

Recitation 4: November 16

Lecturer: Regev Schweiger

Scribe: Regev Schweiger

4.1 EM Example: Mixture of Gaussians

We assume a two stage process for generating each point $\mathbf{x}_1, \dots, \mathbf{x}_n$. In this setting we have a distribution $\mathbf{p} = (p_1, \dots, p_k)$ over k multivariate Gaussians of d dimensions. Let Z_i be the index of the Gaussian from which the i -th point is sampled. Namely, the probability of a sample to originate from the j^{th} Gaussian is $\Pr[Z_i = j] = p_j$.

We limit ourselves in this discussion to Gaussians with covariance matrix of the form ϵI . The points in the j^{th} MVN are generated using $MVN(\boldsymbol{\mu}_j, \epsilon I)$, where $\boldsymbol{\mu}_j \in \mathbb{R}^d$ and I is the identity $d \times d$ matrix. Therefore, the density function of the observation \mathbf{x}_i given that it originates from the j^{th} Gaussian is:

$$f_j(\mathbf{x}_i) = \frac{1}{(\sqrt{2\pi\epsilon})^d} e^{-\frac{1}{2\epsilon} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}$$

Therefore, the parameters of our model are $\boldsymbol{\theta} = (p_1, \dots, p_k, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k)$.

Define $a_{i,j}^t$ as the posterior distribution of Z_i , under the parameters $\boldsymbol{\theta}^t$:

$$a_{i,j}^t = \Pr_{\boldsymbol{\theta}^t} [Z_i = j | \mathbf{X}_i = \mathbf{x}_i] = \frac{p_j^t f_j^t(\mathbf{x}_i)}{\sum_{r=1}^k p_r^t f_r^t(\mathbf{x}_i)}$$

Note that the values of the parameters $\{\boldsymbol{\mu}_j^t\}$ (which are given at the E -Step, as computed by the M -Step of the preceding iteration) appear in $f_j^t(\mathbf{x}_i)$ - this is actually the meaning of the notation t in $f_j^t(\mathbf{x}_i)$.

In the E -step we therefore have, with \mathbf{Z} distributed according to the posterior distribution:

$$\begin{aligned}
 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) &= E_{\mathbf{Z}} \left[\log \Pr_{\boldsymbol{\theta}} [\mathbf{X} = \mathbf{x}, \mathbf{Z}] \right] \\
 &= E_{\mathbf{Z}} \left[\sum_{i=1}^n \log \Pr_{\boldsymbol{\theta}} [\mathbf{X}_i = \mathbf{x}_i, Z_i] \right] \\
 &= E_{\mathbf{Z}} \left[\sum_{i=1}^n \log \Pr_{\boldsymbol{\theta}} [Z_i] + \log \Pr_{\boldsymbol{\theta}} [\mathbf{x}_i | Z_i] \right] \\
 &= E_{\mathbf{Z}} \left[\sum_{i=1}^n \log \Pr_{\boldsymbol{\theta}} [Z_i] + \log \Pr_{\boldsymbol{\theta}} [\mathbf{x}_i | Z_i] \right] \\
 &= E_{\mathbf{Z}} \left[\sum_{i=1}^n \log p_{Z_i} + \left(\text{const} - \frac{1}{2\epsilon} \|\mathbf{x}_i - \boldsymbol{\mu}_{Z_i}\|^2 \right) \right] \\
 &= \sum_{i=1}^n E_{\mathbf{Z}} [\log p_{Z_i}] + \text{const} - \frac{1}{2\epsilon} E_{\mathbf{Z}} [\|\mathbf{x}_i - \boldsymbol{\mu}_{Z_i}\|^2]
 \end{aligned}$$

In the M -step we can separately maximize $\{p_j^{t+1}\}$ and $\{\boldsymbol{\mu}_j^{t+1}\}$. Beginning with $\{p_j^{t+1}\}$, we recall that this is a constrained optimization problem. Also, Q decomposes nicely, so we need only solve:

$$\begin{aligned}
 \mathbf{p}^{t+1} &= \arg \max_{\mathbf{p}} \sum_{i=1}^n E_{\mathbf{Z}} [\log p_{Z_i}] \\
 &= \arg \max_{\mathbf{p}} \sum_{i=1}^n \sum_{j=1}^k a_{i,j}^t \log p_j
 \end{aligned}$$

subject to the optimization $\sum_{j=1}^k p_j = 1$. This can be solved with Lagrange multipliers, with the Lagrangian function

$$\mathcal{L}(p_1, \dots, p_k) = \sum_{i=1}^n \sum_{j=1}^k a_{i,j}^t \log p_j - \lambda \left(\sum_{j=1}^k p_j - 1 \right)$$

