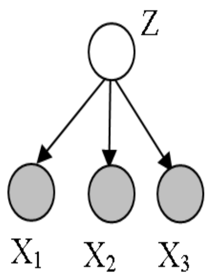
CS281A/Stat241A: Statistical Learning Theory

Expectation Maximization Algorithm (10/19/04)

Lecturer: Michael I. Jordan *Scribes: Ben Wild*



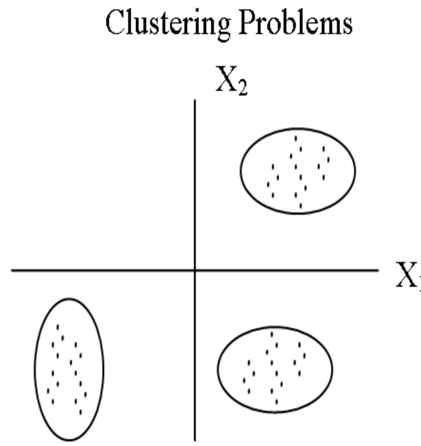


Figure 1: Partially observed case: class labels Z are not observed, but observed points X, shown on the plot on the right, appear to cluster in three groups.

Gaussian Mixtures: The generative mixture model with k mixture components, shown in 1, has a likelihood that can be written out in terms of the prior probability on the hidden (or latent) variable Z and the probability of the observed variables X conditioned the class Z with a Gaussian distribution:

$$p(x|\theta) = \sum_{Z} p(z|\pi)p(x|z, \mu, \Sigma)$$

$$= \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

Data: $X = \{x_n\}_{n=1}^{N}$

The Log Likelihood:

$$\ell(\theta) = \sum_{n} \log \sum_{i} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)$$

Expectation Maximization (Iterative approach for ML estimation):

Complete Log Likelihood:

Assume we have complete data $\{(x_n, z_n)\}_{n=1}^N$

Graphical Model: Becomes fully observed graph.

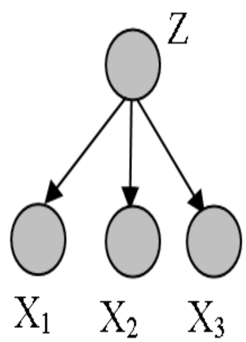


Figure 2: fully observed case

$$\ell_c(\theta) = \sum_n \log p(z_n|\pi) p(x_n|z_n, \mu, \Sigma)$$

where $p(z_n|\pi) = \prod_i \pi_i^{z_n^i}$

and $p(x_n|z_n, \mu, \Sigma) = \prod_i \mathcal{N}(x_n|\mu_i, \Sigma_i)^{z_n^i}$

$$\ell_c(\theta) = \sum_n \sum_i z_n^i \log \pi_i + \sum_n \sum_i z_n^i \log \mathcal{N}(x_n|\mu_i, \Sigma_i)$$

Will get sample means, sample covariances by taking derivates, setting to 0 and solving.

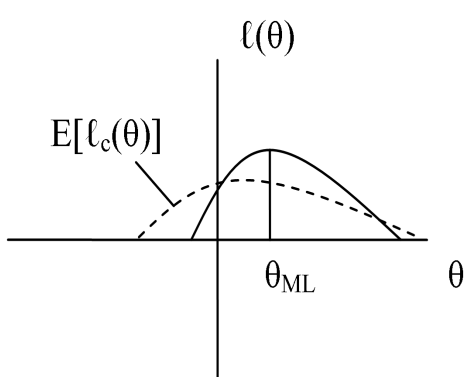


Figure 3: Likelihood function

Don't get to see Z, so you have a random set of likelihoods based on Z.

We get a fixed likelihood curve if we take the expected value of the complete likelihood curve, and then we can find a maximum on that.

$$E_q[\ell_c(\theta)] = \sum_n \sum_i (E_q z_n^i) \log \pi_i + \sum_n \sum_i (E_q z_n^i) \log \mathcal{N}(x_n, \mu_i, \Sigma_i)$$

$$E_q z_n^i = q(z_n^i = 1)$$

EM Algorithm:

The EM algorithm is a coordinate ascent algorithm for a particular choice of $q(z_n)$. Set $q(z_n) = p(z_n|x_n, \theta^{(t)})$

E Step:

Compute expected complete log likelihood using $p(z_n|x_n, \theta^{(t)})$

$$Let \quad \tau_n^{i(t)} = E_{p(z_n|x_n,\theta^{(t)})}[z_n^i]$$

M Step:

$$\theta^{(t+1)} = \arg\max_\theta E_{p(z_n|x_n,\theta^{(t)})}[\ell_c(\theta)]$$

$$\pi_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)}}{N}$$

$$\mu_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} x_n}{\sum_n \tau_n^{i(t)}}$$

$$\Sigma_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} (x_n - \mu_i^{(t+1)})(x_n - \mu_i^{(t+1)})^T}{\sum_n \tau_n^{i(t)}}$$

Picture of what is happening here:

