CS 189 Introduction to Machine Learning
Spring 2018

DIS10

1 One Dimensional Mixture of Two Gaussians

Suppose we have a mixtures of two Gaussians in $\mathbb{R}$ that can be described by a pair of random variables (X, Z) where X takes values in $\mathbb{R}$ and Z takes value in the set $1, 2$. The joint-distribution of the pair (X, Z) is given to us as follows:

$$\begin{aligned} Z &\sim \text{Bernoulli}(\beta), \\ (X|Z = k) &\sim \mathcal{N}(\mu_k, \sigma_k), \quad k \in 1, 2, \end{aligned}$$

We use $\boldsymbol{\theta}$ to denote the set of all parameters $\beta, \mu_1, \sigma_1, \mu_2, \sigma_2$.

(a) Write down the expression for the joint likelihood $p_{\boldsymbol{\theta}}(X = x_i, Z_i = 1)$ and $p_{\boldsymbol{\theta}}(X = x_i, Z_i = 2)$. What is the marginal likelihood $p_{\boldsymbol{\theta}}(X = x_i)$?

Solution:

Joint likelihood:

$$\begin{aligned} p_{\boldsymbol{\theta}}(X = x_i, Z_i = 1) &= p_{\boldsymbol{\theta}}(X = x_i | Z_i = k) p(Z_i = 1) \\ &= \beta \mathcal{N}(x_i | \mu_1, \sigma_1^2) \end{aligned}$$

$$\begin{aligned} p_{\boldsymbol{\theta}}(X = x_i, Z_i = 2) &= p_{\boldsymbol{\theta}}(X = x_i | Z_i = 2) p(Z_i = 2) \\ &= (1 - \beta) \mathcal{N}(x_i | \mu_2, \sigma_2^2) \end{aligned}$$

Marginal likelihood:

$$\begin{aligned} p_{\boldsymbol{\theta}}(X = x_i) &= \sum_k p_{\boldsymbol{\theta}}(X = x_i, Z_i = k) \\ &= \sum_k p_{\boldsymbol{\theta}}(X = x_i | Z_i = k) p(Z_i = k) \\ &= \beta \mathcal{N}(x_i | \mu_1, \sigma_1^2) + (1 - \beta) \mathcal{N}(x_i | \mu_2, \sigma_2^2) \end{aligned}$$

(b) What is the log-likelihood $\ell_{\boldsymbol{\theta}}(\mathbf{x})$? Why is this hard to optimize?

Solution:

Log-likelihood:

$$\ell_{\boldsymbol{\theta}}(\mathbf{x}) = \log\big(p_{\boldsymbol{\theta}}(X = x_1, \ldots, X = x_n)\big)$$

$$= \sum_{i=1}^{n} \log\big(p_{\boldsymbol{\theta}}(X = x_i)\big)$$

$$= \sum_{i=1}^{n} \log\left[\beta\mathcal{N}(x_i|\mu_1, \sigma_1^2) + (1-\beta)\mathcal{N}(x_i|\mu_2, \sigma_2^2)\right]$$

Taking the derivative with respect to μ_1, for example, would give:

$$\frac{\partial \ell_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \mu_1} = \sum_{i=1}^{n} \frac{\beta\mathcal{N}(x_i|\mu_1, \sigma_1^2)}{\beta\mathcal{N}(x_i|\mu_1, \sigma_1^2) + (1-\beta)\mathcal{N}(x_i|\mu_2, \sigma_2^2)} (\frac{x_i - \mu_1}{\sigma_1^2})$$

This ratio of exponentials and linear terms makes it difficult to solve for a maximum likelihood expression. Recall that there is no rule for splitting up $\log(a+b)$ which prevents us from applying the log to the exponential.

(c) (Optional) You'd like to optimize the log likelihood: $\ell_{\boldsymbol{\theta}}(x)$. However, we just saw this can be hard to solve for an MLE closed form solution. Show that a lower bound for the log likelihood is $\ell_{\boldsymbol{\theta}}(x_i) \geq \mathbb{E}_q\left[\log\left(\frac{p_{\boldsymbol{\theta}}(X=x_i, Z_i=k)}{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}\right)\right]$.

