Ch 9: Mixture Models and EM

ifding

Abstract

K-means Clustering, Mixtures of Gaussians

If we define a joint distribution over observed and latent variables, the corresponding distribution of the observed variables alone is obtained by marginalization.

1. K-means Clustering

Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ consisting of N observations of a random D-dimensional Euclidean variable $\mathbf{x}$. Our goal is to partition the data set into some number K of clusters. A set of $D-$dimensional vectors $\boldsymbol{\mu}_k$, where $k = 1, \ldots, K$, in which $\boldsymbol{\mu}_k$ is a prototype associated with the k^{th} cluster.

For each data point $\mathbf{x}_n$, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, \ldots, K$ describing which of the K clusters the data point $\mathbf{x}_n$ is assigned to, so that if data point $\mathbf{x}_n$ is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \boldsymbol{\mu}_k \right\|^2 \tag{1}$$

which represents the sum of the squares of the distances of each data point to its assigned vector $\boldsymbol{\mu}_k$. Our goal is to find values for the $\{r_{nk}\}$ and the $\{\boldsymbol{\mu}_k\}$ so as to minimize J.

We can do this through an iterative procedure in which each iteration involves two successive steps. First we choose some initial values for the $\boldsymbol{\mu}_k$. Then in the first phase we minimize J with respect to the r_{nk}, keeping the $\boldsymbol{\mu}_k$ fixed. In the second phase we minimize J with respect to the $\boldsymbol{\mu}_k$, keeping r_{nk} fixed. This two-stage optimization is then repeated until convergence. These two stages of updating r_{nk} and updating $\boldsymbol{\mu}_k$ correspond respectively to the E (expectation) and M (maximization) steps of the EM algorithm.

Consider first the determination of the r_{nk}, we simply assign the n^{th} data point to the closest cluster centre.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \left\| \mathbf{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Now consider the optimization of the $\boldsymbol{\mu}_k$ with the r_{nk} held fixed. Set its derivative with respect to $\boldsymbol{\mu}_k$ to zero giving

$$2 \sum_{n=1}^{N} r_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_k \right) = 0 \tag{3}$$

which we can easily solve for $\boldsymbol{\mu}_k$ to give

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \tag{4}$$

The denominator in this expression is equal to the number of points assigned to cluster k, and set $\boldsymbol{\mu}_k$ equal to the mean of all of the data points $\mathbf{x}_n$ assigned to cluster k.

2. Mixtures of Gaussians

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

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N} \left(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right) \tag{5}$$

Let us introduce a $K-$dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation, the values of z_k satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. There are K possible states for the vector $\mathbf{z}$ according to which element is nonzero.

We shall define the joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of a marginal distribution $p(\mathbf{z})$ and a conditional distribution $p(\mathbf{x}|\mathbf{z})$, corresponding to the graphical model in Figure 1.

Figure 1: Graphical representation of a mixture model, $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.

The marginal distribution over $\mathbf{z}$ is specified in terms of the mixing coefficients π_k, such that

$$p\left(z_k = 1\right) = \pi_k \tag{6}$$

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

$$0 \leqslant \pi_k \leqslant 1 \tag{7}$$

together with

$$\sum_{k=1}^{K} \pi_k = 1 \tag{8}$$

in order to be valid probabilities. Because $\mathbf{z}$ uses a 1-of-K representation, we can write the distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{9}$$

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

$$p\left(\mathbf{x}|z_k = 1\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{10}$$

which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_k} \tag{11}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{z_k} \tag{12}$$

Exploiting the 1-of-K representation for $\mathbf{z}$, we can re-write the r.h.s. as

$$\sum_{j=1}^{K} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{I_{kj}} = \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{13}$$

where $\mathbf{I}_{kj} = 1$ if $k = j$ and 0 otherwise. The marginal distribution of $\mathbf{x}$ is a Gaussian mixture.

The conditional probability of $\mathbf{z}$ given $\mathbf{x}$, we shall use $\gamma(z_k)$ to denote $p(z_k = 1|\mathbf{x})$,

$$\begin{aligned} \gamma\left(z_k\right) \equiv p\left(z_k = 1|\mathbf{x}\right) =& \frac{p\left(z_k = 1\right) p\left(\mathbf{x}|z_k = 1\right)}{\sum_{j=1}^{K} p\left(z_j = 1\right) p\left(\mathbf{x}|z_j = 1\right)} \\ =& \frac{\pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \end{aligned} \tag{14}$$

We shall view π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed $\mathbf{x}$. $\gamma(z_k)$ can also be viewed as the *responsibility* that component k takes for 'explaining' the observation $\mathbf{x}$.