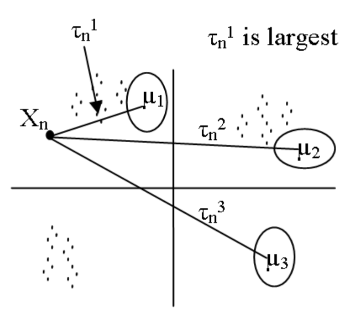


Figure 4: EM Algorithm

In particular, for the E step:

$$\tau_n^{i(t)} = p(z_n^i = 1|x_n, \theta^{(t)})$$

Bayes Rule:

$$\tau_n^{i(t)} = \frac{p(x_n|z_n^i = 1, \theta^{(t)})p(z_n^i = 1|\theta^{(t)})}{\sum_i p(x_n|z_n^i = 1, \theta^{(t)})p(z_n^i = 1|\theta^{(t)})}$$

After a few iterations this will converge to the correct values.

A note about the K-Means Algorithm: (ad hoc algorithm for clustering)

Assumes that you know the number of clusters. Algorithm consists of:

-partition data, update estimates

-re-partitioning data

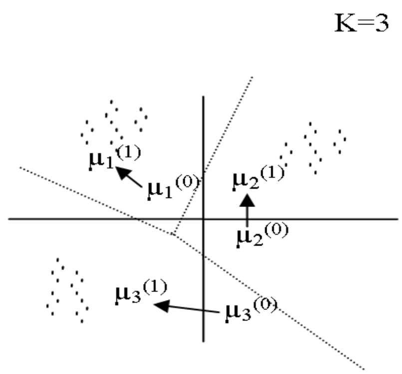


Figure 5: K-means algorithm

K-Means doesn't have covariance structure, doesn't take into account the π parameters representing the relative frequency of each class. K-Means will get hung up on some data sets.

Can use K-means for initialization and then run the EM algorithm.

Back to EM algorithm:

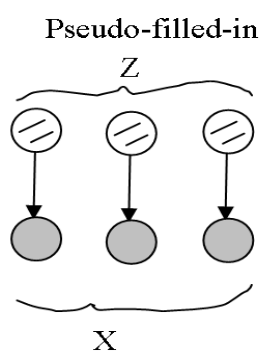


Figure 6: Pseudo filled in variables

E step $\equiv$ Inference algorithm (e.g. Sum Product to get sufficient statistics)

M step $\equiv$ Maximum likelihood algorithm

-EM is coordinate ascent in $\ell(\theta)$

Jensen's Inequality:

For convex f:

$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i)$ with equality iff f is linear.

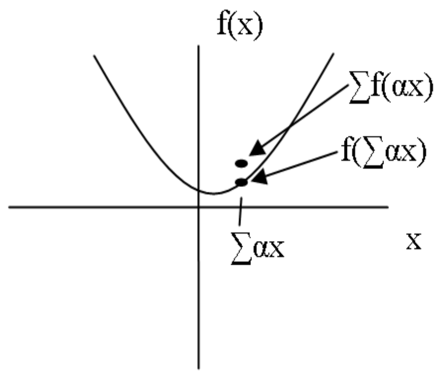


Figure 7: Jensen's Inequality

Examples:

$$\exp\left\{\sum_i \alpha_i x_i\right\} \leq \sum_i \alpha_i \exp\{x_i\}$$

$$\log\left\{\sum_i \alpha_i x_i\right\} \geq \sum_i \alpha_i \log\{x_i\}$$

since the log function is concave.

General setting

X- Observed variables

Z – Latent variables

Probability model : $p(x, z|\theta)$

Log likelihood: $\ell(\theta) = \log p(x|\theta)$

$$\ell(\theta) = \log \sum_z p(x, z|\theta)$$

$$= \log \sum_z q(z|x) \tfrac{p(x,z|\theta)}{q(z|x)}, \text{ q is a free parameter}$$

$$\log \sum_z q(z|x) \frac{p(x, z|\theta)}{q(z|x)} \geq \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)}$$

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

In other words, for any q:

$$\ell(\theta) \geq \mathcal{L}(q,\theta)$$

EM

-Coordinate ascent in $\mathcal{L}(q,\theta)$

E step:

$$q^{(t+1)}(z|x) = \arg\max_q \mathcal{L}(q,\theta^{(t)})$$

M step:

$$\theta^{(t+1)}(z|x) = \arg\max_\theta \mathcal{L}(q^{(t+1)},\theta)$$

So for each step of the EM algorithm, take one step uphill on one coordinate, then a step up hill on the other.

Another way to write the M step:

$$\mathcal{L}(q^{(t+1)},\theta) = \sum_z q^{(t+1)} \log q^{(t+1)} + \sum_z q^{(t+1)}(z|x) \log p(x,z|\theta)$$

$$\theta^{(t+1)}(z|x) = \arg\max_\theta \sum_z q^{(t+1)}(z|x) \log p(x,z|\theta)$$

$$= \arg\max_\theta E_{q^{(t+1)}}[\log p(x,z|\theta)]$$

$E_{q^{(t+1)}}[\log p(x,z|\theta)]$ is the expected complete log likelihood, or the expected value of the complete log likelihood.