that entropy on a discrete random variable is maximized by a uniform distribution ($\max_\theta H(X)$ where $\sum_n \theta_n = 1$ using the Lagrange equation).

To minimize $H(X)$ we want $P(x_i) = 1$ for some i (with all other $P(x_j)$ being zero[1]) giving $H(X) = \sum_{1 \leq j \leq K, j \neq i} 0 \log \frac{1}{0} + 1 \log 1 = 0$. We see then that:

$$0 \leq H(X) \leq \log K$$

If we consider some distribution we can see that if we cut up the "shape" of the distribution and add in gaps that the gaps that are added do not contribute to $P(x) \log \frac{1}{P(x)}$.

[1] What about $0 \cdot \log \frac{1}{0}$? It is standard to define this as equal to zero (justified by the limit being zero).

1.2 Further Entropy Equations

$$
\begin{aligned}
H(X,Y) &= \sum P(x,y) \log \frac{1}{P(x,y)} \\
H(X|Y) &= \sum_{x,y} P(x|y)P(y) \log \frac{1}{P(x|y)} \\
&= \mathrm{E}_{XY}\left[\log \frac{1}{P(x|y)}\right] \\
&= \sum_{x,y} P(x,y) \log \frac{1}{P(x|y)}
\end{aligned}
$$

2 Mutual Information

Mutual information attempts to measure how correlated two variables are with each other:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\
&= \mathrm{E}\left[\log \frac{1}{P(x)} + \log \frac{1}{P(y)} - \log \frac{1}{P(x,y)}\right] \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

Consider communicating the values of two variables. The mutual information of these two variables is the difference beween the entropy of communicating these varibales individually and the entropy if we can send them together. For example if X and Y are the same then $H(X) + H(Y) = 2H(X)$ while $H(X,Y) = H(X)$ (since we know Y if we are given X). So $I(X;Y) = 2H(X) - H(X) = H(X)$.

2.1 Covariance

A number version of mutual information is covariance:

$$
\begin{aligned}
\mathrm{Covar}[X,Y] &= \sum_{x,y} P(x,y)(x - \bar{X})(y - \bar{Y}) \\
\mathrm{Covar}[X,X] &= \mathrm{Var}[X]
\end{aligned}
$$

Covariance indicates the high level trend, so if both X and Y are generally increasing, or both generally decreasing, then the covariance will be positive. If one is generally increasing, but the other is generally decreasing, then the covariance will be negative. Two variables can have a high amount of mutual information but no general related trend and the covariance will not indicate much (probably be around zero).

2.2 KL divergence

Kullback–Leibler (KL) divergence compares two distributions over some variable:

$$
\begin{aligned}
D(P \parallel Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\
&= \mathrm{E}_P \left[\log \frac{1}{Q(x)} - \log \frac{1}{P(x)} \right] \\
&= \underbrace{H_P(Q)}_{\text{Cross Entropy}} - \underbrace{H(P)}_{\text{Entropy}}
\end{aligned}
$$

If we have the same distribution then the there is non divergence $D(P \parallel P) = 0$. In general the KL divergence is non-symetric $D(P \parallel Q) \neq D(Q \parallel P)$. If neither distribution is "special" the average $\frac{1}{2}[D(P \parallel Q) + D(Q \parallel P)]$ is sometimes used and is symetric. The units of KL divergence are log probability.

The cross entropy has an information interpretation quantifying how many bits are wasted by using the wrong code:

$$
H_P(Q) = \sum_x \underbrace{P(x)}_{\text{Sending } P} \overbrace{\log \frac{1}{Q(x)}}^{\text{code for } Q}
$$

2.3 Lower Bound for KL divergence

We will show that KL divergence is always greater or equal to zero using Jensen's inequality. First we need a definition of convex. A function f is convex if for all x_1, x_2 and θ where $0 \leq \theta \leq 1$, $f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$. This is saying that any chord on the function is above the function itself on the same interval.

Some examples of convex include a straight line and $f(x) = x^2$. If the Hessian exists for a function then $\nabla^2 f \succeq 0$ (the Hessian is positive semidefinite) indicates that f is convex. This works for a line, but not something like $f(x) = |x|$.

Jensen's inequality states that if f is convex then $\mathrm{E}[f(X)] \geq f(E[X])$.

Proof.

$$
\begin{aligned}
D(P \parallel Q) &= \mathrm{E}_P \left[\log \frac{P(x)}{Q(x)} \right] \\
&= \mathrm{E}_P \left[-\log \frac{Q(x)}{P(x)} \right]
\end{aligned}
$$

To apply Jensen's inequality we will let $-\log$ be our function and $\frac{Q(x)}{P(x)}$ be our x (note that this ratio is a number so we can push the E_P inside).

$$
\begin{aligned}
\mathrm{E}_P \left[-\log \frac{Q(x)}{P(x)} \right] &\geq -\log \mathrm{E}_P \left[\frac{Q(x)}{P(x)} \right] \\
&= -\log \sum_x P(x) \frac{Q(x)}{P(x)} \\
&= -\log 1 = 0
\end{aligned}
$$

□

Thinking of our information interpretation, we see that we always pay some cost for using the wrong code. Also note that $\log \frac{P(x)}{Q(x)}$ is sometimes positive and sometimes negative (P and Q both sum to one), yet $D(P \parallel Q) \geq 0$.

2.4 L_1 norm

The L_1 norm is defined as:

$$\|P - Q\|_1 = \sum_x |P(x) - Q(x)|$$

It can be thought of as "how much earth has to be moved" to match the distributions.

Because P and Q sum to one we quickly see that $0 \leq \|P - Q\|_1 \leq 2$. This property can be advantagous when bounds are needed.

3 Naive Bayes

Consider the example of predicting whether a congress member is a Republican or a Democrat based on votes on bills. If we let Y be the classification variable and X be the features we see then we are trying to solve:

$$\operatorname*{argmax}_y P(Y|X)$$

Using the product rule we have $P(Y|X) = \frac{P(Y,X)}{P(X)} = \frac{c(Y,X)}{c(X)}$ for counts c. But we cannot expect that congress members are going to vote exactly the same making the denominator likely to be zero. Naive Bayes takes the expansion of joint probability in the chain rule and applies an assumption of independence: $P(X_i, X_j|Y) = P(X_i|Y)P(X_j|Y)$. This gives:

$$\begin{aligned} P(Y, X_1, X_2, \ldots, X_N) &= P(Y)P(X_1|Y)P(X_2|Y, X_1)\cdots P(X_N|Y, X_1, \ldots, X_N) \\ &= P(Y)P(X_1|Y)\cdots P(X_N|Y) \end{aligned}$$

Now we can use counts without issue. We note that $P(Y|X_1^N) = \frac{P(Y,X_1^N)}{P(X_1^N)}$ has a constant denominator, so we can just maximize on $P(Y, X_1^N)$.

$$\begin{aligned} \operatorname*{argmax}_y P(Y|X_1^N) &= \operatorname*{argmax}_y P(Y, X_1^N) \\ &\simeq \operatorname*{argmax}_y P(Y) \prod_{n=1}^{N} P(X_i|Y) \end{aligned}$$

One can compute $P(X_i|Y)$ using counting on training data.

Mutual information $I(Y; X_i)$ can be used for feature selection according to the following procedure: sort features by mutual information, and keep the top K features. Choose K by optimizing classification error on held-out development data.

Ryan Yates 1/12; DG 1/13