Solving this gives the solution:

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

For the values of $\boldsymbol{\mu}^{t+1}$ we have

$$\begin{aligned}
 \boldsymbol{\mu}^{t+1} &= \arg \max_{\boldsymbol{\mu}} \sum_{i=1}^n \left(-\frac{1}{2\epsilon} E_{\mathbf{Z}} [\|\mathbf{x}_i - \boldsymbol{\mu}_{Z_i}\|^2] \right) \\
 &= \arg \max_{\boldsymbol{\mu}} \sum_{i=1}^n \left(-\frac{1}{2\epsilon} \sum_{j=1}^k a_{i,j}^t \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right) \\
 &= \arg \min_{\boldsymbol{\mu}} \sum_{i=1}^n \sum_{j=1}^k a_{i,j}^t \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \\
 &= \arg \min_{\boldsymbol{\mu}} F(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k)
 \end{aligned}$$

We need to optimize this for each coordinate of each $\boldsymbol{\mu}_j$. However, using matrix calculus we can write this more simply as a derivative according to a vector:

$$\begin{aligned}
 \frac{\partial F}{\partial \boldsymbol{\mu}_j} &= \frac{\partial}{\partial \boldsymbol{\mu}_j} \left(\sum_{i=1}^n \sum_{j=1}^k a_{i,j}^t \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right) \\
 &= 2 \sum_{i=1}^n a_{i,j}^t (\mathbf{x}_i - \boldsymbol{\mu}_j) = 0 \Rightarrow \\
 \boldsymbol{\mu}_j^{t+1} &= \frac{\sum_{i=1}^n a_{i,j}^t \mathbf{x}_i}{\sum_{i=1}^n a_{i,j}^t}
 \end{aligned}$$

4.2 Back to k -means

Recall the iterative update rules of k -means:

Assign: Set each point to its closest center:

$$C_i^{t+1} = \arg \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j^t\|^2, S_j^{t+1} = \{i | C_i^{t+1} = j\}$$

Update: Minimize sum of distances by re-computing the centers:

$$\boldsymbol{\mu}_j^{t+1} = \frac{\sum_{i \in S_j^{t+1}} \mathbf{x}_i}{|S_j^{t+1}|}$$

Compare this to the iterative update rules of EM in the case of GMM:

E-Step:

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

M-Step:

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

$$\boldsymbol{\mu}_j^{t+1} = \frac{\sum_{i=1}^n a_{i,j}^t \mathbf{x}_i}{\sum_{i=1}^n a_{i,j}^t}$$

We can see that k -means in this case is a limiting case of EM, where the posterior probabilities $a_{i,j}^t$ are either 0 or 1.

Let us see this more formally. At a given iteration t , fix an \mathbf{x}_i and $\boldsymbol{\mu}_j^t$, and examine $a_{i,j}^t$. Suppose that, w.l.o.g., \mathbf{x}_i is closer to the first cluster's centre:

$$\|\mathbf{x}_i - \boldsymbol{\mu}_1^t\|^2 < \|\mathbf{x}_i - \boldsymbol{\mu}_2^t\|^2, \dots, \|\mathbf{x}_i - \boldsymbol{\mu}_k^t\|^2$$

Compare $a_{i,1}^t$ and the other $a_{i,j}^t$ ($j > 1$), as a function of ϵ , e.g.:

$$\begin{aligned} \frac{a_{i,j}^t}{a_{i,1}^t} &= \frac{p_2^t \cdot f_2(\mathbf{x})}{p_1^t \cdot f_1(\mathbf{x})} \\ &= \frac{p_2^t}{p_1^t} \cdot \frac{(\sqrt{2\pi\epsilon})^{-d} \exp\left(-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_2^t\|^2\right)}{(\sqrt{2\pi\epsilon})^{-d} \exp\left(-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_1^t\|^2\right)} \\ &= \frac{p_2^t}{p_1^t} \cdot \exp\left(-\frac{1}{2\epsilon}(\|\mathbf{x} - \boldsymbol{\mu}_2^t\|^2 - \|\mathbf{x} - \boldsymbol{\mu}_1^t\|^2)\right) \end{aligned}$$

When $\epsilon \rightarrow 0$, the ratio quickly converges to 0, while their sum is bounded. Therefore, we will get $a_1^t \rightarrow 1$ and $a_j^t \rightarrow 0$ for $j > 1$.