Solution:

$$\begin{aligned}
\ell_{\boldsymbol{\theta}}(x_i) &= \log\left(\sum_k p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)\right) && \text{Marginalizing over possible Gaussian labels} \\
&= \log\left(\sum_k \frac{q_{\boldsymbol{\theta}}(Z_i = k|X = x_i) p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)}{q_{\boldsymbol{\theta}}(Z_i = k|X = x_i)}\right) && \text{Introducing arbitrary distribution q} \\
&= \log\left(\mathbb{E}_q\left[\frac{p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)}{q_{\boldsymbol{\theta}}(Z_i = k|X = x_i)}\right]\right) && \text{Rewriting as expectation} \\
&\geq \mathbb{E}_q\left[\log\left(\frac{p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)}{q_{\boldsymbol{\theta}}(Z_i = k|X = x_i)}\right)\right] && \text{Using Jensen's inequality}
\end{aligned}$$

where Jensen's inequality says $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$ for convex function ϕ.

(d) (Optional) The EM algorithm first initially starts with two randomly placed Gaussians (μ_1, σ_1) and (μ_2, σ_2), which are both particular realizations of $\boldsymbol{\theta}$.

- E-step: $\mathbf{q}_{i,k}^{t+1} = p_{\boldsymbol{\theta}}(Z_i = k|X = x_i)$. For each data point, determine which Gaussian generated it, being either (μ_1, σ_1) or (μ_2, σ_2).
- M-step: : $\boldsymbol{\theta}^{t+1} = \text{argmax}_{\theta} \sum_{i=1}^{n} \mathbb{E}_q\left[\log\big(p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)\big)\right]$. After labeling all data-points in the E-step, adjust (μ_1, σ_1) and (μ_2, σ_2).

Why does alternating between the E-step and M-step result in maximizing the lower bound?

Solution: To show the M-step (so-called because we are maximizing with respect to the parameters) is maximizing the lower bound:

$$\mathbb{E}_q\left[\log\left(\frac{p_{\boldsymbol{\theta}}(X=x_i, Z_i=k)}{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}\right)\right]$$
$$=\mathbb{E}_q\left[\log\big(p_{\boldsymbol{\theta}}(X=x_i, Z_i=k)\big)\right]-\mathbb{E}_q\left[\log\big(q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)\big)\right]$$

The M-step is maximizing the first term.

To show the E-step is maximizing the bound we can rewrite the lower bound as:

$$\mathbb{E}_q\left[\log\left(\frac{p_{\boldsymbol{\theta}}(X=x_i)p_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}\right)\right]$$
$$=\mathbb{E}_q\left[\log\big(p_{\boldsymbol{\theta}}(X=x_i)\big)\right]-\mathbb{E}_q\left[\log\left(\frac{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}{p_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}\right)\right]$$

This expression is minimized if the second term is 0, which occurs when $q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)=p(Z_i=k|X=x_i)$.

(e) E-step: What are expressions for probabilistically imputing the classes for all the datapoints, i.e. $q_{i,1}^{t+1}$ and $q_{i,2}^{t+1}$?

Solution:

$$q_{i,1}^{t+1}=P(Z=1|X=x_i;\boldsymbol{\theta}^t)=\frac{P(x_i|Z=1;\boldsymbol{\theta}^t)P(Z=1)}{P(x_i|Z=1;\boldsymbol{\theta}^t)P(Z=1)+P(x_i|Z=2;\boldsymbol{\theta}^t)P(Z=2)}$$
$$q_{i,2}^{t+1}=P(Z=2|X=x_i;\boldsymbol{\theta}^t)=\frac{P(x_i|Z=2;\boldsymbol{\theta}^t)P(Z=2)}{P(x_i|Z=1;\boldsymbol{\theta}^t)P(Z=1)+P(x_i|Z=2;\boldsymbol{\theta}^t)P(Z=2)}$$

where $P(x_i|Z=1)=\frac{1}{\sqrt{2\pi\sigma_1}}exp(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2})$

To be clear, you would have to compute nC such $q_{i,k}$ values at each time step where C is the number of classes. Here, C=2.