2.1. 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 set as an $N \times D$ matrix $\mathbf{X}$ in which the n^{th} rows is given by $\mathbf{x}_n^{\mathrm{T}}$. The corresponding latent variables will be denoted by an $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^{\mathrm{T}}$. The log of likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\} \tag{15}$$

The presence of **singularities**, for simplicity, consider a Gaussian mixture whose components have covariance matrices given by $\sum_k = \sigma^2 \mathbf{I}$, which $\mathbf{I}$ is the unit matrix. Suppose that one of the components of the mixture model, the j^{th} component, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some value of n.

$$\mathcal{N}\left(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2 \mathbf{I}\right) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j} \tag{16}$$

If we consider the limit limit $\sigma_j \to 0$, the log likelihood function will go to infinity. Thus the maximization of the log likelihood function is not a well posed problem because such singularities will always be present and will occur whenever one of the Gaussian components 'collapses' onto a specific data point.

2.2. EM for Gaussian mixtures

Let us begin by writing down the conditions that must be satisfied at a maximum of the likelihood function. Setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the

mean $\boldsymbol{\mu}_k$ to zero,

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \tag{17}$$

Multiplying by Σ_k^{-1} and rearranging we obtain

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n \tag{18}$$

where we have defined

$$N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \tag{19}$$

We can interpret N_k as the effective number of points assigned to cluster k. The mean $\boldsymbol{\mu}_k$ for the k^{th} Gaussian component is obtained by taking a weighted mean of all of the points in the data set, in which the weighting factor is given by the posterior probability $\gamma(z_{nk})$ that component k was responsible for generating $\mathbf{x}_n$.

If we set the derivative of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}_k$ to zero,

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \tag{20}$$

Finally, we maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k, and take account of the constraint that the mixing coefficients sum to one.

$$\pi_k = \frac{N_k}{N} \tag{21}$$

so that the mixing coefficient for the k^{th} component is given by the average responsibility which that component takes for explaining the data points.

2.3. EM algorithm

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $boldsymbolSigma_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.

2. **E step.** Evaluate the responsibilities using the current parameter values

$$\gamma\left(z_{nk}\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \tag{22}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n \tag{23}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{24}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{25}$$

where

$$N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \tag{26}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\} \tag{27}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

3. An Alternative View of EM

The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables $\mathbf{Z}$, the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\} \tag{28}$$

Note that the discussion will apply equally well to continuous latent variables simply replacing the sum over $\mathbf{Z}$ with an integral.

Suppose that, for each observation in $\mathbf{X}$, we were told the corresponding value of the latent variable $\mathbf{Z}$, $\{\mathbf{X}, \mathbf{Z}\}$ is called as the complete data set. In practice, we are only given the incomplete data $\mathbf{X}$. The values of the latent variables in $\mathbf{Z}$ is given by the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$. Because we cannot use the complete data log likelihood $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, we consider instead its expected value under the posterior distribution of the latent variable, which corresponds to the E step.

In the E step, we use the current parameter values $\boldsymbol{\theta}^{old}$ to find the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$. We then use this posterior distribution to find the expectation of the

complete data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$.

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \sum_{\mathbf{Z}} p\left(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{29}$$

In the M step, we determine the revised parameter estimate $\boldsymbol{\theta}^{new}$ by maximizing this function

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) \tag{30}$$

3.1. Gaussian mixtures revisited

Suppose then that in addition to the observed data set $\mathbf{X}$, we were also given the values of the corresponding discrete variables $\mathbf{Z}$. The graphical model for the complete data is shown in Figure 2.

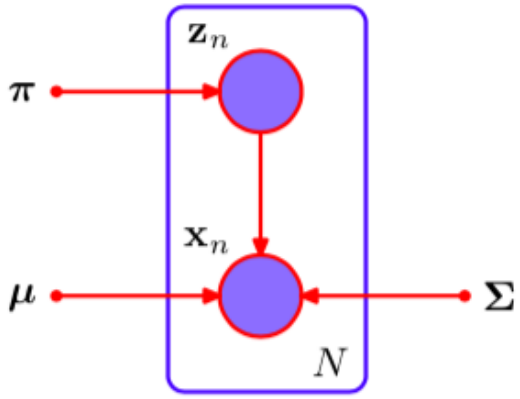

Figure 2: The discrete variables z_n are observed, as well as the data variables $\mathbf{x}_n$.

Now consider the problem of maximizing the likelihood for the complete data set $\{\mathbf{X}, \mathbf{Z}\}$.

