CSCI 590: Machine Learning

Lecture 21: EM for Mixture of Gaussians

Instructor: Murat Dundar

Acknowledgement:

1. PRML by Chris Bishop

Mixtures of Gaussians (1)

The Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

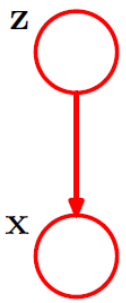
Let us introduce a K-dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0.

The values of z_k therefore satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$

We shall define the joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of a marginal distribution $p(\mathbf{z})$ and a conditional distribution $p(\mathbf{x}|\mathbf{z})$

Mixtures of Gaussians (2)

Graphical representation of a mixture model, in which the joint distribution is expressed in the form $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.



The marginal distribution over **z** is specified in terms of the mixing coefficients π_k, such that

$$p(z_k = 1) = \pi_k$$

where the parameters $\{\pi_k\}$ must satisfy

$$0 \leqslant \pi_k \leqslant 1$$

together with

$$\sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Gaussians (3)

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{9.10}$$

Similarly, the conditional distribution of $\mathbf{x}$ given a particular value for $\mathbf{z}$ is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{9.11}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

An equivalent formulation of the Gaussian mixture with a latent variable.

Mixtures of Gaussians (4)

Conditional probability of z given x:

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

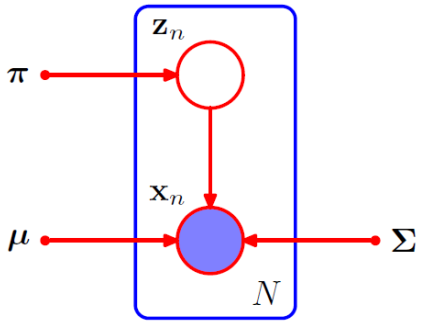
Mixtures of Gaussians (5)

Maximum Likelihood: Suppose we have a data set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and we wish to model this data using a mixture of Gaussians.

We can represent this data as a matrix X in which the nth row is given by x_n^T. Similarly, the corresponding latent variables will be denoted by an N x K matrix Z.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{\mathbf{x}_n\}$, with corresponding latent points $\{\mathbf{z}_n\}$, where $n = 1, \ldots, N$.



Mixtures of Gaussians (6)

Recall: $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

Suppose that in addition to the observed data set X, we were also given the values of the corresponding discrete variables Z. Consider the problem of maximizing the likelihood for the complete data set {X,Z}. From 9.10 and 9.11 this likelihood takes the form

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

where z_{nk} denotes the k_{th} component of z_n. Taking the logarithm, we obtain

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

Mixtures of Gaussians (7)

Recall: $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

If Z is known the complete-data loglikelihood function can be maximized trivially to obtain μ, π, Σ for each component.

In practice, however, we don't have values for the latent variables so, we consider the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables.

Using (9.10) and (9.11) together with Bayes' theorem, we see that this posterior distribution takes the form

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} . \tag{9.38}$$

EM Algorithm in General (1)

$$\log p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) = \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) + \log p_{\boldsymbol{X}}(\boldsymbol{x})$$

$$\log p_{\boldsymbol{X}}(\boldsymbol{x}) = \log p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) - \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})$$

$$\log a - \log b = \log \frac{a}{c} - \log \frac{b}{c}$$

$$\log p_{\boldsymbol{X}}(\boldsymbol{x}) = \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} - \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})},$$

and multiplying both sides by $q_{\boldsymbol{Z}}(\boldsymbol{z})$ we obtain

$$q_{\boldsymbol{Z}}(\boldsymbol{z}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) = q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} - q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}$$

EM Algorithm in General (2)

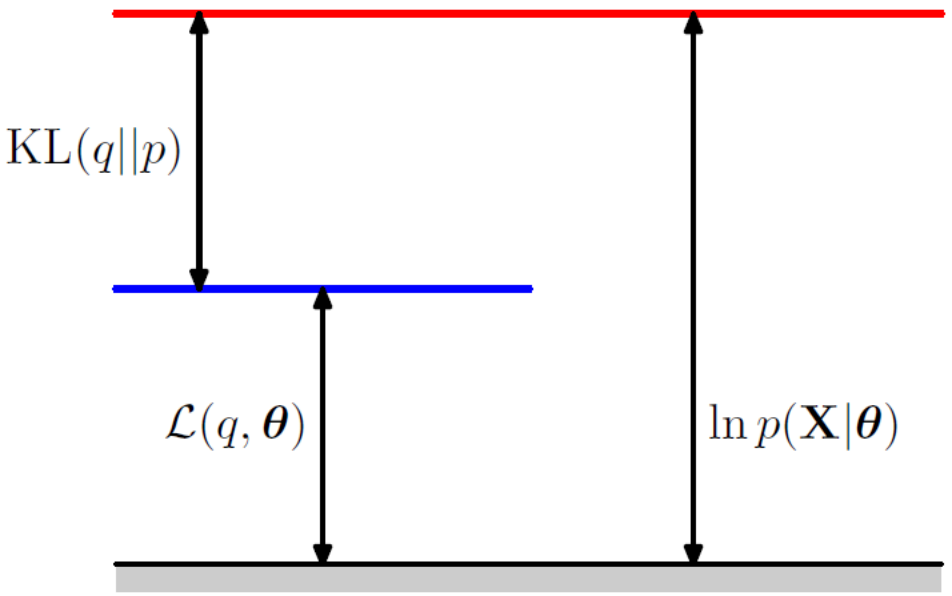
By integrating with respect to $\boldsymbol{z}$, we have