(f) What is the expression for μ_1^{t+1} for the M-step?

Solution: From Homework 10, we know that

$$\mu_1^{t+1}=\frac{\sum_{i=1}^n q_{i,1}^{t+1}x_i}{\sum_{i=1}^n q_{i,1}^{t+1}}=\frac{q_{1,1}^{t+1}x_1+q_{2,1}^{t+1}x_2+\cdots+q_{n,1}^{t+1}x_n}{q_{1,1}^{t+1}+q_{2,1}^{t+1}+\cdots+q_{n,1}^{t+1}}$$
$$\mu_2^{t+1}=\frac{\sum_{i=1}^n q_{i,2}^{t+1}x_i}{\sum_{i=1}^n q_{i,2}^{t+1}}=\frac{q_{1,2}^{t+1}x_1+q_{2,2}^{t+1}x_2+\cdots+q_{n,2}^{t+1}x_n}{q_{1,2}^{t+1}+q_{2,2}^{t+1}+\cdots+q_{n,2}^{t+1}}$$
$$(\sigma_1^2)^{(t+1)}=\frac{\sum_{i=1}^n q_{i,1}^{t+1}(x_i-\mu_1^{t+1})^2}{\sum_{i=1}^n q_{i,1}^{t+1}}$$

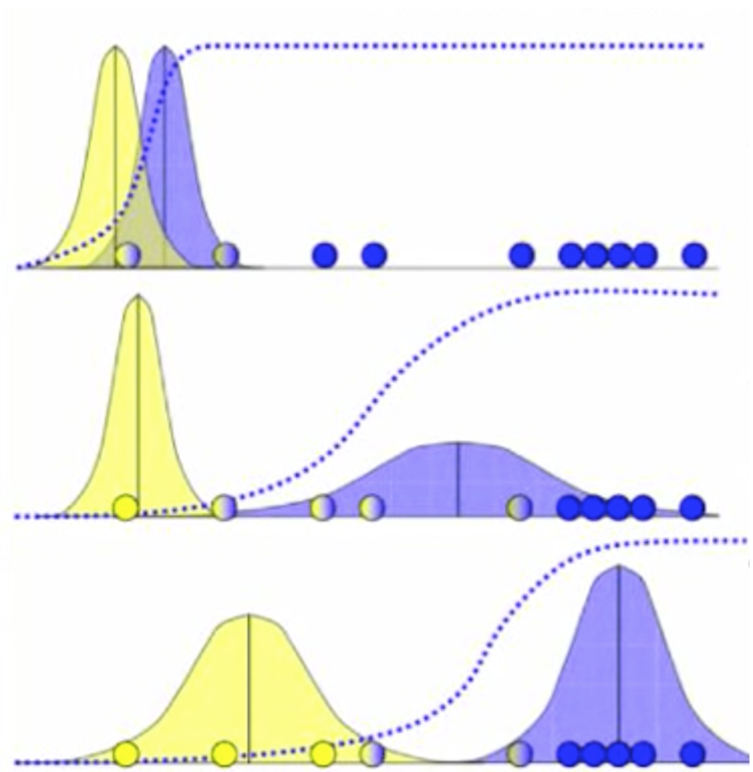

Figure 1: EM examples in 1D for two clusters (yellow and blue). The shadings of the datapoints (circles) indicate the respective estimated probabilities of coming from either the yellow or blue cluster.

$$(\sigma_2^2)^{(t+1)} = \frac{\sum_{i=1}^n q_{i,2}^{t+1}(x_i - \mu_2^{t+1})^2}{\sum_{i=1}^n q_{i,2}^{t+1}}$$

We show how to obtain μ_1^{t+1} as an example:

