Recitation 3: October 27

Lecturer: Mariano Schain *Scribe: ym*

3.1 Expectation Maximization (EM) algorithm

We assume a two stage process for generating point x_i. First, given the unknown parameters θ_1 we generate c_i (c_i is unobserved). Given c_i and theunknown parameters θ_2 we generate the observation x_i.

$$\xrightarrow{\theta_1} c_i \xrightarrow{\theta_2} x_i$$

Therefore, the model consists of two parametrised distributions:
$\Pr[c;\theta_1]$ the distribution of c when its parameter is θ_1.
$\Pr[x|c;\theta_2]$ the conditional distribution of the observation x given the hidden c, for a parameter θ_2. Our goal is to recover the parameters $\theta = (\theta_1, \theta_2)$ based on the observations $\{x_i\}_{i=1}^n$.

Consider the log-likelihood:

$$\begin{aligned}\ell(\theta|\{x_i\}) &= \log \Pr[\{x_i\}_{i=1}^n|\theta] \\ &= \sum_{i=1}^n \log \Pr[x_i|\theta] \\ &= \sum_{i=1}^n \log \left(\sum_c \Pr[c_i = c;\theta_1] \Pr[x_i|c_i = c;\theta_2] \right)\end{aligned}$$

Maximizing $\ell(\theta|\{x_i\})$ above is difficult in general due to the sum within the log (note however that the form of the probabilities within the log is exactly given by our model). Therefore, we do the following:

For a given x_i we will define the probability of the hidden result being $c_i = j$ (we assume that the hidden result is one of K possible results).

$$a_{i,j}^t = \Pr[c_i = j|x_i;\theta^t]$$

The t indicates that $a_{i,j}^t$ is computed at iteration t, based on the parameter values θ^t that were already computed at the end of iteration $t-1$.

Recall that

$$\Pr[x|y] = \frac{\Pr[x,y]}{\Pr[y]} = \frac{\Pr[y|x]\Pr[x]}{\sum_x \Pr[x,y]}$$

Therefore,

$$a_{i,j}^t = \Pr[c_i = j | x_i; \theta^t] = \frac{\Pr[x_i | c_i = j; \theta_2^t] \Pr[c_i = j; \theta_1^t]}{\sum_c \Pr[x_i | c_i = c; \theta_2^t] \Pr[c_i = c; \theta_1^t]}$$

Note again that all forms of the above probabilities are exactly given by our model, and assuming the parameters θ^t were computed in the previous iteration the algorithm can directly plug the computed parameters θ^t above and compute $a_{i,j}^t$. Also note that $\sum_j a_{i,j}^t = 1$.

The EM algorithm alternates between an E-step and an M-step. In the E-step, we define a function $Q(\theta|\theta^t)$ as the average likelyhood (over the probabilities $a_{i,j}^t$ of the unobserved outcomes) of our observations $\{x_i\}$:

$$\begin{aligned} \textbf{E-step}: \quad Q(\theta|\theta^t) &= \sum_{i=1}^{n} \sum_{j=1}^{k} a_{i,j}^t \log \Pr[x_i, c_i = j; \theta] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} a_{i,j}^t \left(\log \Pr[x_i | c_i = j; \theta_2] + \log \Pr[c_i = j; \theta_1] \right) \end{aligned}$$

Note that the influence of θ^t in Q is through the coefficients $a_{i,j}^t$. Also note that Q above is a function of the model parameters $\theta = (\theta_1, \theta_2)$ (since the coefficients $a_{i,j}^t$ are previously computed constants). Terefore, in the M-Step of the EM algorithm we find the parameters θ that maximize Q.

$$\textbf{M-step}: \qquad \theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t)$$

This time the maximization may be easy due to the log applied to each of the model probabilities, as illustrated in the following sections.

3.2 Example 1: Three coins

In the first step we flip a coin with bias λ which returns either 1 or 2. More precisely, $\Pr[c_i = 1] = \lambda$ and $\Pr[c_i = 2] = 1 - \lambda$. If $c_i = 1$ then we flip a coin with bias p_1 to set x_i and If $c_i = 2$ then we flip a coin with bias p_2 to set x_i. The flow of information:

$$\xrightarrow{\lambda} c_i \xrightarrow[\{1,2\}]{p_1, p_2} x_i$$

The model has:
(1) $\Pr[c_i = 1] = \lambda$, (2) $\Pr[x_i = 1 | c_i = 1] = p_1$, and (3) $\Pr[x_i = 1 | c_i = 2] = p_2$.

We observe the sequence $x = \{x_i\}_{i=1}^n$, for example $x = (0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0)$ for $n = 13$. We would like to run EM to recover the missing parameters $\theta = \{\lambda, p_1, p_2\}$.

For the$E$-Step at iteration t, assume we have the model parameters $\theta^t = \{\lambda^t, p_1^t, p_2^t\}$ and compute $a_{i,j}^t$ (we only need $a_{i,1}^t$ since $a_{i,2}^t = 1 - a_{i,1}^t$):

$$a_{i,1}^t = \frac{\lambda^t (p_1^t)_i^x (1-p_1^t)^{1-x_i}}{\underbrace{\lambda^t (p_1^t)_i^x (1-p_1^t)^{1-x_i}}_{c_i=1} + \underbrace{(1-\lambda^t)(p_2^t)_i^x (1-p_2^t)^{1-x_i}}_{c_i=2}}$$