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_{nk}} \tag{31}$$

where z_{nk} denotes the k^{th} component of $\mathbf{z}_n$. Taking the logarithm

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\} \tag{32}$$

Consider first the maximization with respect to the means and covariances. Because $\mathbf{z}_n$ is a K-dimensional vector with all elements equal to 0 except for a single element having the value 1, the complete data log likelihood function is simply a sum of K independent contributions, one for each mixture component. Thus the maximization with respect to a mean or a covariance is exactly as for a single Gaussian, except that it involves only the subset of data points that are 'assigned' to that component. For the maximization with respect to the mixing coefficients, this can be enforced using a Lagrange multiplier,

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk} \tag{33}$$

the mixing coefficients are equal to the fractions of data points assigned to the corresponding components.

The complete-data log likelihood function can be maximized trivially in closed form. In practice, however, we do not have values for the latent variables. We consider the expectation, with respect to the posterior distribution of the latent variables, of the complete-data log likelihood. The posterior distribution takes the form

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right]^{z_{nk}} \tag{34}$$

To prove it, we only need to prove $p(\mathbf{Z}|\mathbf{X})$ can be written as the product of $p(\mathbf{z}_n|\mathbf{x}_n)$. Notice that the condition on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ can be omitted.

$$p(\mathbf{X}, \mathbf{Z}) = p\left(\mathbf{x}_1, \mathbf{z}_1\right) \ldots p\left(\mathbf{x}_N, \mathbf{z}_N\right) \tag{35}$$

Since there is no link from $\mathbf{z}_m$ to $\mathbf{z}_n$, from $\mathbf{x}_m$ to $\mathbf{x}_n$, and from $\mathbf{z}_m$ to $\mathbf{x}_n$ (m $\neq$ n),

$$p(\mathbf{Z}) = p\left(\mathbf{z}_1\right) \ldots p\left(\mathbf{z}_N\right), \quad p(\mathbf{X}) = p\left(\mathbf{x}_1\right) \ldots p\left(\mathbf{x}_N\right) \tag{36}$$

The marginal distribution from $p(\mathbf{X}, \mathbf{Z})$

$$\begin{aligned} p(\mathbf{Z}) = \sum_{\mathbf{X}} p(\mathbf{X}, \mathbf{Z}) &= \sum_{\mathbf{x}_1, \ldots, \mathbf{x}_N} p\left(\mathbf{x}_1, \mathbf{z}_1\right) \ldots p\left(\mathbf{x}_N, \mathbf{z}_N\right) \\ &= p\left(\mathbf{z}_1\right) \ldots p\left(\mathbf{z}_N\right) \end{aligned} \tag{37}$$

According to Bayes' Theorem, we have

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X}) &= \frac{p(\mathbf{X}|\mathbf{Z}) p(\mathbf{Z})}{p(\mathbf{X})} \\ &= \frac{\left[\prod_{n=1}^{N} p\left(\mathbf{x}_n|\mathbf{z}_n\right)\right] \left[\prod_{n=1}^{N} p\left(\mathbf{z}_n\right)\right]}{\prod_{n=1}^{N} p\left(\mathbf{x}_n\right)} \\ &= \prod_{n=1}^{N} \frac{p\left(\mathbf{x}_n|\mathbf{z}_n\right) p\left(\mathbf{z}_n\right)}{p\left(\mathbf{x}_n\right)} \\ &= \prod_{n=1}^{N} p\left(\mathbf{z}_n|\mathbf{x}_n\right) \end{aligned} \tag{38}$$

3.2. Relation to K-means

Whereas the K-means algorithm performs a *hard* assignment of data points to clusters, in which each data point is associated uniquely with one cluster, the EM algorithm makes a *soft* assignment based on the posterior probabilities.

Consider a Gassian mixture model in which the covariance matrices of the mixture components are given by $\epsilon\mathbf{I}$, where ϵ is a variance parameter that is shared by all of the components, and $\mathbf{I}$ is the identify matrix, so that

$$p\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\left\|\mathbf{x}-\boldsymbol{\mu}_k\right\|^2\right\} \tag{39}$$

We now consider the EM algorithm for a mixture of K gaussians of this form in which we treat ϵ as a fixed constant. The posterior probabilities, or responsibilities, for a particular data point $\mathbf{x}_n$, are given by

$$\gamma\left(z_{nk}\right) = \frac{\pi_k \exp\left\{-\left\|\mathbf{x}_n-\boldsymbol{\mu}_k\right\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\left\|\mathbf{x}_n-\boldsymbol{\mu}_j\right\|^2/2\epsilon\right\}} \tag{40}$$

If we consider the limit $\epsilon \to 0$, in the denominator the term for which $\left\|\mathbf{x}_n-\boldsymbol{\mu}_j\right\|^2$ is smallest will go to zero most slowly, and hence the responsibilities $\gamma\left(z_{nk}\right)$ for the data point $\mathbf{x}_n$ all go to zero except for term j, for which the responsibility $\gamma\left(z_{nj}\right)$ will go to 1. In this limit, we obtain a hard assignment of data points to clusters, just as in the K-means algorithm. Finally, in the limit $\epsilon \to 0$ the expected complete-data likelihood becomes

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|\mathbf{x}_n-\boldsymbol{\mu}_k\right\|^2 + \text{const.} \tag{41}$$

Thus, maximizing the expected complete-data log likelihood is equivalent to minimizing the distortion measure J for the K-means algorithm.

