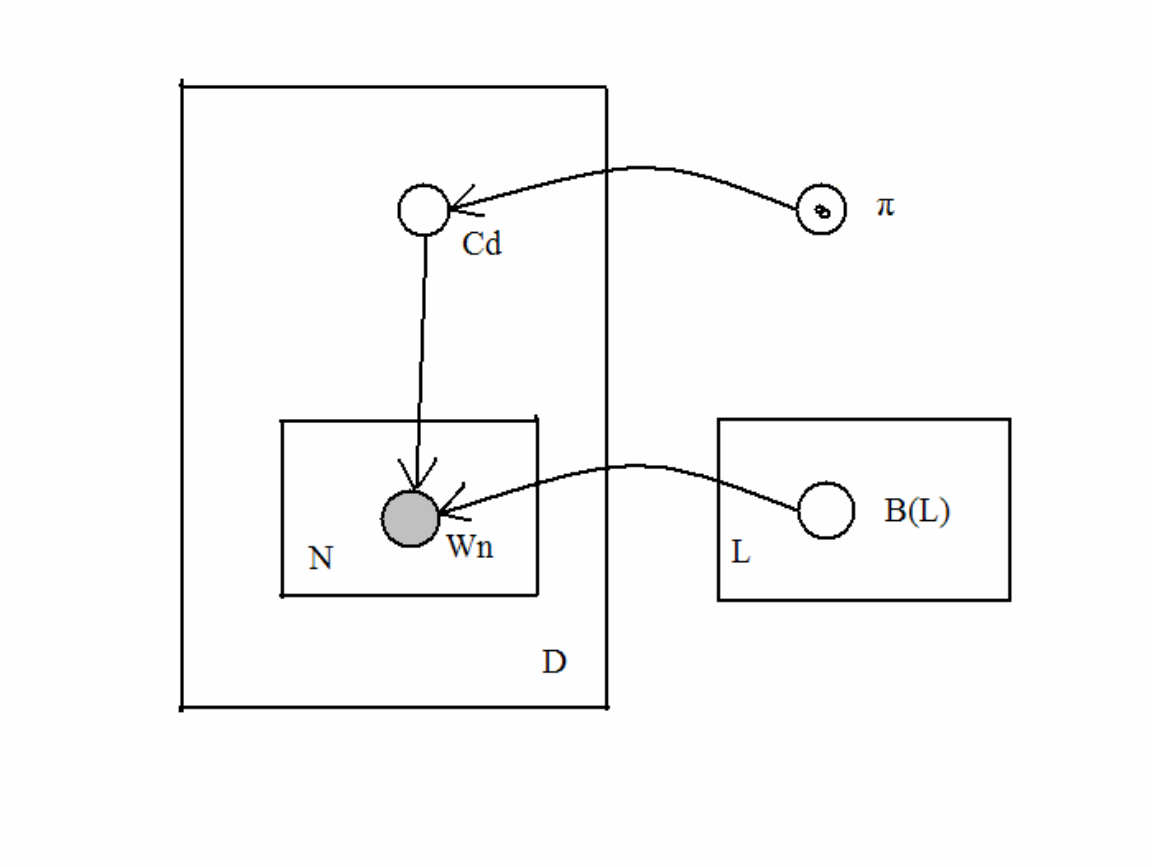
Mixture Models

Tuesday March 27th
David Blei



L – group
D – documents
N – words per document

Goal: find MLE on B and π

That is find the most likely distribution of words B_l for each of the L classes and the most likely distribution π of documents among classes, given the N words in each of the D documents.

$$L(\beta_{1:L}, \pi, D) = \sum_{d=1}^{D} \log \sum_{l=1}^{L} p(c = l \mid \Pi) \prod_{n=1}^{N} p(w_n \mid \beta_l)$$

Where $D = \{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_D\}$

Explanation of above equation:

First the above equation seeks to calculate
Max P(documents)
Therefore it computes

$$\max \sum_{d=1}^{D} \log(P(document_d))$$

Probability of a document = $p(w_{1:N} \mid \beta, \pi) = \sum_{l=1}^{L} p(c = l \mid \pi) \prod_{n=1}^{N} p(w_n \mid \beta_l)$

Here we marginalize out B and π so that we may actually calculate the probability of a document.

$\sum_{l=1}^{L} p(c = l \mid \pi)$ = probability of the class l

$\prod_{n=1}^{N} p(w_n \mid \beta_l)$ = probability of all the words in the document for the given class

Therefore:

$$\sum_{l=1}^{L} p(c = l \mid \pi) \prod_{n=1}^{N} p(w_n \mid \beta_l)$$

is the probability of the class times the probability of all the words given the class. Therefore we have marginalized with respect to B and π. That is we have summed the probability out for each class in the distribution π and for each vocabulary B associated with that class.

