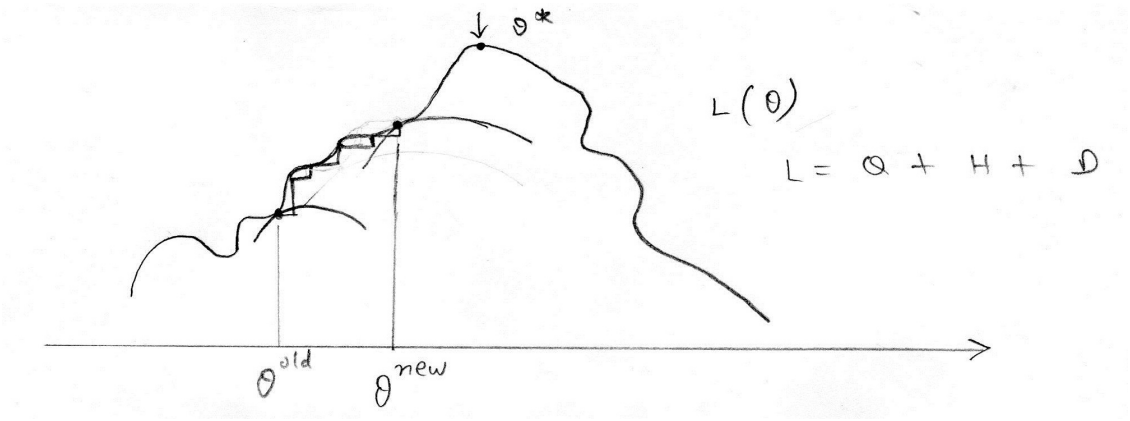
CSC 446 Notes: Lecture 12

Typed by Titas De

March 30, 2012

Expectation Maximisation



1 Gradient Ascent ($\frac{\partial L}{\partial \theta}$)

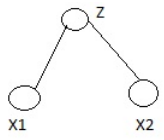
Gradient Ascent $\frac{\partial L}{\partial \theta}$ means the gradient of likelihood function w.r.t. parameters.

Using the gradient ascent, our new parameter θ^{new} is given by :

$$\theta^{new} = \theta^{old} + \eta \frac{\partial L}{\partial \theta}$$

EM (expectation maximization) generally finds a local optimum of θ more quickly, because in this method we optimize the q-step before making a jump. It is similar to saying that we make a big jump first and then take smaller steps accordingly. But for Gradient Ascent method, the step size varies with the likelihood gradient, and so it requires more steps when the gradient is not that steep.

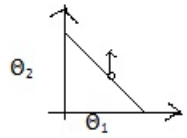
We are required to find $E_{P(Z|X,\theta)}[Z|X]$



$$\text{Our } \theta = \begin{cases} P(Z) \\ P(X_1|Z) \\ P(X_2|Z) \end{cases}$$

which is a long list of all the possible probabilities, after unfolding each of them.

We should make sure that the parameters ($\theta = [\theta_1, \theta_2]$, for only 2 parameters) always stays within the straight line shown below. It should never go out of the line.



$$P(Z = k) = \theta_k = \frac{e_k^{\lambda}}{\sum_{k'} e_{k'}^{\lambda}}$$

In this case, the gradient ascent is used to find the new value of λ

$$\lambda_{new} = \lambda_{old} + \eta \frac{\partial L}{\partial \lambda}$$

2 Newton's Method

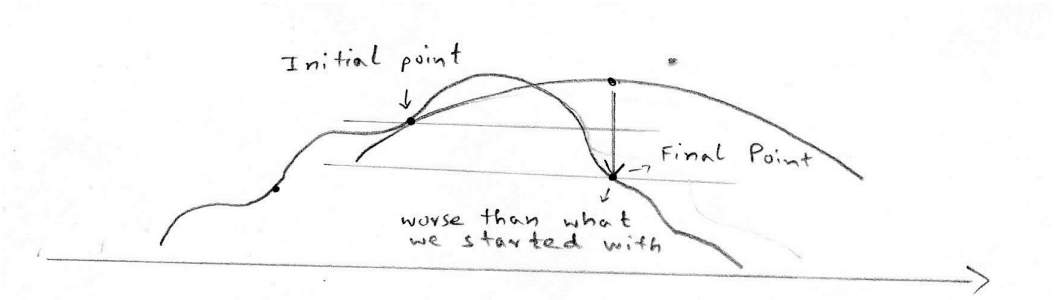
$$\theta^{new} = \theta^{old} + (\nabla_\theta^2 L)^{-1}(\nabla_\theta L)$$

In Newton' Method, we approximate the curve with a quadratic function, and we jump to the maximum of the approximation in each step.

Differences between Newton's Method and EM :

a) Newton's Method takes a lot of time, because we need to calculate the Hessian Matrix, which is the 2nd derivative.

b) Since there is no KL divergence in Newton's Method, there is always a chance that we take a big jump, and arrive at a point far away from the global optimum and in fact worse than where we started.



3 Variational Method

In this method, at first we trace the Likelihood function by a new parameter, then fix and optimize that parameter, and then once again start the iteration, until we get the optimum value of the Likelihood function.