4. The EM Algorithm in General

Consider a probabilistic model in which we collectively denote all of the observed variables by $\mathbf{X}$ and all of the hidden variables by $\mathbf{Z}$. The joint distribution $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ is governed by a set of parameters denoted $\boldsymbol{\theta}$. Our goal is to maximize the likelihood function that is given by

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) \tag{42}$$

Here we are assuming $\mathbf{Z}$ is discrete, although the discussion is identical if $\mathbf{Z}$ comprises continuous variables.

We shall suppose that direct optimization of $p(\mathbf{X}|\boldsymbol{\theta})$ is difficult, but that optimization of $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ is easier. Next we introduce a distribution $q(\mathbf{Z})$ defined over the latent variables,

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{43}$$

where we have defined

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}\end{aligned} \tag{44}$$

$\mathcal{L}(q,\boldsymbol{\theta})$ contains the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ while $\mathrm{KL}(q\|p)$ contains the conditional distribution of given $\mathbf{X}$. To verify the decomposition, we first make use of the product rule of probability to give

$$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta}) \tag{45}$$

which we then substitute into expression of $\mathcal{L}(q,\boldsymbol{\theta})$. This fives rise to two terms, one of which cancels $\mathrm{KL}(q|p)$ while the other gives the required log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ after nothing that $q(\mathbf{Z})$ is a normalized distribution that sums to 1.

The KL divergence satisfies $\mathrm{KL}(q||p) \geq 0$, with equality if, and only if, $q(\mathbf{Z}) = p((\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$. $\mathcal{L}(q,\boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$, in other words that $\mathcal{L}(q,\boldsymbol{\theta})$ is a lower bound on $\ln p(\mathbf{X}|\boldsymbol{\theta})$. The decomposition (Eqn.43) is illustrated in Figure 3.

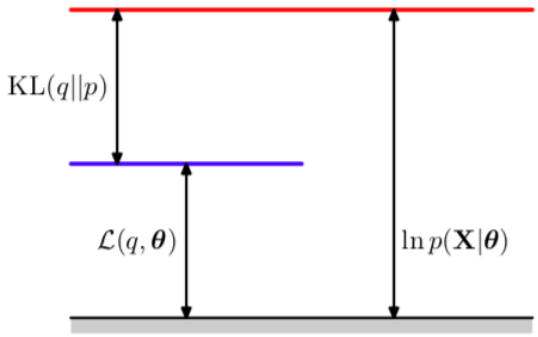

Figure 3: Illustration of the decomposition given by Eqn. 43, which holds for any choice of distribution $q(\mathbf{Z})$.

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions. Suppose that the current value of the parameter vector is $\boldsymbol{\theta}^{\text{old}}$.

In the E step, the lower bound $\mathcal{L}(q,\boldsymbol{\theta}^{\text{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\text{old}}$ fixed. The value

of $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ does not depend on $q(\mathbf{Z})$ and so the largest value of $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ will occur when the KL divergence vanishes, in other words when $q(\mathbf{Z})$ is equal to the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$. In this case, the lower bound will equal the log likelihood, as illustrated in Figure 4.

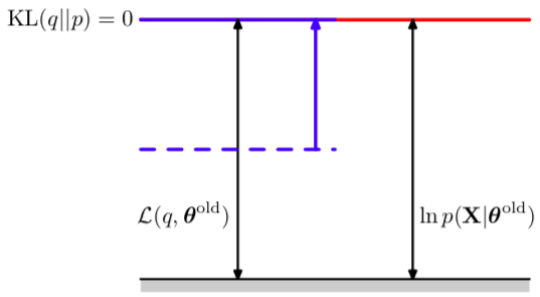

Figure 4: Illustration of the E step of the EM algorithm.

In the subsequent M step, the distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$. This will cause the lower bound $\mathcal{L}$ to increase, which will necessarily cause the corresponding log likelihood function to increase. Because the distribution q is determined using the old parameter values rather than the new values and is held fixed during the M step, it will not equal the new posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})$, and hence there will be a nonzero KL divergence.

References

[1] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.