Remember from previous lectures that $p(A) = \sum_{B} p(A, B)$

Because these probabilities are tiny we take the log. We do this for all documents 1 to D. Giving us.

$$L(\beta_{1:L}, \pi, D) = \sum_{d=1}^{D} \log \sum_{l=1}^{L} p(c = l \mid \Pi) \prod_{n=1}^{N} p(w_n \mid \beta_l)$$

Which is just the log likelihood of seeing all the documents for a given π and B.

L–means:

1. partition the data according to current means
2. re-estimate the means

If we knew the classes,

$$\hat{\pi} = \sum_d C_d / D \quad \hat{\pi} = \text{\# of docs in class / \# of docs} \qquad \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 1 \\ . \\ . \\ 0 \end{pmatrix} = C$$

$$\hat{\beta}_l = (\sum_d C_d^l \sum_m w_{d,n}) / \sum_d C_d^l N$$

$\hat{\beta}_l =$ sum word counts in ℓth cluster / words in ℓth cluster

Summing w vectors to get word counts

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

That is the solution to the MLE, given the classes, is to just calculate the straightforward arithmetic estimates. Because, as was shown in the first homework, these arithmetic estimates yield the maximum likelihood for the observed data.

EM Algorithm:

Iterate to obtain: $\hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}$

Replace C_d with $E[C_d \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}]$

$$C_d^1 \in \{0,1\}$$
$$C_d^2 \in \{0,1\}$$
$$C_d^3 \in \{0,1\}$$

E[ ] is a vector of probabilities that d is in each cluster $\begin{pmatrix} .03 \\ .05 \\ .11 \\ . \\ . \\ \end{pmatrix}$

$$\hat{\pi} = \sum_d \mathrm{E}[c_d \mid w_d] \Big/ D$$

$$\hat{\beta}_l = \sum_d \mathrm{E}[c_d^l \mid \vec{w}_d] \sum_n w_{d,n} \Big/ \sum_d \mathrm{E}[c_d^l \mid w_d] N$$

We are now weighting documents in classes rather than considering them fixed in one class. That is rather than considering a document to be either in class i or j the way L-means (also k-means although in this lecture it was called L-means for consistency with the above equation's notation), it can be considered to be in class i with probability x and j with probability y. Therefore each document is associated with a vector E[] that contains the probabilities that it belongs in each cluster.

E-step:

Replace C_d with $\mathrm{E}[C_d \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}]$

M-step:

Compute a weighted MLE

According to $\mathrm{E}[C_d \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}]$

Or written out more fully

M-step

$$\hat{\pi}^{(t+1)} = \frac{\sum_d \mathrm{E}[c_d \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}]}{D}$$

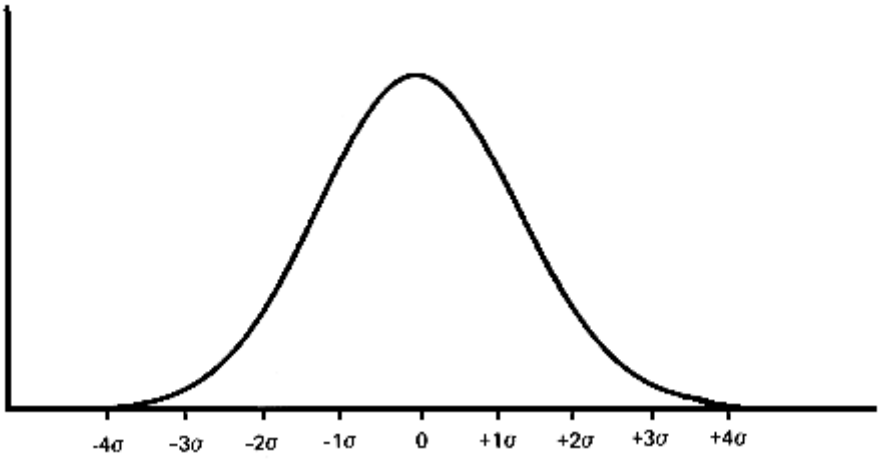
$$\hat{\beta}_l^{(t+1)} = \frac{\sum_d \mathrm{E}[c_d^l \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}] \sum_n w_{d,n}}{\sum_d \mathrm{E}[c_d^l \mid \vec{w}_d, \hat{\pi}^{(t)}, \hat{\beta}_{1:L}^{(t)}] \cdot N}$$