Now, the resulting form of $Q(\theta|\theta^t)$ is:

$$\begin{aligned}
Q(\theta|\theta^t) &= \sum_{i=1}^{n} a_{i,1}^t \left(\log\lambda + x_i \log p_1 + (1-x_i)\log(1-p_1)\right) \\
&\quad + (1-a_{i,1}^t)\left(\log(1-\lambda) + x_i \log p_2 + (1-x_i)\log(1-p_2)\right) \\
&= \left(\sum_{i=1}^{n} a_{i,1}^t \log\lambda + (1-a_{i,1}^t)\log(1-\lambda)\right) \\
&\quad + \left(\sum_{i=1}^{n} a_{i,1}^t [x_i \log p_1 + (1-x_i)\log(1-p_1)]\right) \\
&\quad + \left(\sum_{i=1}^{n} (1-a_{i,1}^t)[x_i \log p_2 + (1-x_i)\log(1-p_2)]\right)
\end{aligned}$$

In the M-step we are maximizing Q:

$$\theta^{t+1} = (\lambda^{t+1}, p_1^{t+1}, p_2^{t+1}) = \arg\max_{\theta=(\lambda,p_1,p_2)} Q(\theta|\theta^t)$$

Fortunately, this breaks up to three optimization problems

$$\lambda^{t+1} = \arg\max_{\lambda} \left(\sum_{i=1}^{n} a_{i,1}^t\right)\log\lambda + \left(\sum_{i=1}^{n}(1-a_{i,1}^t)\right)\log(1-\lambda) = F(\lambda)$$

We compute

$$F'(\lambda) = \frac{\sum_{i=1}^n a_{i,1}^t}{\lambda} - \frac{\sum_{i=1}^n (1-a_{i,1}^t)}{1-\lambda} = 0$$

and we get

$$\lambda^{t+1} = \frac{\sum_{i=1}^n a_{i,1}^t}{n}$$

We need to verify that this is the maximum, by checking the second derivative

$$F''(x) = -\frac{\sum_{i=1}^n a_{i,1}^t}{\lambda^2} - \frac{\sum_{i=1}^n (1-a_{i,1}^t)}{(1-\lambda)^2} < 0$$

Similarly we maximize p_1 and p_2 and get

$$p_1^{t+1} = \arg\max_{p_1} \sum_{i=1}^{n} a_{i,1}^t [x_i \log p_1 + (1 - x_i) \log(1 - p_1)] = F_1(p_1)$$

and get

$$p_1^{t+1} = \frac{\sum_{i=1}^n a_{i,1}^t x_i}{\sum_{i=1}^n a_{i,1}^t}$$

and similarly,

$$p_2^{t+1} = \frac{\sum_{i=1}^n (1 - a_{i,1}^t) x_i}{\sum_{i=1}^n (1 - a_{i,1}^t)}$$

3.3 Example 2: Mixture of Gaussians

In this setting we have a distribution $p = (p_1, \ldots, p_k)$ over k multivariate Gaussians of d dimensions. Namely, the probability of a sample to originate from the j^{th} Gaussian is $\Pr[c_i = j] = p_j$. The points in the jth MVN are generated using $MVN(\mu_j, \epsilon I)$, where $\mu_j \in \mathbb{R}^d$ and I is the identity $d \times d$ matrix. Therefore, the density function of the observation x_i given that it originates from the j^{th} Gaussian is:

$$f_j(x_i) = \frac{1}{(\sqrt{2\pi}\epsilon)^d} e^{-\frac{1}{2\epsilon^2}\|x_i - \mu_j\|^2}$$

Therefore, our model is $\theta = (\{p_j\}, \{\mu_j\})$.

We set the $a_{i,j}^t$ as follows

$$a_{i,j}^t = \frac{p_j^t f_j^t(x_i)}{\sum_{r=1}^k p_r^t f_r^t(x_i)}$$

Note that the values of the parameters $\{\mu_j^t\}$ (which are given at the E-Step, as computed by the M-Step of the preceeding iteration) appear in $f_j^t(x_i)$ - this is actually the meaning of the notation t in $f_j^t(x_i)$.

In the E-step we therefore have

$$Q(\theta|\theta^t) = Q((\{p_j\}, \{\mu_j\})|\theta^t) = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{i,j}^t \left(\log p_j + \text{const} - \frac{1}{2\epsilon^2} \|x_i - \mu_j\|^2 \right)$$

In the M-step we can separately maximize p^{t+1} and μ^{t+1}.

$$\begin{aligned} p^{t+1} &= \arg\max_p \sum_{i=1}^{n}\sum_{j=1}^{k} a_{i,j}^t \log p_j \\ &= \arg\max_p \sum_{j=1}^{k}(\sum_{i=1}^{n} a_{i,j}^t)\log p_j \end{aligned}$$

Recall that we have the constraint that $\sum_{j=1}^{k} p_j = 1$. As we saw a few times, the maximizer is,

$$p_j^{t+1} = \frac{\sum_{i=1}^{n} a_{i,j}^t}{\sum_{j=1}^{k}\sum_{i=1}^{n} a_{i,j}^t} = \frac{\sum_{i=1}^{n} a_{i,j}^t}{n}$$

For the values of μ^{t+1} we have

$$\begin{aligned} \mu^{t+1} &= \arg\max_\mu \sum_{i=1}^{n}\sum_{j=1}^{k} -\frac{1}{2\epsilon^2}\|x_i - \mu_j\|^2 \\ &= \arg\min_\mu \sum_{i=1}^{n}\sum_{j=1}^{k} \|x_i - \mu_j\|^2 \end{aligned}$$

As we saw in the k-means, the minimizer is,

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{n} a_{i,j}^t x_i}{\sum_{i=1}^{n} a_{i,j}^t}$$

In the next recitation we will review the connection of this setting to the K-means algorithm and the related interpretation of the ϵI covariance matrix.