$$\int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log p_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{z} = \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\, d\boldsymbol{z} - \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\, d\boldsymbol{z}$$

$$\begin{aligned}
\mathcal{L}(q_{\boldsymbol{Z}}) &:= \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\, d\boldsymbol{z} \\
&= \mathbb{E}_{q_{\boldsymbol{Z}}}\left[\log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{X},\boldsymbol{Z})}{q_{\boldsymbol{Z}}(\boldsymbol{Z})}\right] \\
\mathrm{KL}\left(q_{\boldsymbol{Z}} || p_{\boldsymbol{Z}|\boldsymbol{X}}\right) &:= -\int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\, d\boldsymbol{z} \\
&= \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{q_{\boldsymbol{Z}}(\boldsymbol{z})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}\, d\boldsymbol{z} \\
&= \mathbb{E}_{q_{\boldsymbol{Z}}}\left[\log \frac{q_{\boldsymbol{Z}}(\boldsymbol{Z})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{X})}\right],
\end{aligned}$$

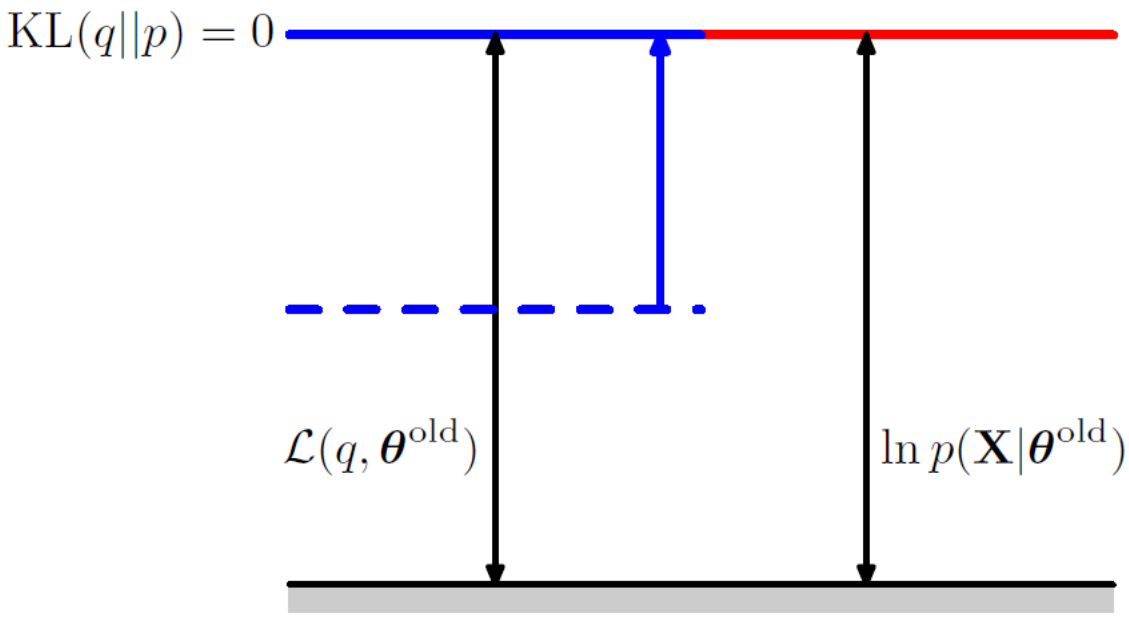
EM Algorithm in General (3)

Illustration of the decomposition given by (9.70), which holds for any choice of distribution $q(\mathbf{Z})$. Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$, we see that the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.



EM Algorithm in General (4)

Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\text{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.



EM Algorithm in General (5)

Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.