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}_q \left[\log\big(p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)\big)\right] \\
&= \sum_{i=1}^n \left[q_{i,1}^{t+1} \log\big(\beta \mathcal{N}(x_i|\mu_1, \sigma_1^2)\big) + q_{i,2}^{t+1} \log\left((1-\beta)\mathcal{N}(x_i|\mu_2, \sigma_2^2)\right)\right] \\
&= \sum_{i=1}^n \left[q_{i,1}^{t+1}\left(\log(\beta) - \frac{(x_i-\mu_1)^2}{2\sigma_1^2} - \log(\sigma_1)\right) + q_{i,2}^{t+1}\left(\log(1-\beta) - \frac{(x_i-\mu_2)^2}{2\sigma_2^2} - \log(\sigma_2)\right)\right] + \text{cons}
\end{aligned}$$

Taking a derivative with respect to μ_1 and setting to 0 to obtain the maximum gives:

$$\begin{aligned}
\sum_{i=1}^n q_{i,1}^{t+1}\left(\frac{(x_i - \mu_1)}{\sigma_1^2}\right) &= 0 \\
\sum_{i=1}^n q_{i,1}^{t+1} x_i - \sum_{i=1}^n q_{i,1}^{t+1}\mu_1 &= 0 \\
\mu_1 &= \frac{\sum_{i=1}^n q_{i,1}^{t+1} x_i}{\sum_{i=1}^n q_{i,1}^{t+1}}
\end{aligned}$$

(g) Compare and contrast k-means, soft k-means, and mixture of Gaussians fit with EM.

Solution: For k-means, we implicitly assume clusters are spherical and so this doesn't work for complex geometrical shaped data. Additionally, it uses hard assignment, meaning the $q_{i,1}$ probabilities are 0 or 1. This can be easier to interpret, but doesn't incorporate information from all data points to update each centroid. K-means will also usually have trouble with clusters that have large overlap (see Figure 2)

For soft k-means and EM we have soft assignments. For soft k-means, the weighted mean amounts to

$$\begin{aligned}
r_{i,1} &= \frac{\exp\left\{-B||x_i - \mu_1||^2\right\}}{\exp\left\{-B||x_i - \mu_1||^2\right\} + \exp\left\{-B||x_i - \mu_2||^2\right\}} \\
r_{i,2} &= \frac{\exp\left\{-B||x_i - \mu_2||^2\right\}}{\exp\left\{-B||x_i - \mu_1||^2\right\} + \exp\left\{-B||x_i - \mu_2||^2\right\}}
\end{aligned}$$

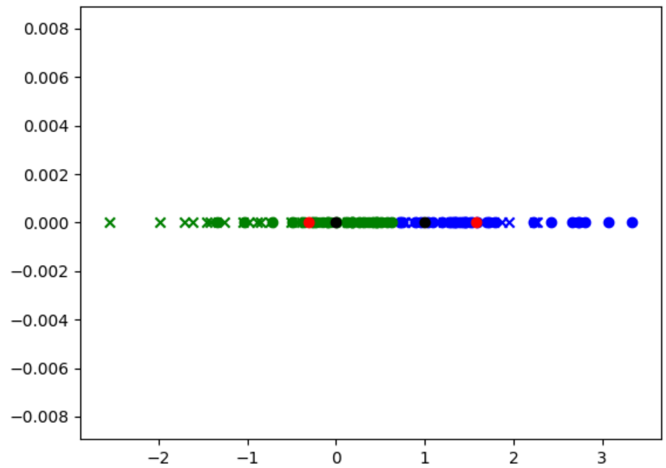

Figure 2: K-means for two clusters in 1D. 'x' points indicate coming from the μ_1 while 'o' indicates points coming from μ_2. The colors blue and green indicate the predicted clustering. Black dots indicate the true means, while red indicates the predicted means.

$$\mu_1^{t+1} = \frac{\sum_{i=1}^n r_{i,1}^{t+1} x_i}{\sum_{i=1}^n r_{i,1}^{t+1}}$$

$$\mu_2^{t+1} = \frac{\sum_{i=1}^n r_{i,2}^{t+1} x_i}{\sum_{i=1}^n r_{i,2}^{t+1}}$$

where we have a stiffness parameter β, which can be intrepreted as the inverse variance. In cases where the clusters have different geometry, one might resort to EM. Note that EM is not unrelated to LDA/QDA. The setup is similar in that we probablistically determine the probabilities of coming from cluster k, but LDA/QDA does hard classification, EM probabilistic performs soft classification.