4 Mixture of Gaussians

$$\begin{aligned} P(X) &= \sum_k P(Z = k) N(X|\mu_k, \Sigma_k) \\ &= \sum_k \lambda_k N(X|\mu_k, \Sigma_k), \end{aligned}$$

where μ and Σ are the mean vector and co-variance matrix respectively.

$$N(X|\mu, \Sigma) = \frac{e^{-0.5(X-\mu)^T \Sigma^{-1}(X-\mu)}}{(2\pi)^{D/2}|\Sigma|^{0.5}} \tag{1}$$

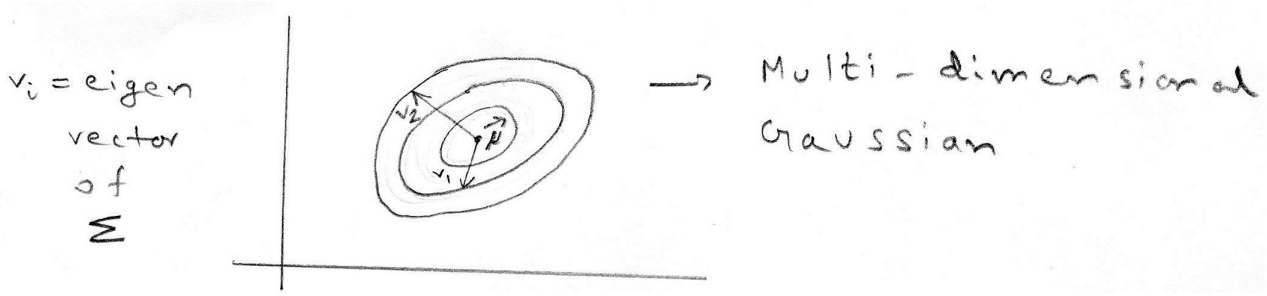
Here X and μ are vectors, Σ is a 2-D matrix, and D is the dimensionality of data X. The left side side of equation (1) refers to :

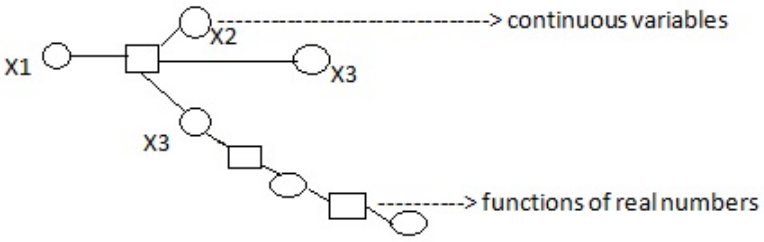
$$N\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \middle| \mu, \Sigma\right)$$

For 1-D data,

$$N(X|\mu,\sigma) = \frac{e^{-0.5(X-\mu)^2/\sigma^2}}{(2\pi)^{0.5}\sigma}$$

Equation (1) is similar to writing as $f(X) = X^T AX$; wherein we are stretching the vector space.





For continuous variables the equation :

$$r_{m\to n} = \sum_{\vec{X}_m \setminus X_n} f(\vec{X_m}) \prod_{n \in N(m)\setminus n} q_{n\to m}$$

changes to

$$r_{m\to n} = \int f(\vec{X}_m) \prod_{n \in N(m)\setminus n} q_{n\to m}\; d(\vec{X}_m \setminus X_n)$$

From Assignment 1, we obtained the following formulas:

$$\mu^* = \frac{\sum_n X_n}{N}$$

$$\sigma^* = \sqrt{\frac{\sum_n (X_n - \mu^*)^2}{N}}$$

$$\Sigma = \frac{\sum_n (X_n - \mu^*)(X_n - \mu^*)^T}{N}\text{, where } \Sigma \text{ is the covariance matrix}$$

$$\Sigma_{ij} = \frac{\sum_n (X_n^i - \mu_i)(X_n^j - \mu_j)}{N} = \text{covariance}(\mathrm{X_i}, \mathrm{X_j})$$

Θ (parameters of the model) $= \lambda, \mu_k, \Sigma$, where λ and μ_k are vectors

E-step

for n=1,2,. . . ,N

$$P(Z|X^n) = \frac{P(Z)P(X^n|Z)}{P(X^n)}$$
$$= \frac{(\lambda_k/z_k)\exp[-0.5(X^n-\mu_k)^T\Sigma^{-1}(X^n-\mu_k)]}{\sum_{k'}(1/z_{k'})\exp[-0.5(X^n-\mu_{k'})^T\Sigma^{-1}(X^n-\mu_{k'})]}$$

M-step

for k=1,2,. . . ,K (total number of hidden variables)

$$\lambda_k = \frac{\sum_{n=1}^{N} P(Z=k|X^n)}{N}$$

$$\mu_k = \frac{\sum_n P(Z=k|X^n)X^n}{\sum_n P(Z=k|X^n)}$$

$$\Sigma_k = \frac{\sum_n P(Z=k|X^n)(X^n-\mu_k)(X^n-\mu_k)^T}{\sum_n P(Z=k|X^n)}$$