EM is finding a fixed point of the expected complete log likelihood

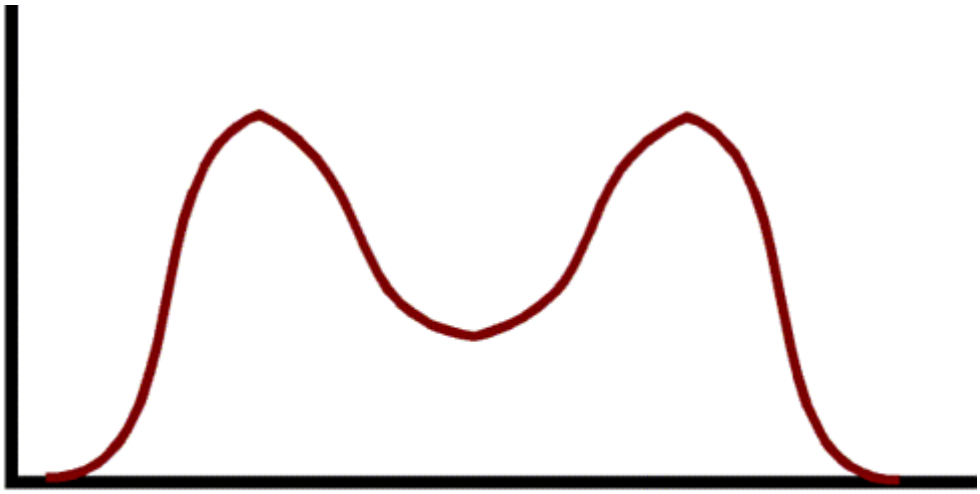
$$\mathrm{E}[\log p(C_{1:D}, \vec{w}_{1:D} \mid \pi, \beta_{1:L})] = \sum_{d=1}^{D} \sum_{l=1}^{L} \mathrm{E}[c_d^l \mid \vec{w}_d] \log p(c_d^l \mid \pi_l) + \sum_{d=1}^{D} \sum_{l=1}^{L} \mathrm{E}[c_d^l \mid w_d] \log p(\vec{w}_d \mid \beta_l)$$

In plain English, EM is computing the posterior probability vectors E[] and then computing the maximum log likelihood of expectation given these vectors. That is E estimates the posterior probabilities E[] and M computes the new means (or in our example the most probable distributions π and B) given E[]. At which point, E recomputes new probabilities $\mathrm{E}^{(t+1)}$[] based on the new means (values of π and B in our case), etc., etc.

If Data are Gaussian



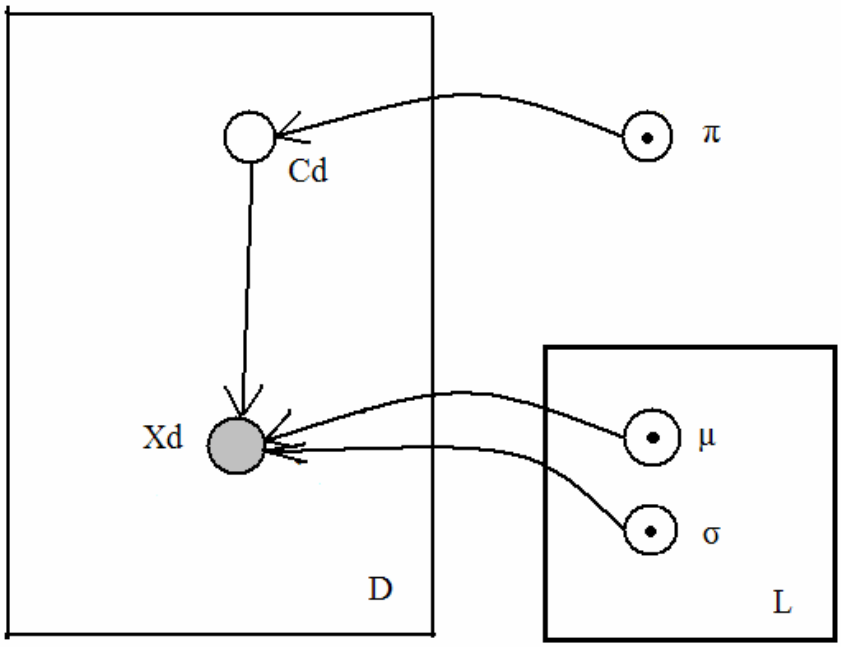
$$p(x \mid \mu, \sigma)$$



$$p(x \mid \pi, \mu_1, \sigma_1, \mu_2, \sigma_2)$$

Marginalize: $\pi_1 p(x \mid \mu_1, \sigma_1) + \pi_2 p(x \mid \mu_2, \sigma_2)$

Gaussian Mixture Model



EM is a way of fitting parameters in latent variable models

E-step – values of latent variables are “filled in” (expectation)
M-step – parameters are fit to match filled in variables (maximization)