

# PRML Note

## C09 Mixture Models and EM

Yang Zhao

Department of Automation, Tsinghua University

### 1 K-means Clustering

- Suppose we have a data set  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  and our goal is to partition the data set into some number  $K$  of clusters, where we shall suppose for the moment that the value of  $K$  is given.
- We might think of a cluster as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster. To do this, we introduce a set of  $D$ -dimensional vectors  $\boldsymbol{\mu}_k$ , where  $k = 1, \dots, K$ , in which  $\boldsymbol{\mu}_k$  is a prototype associated with the  $k^{th}$  cluster.
- For each data point, we introduce a corresponding set of binary indicator variables  $r_{nk} \in \{0, 1\}$ , where  $k = 1, \dots, K$  describing which of  $K$  clusters the data point  $\mathbf{x}_n$  is designed to, so that if data point  $\mathbf{x}_n$  is designed to cluster  $k$ , then  $r_{nk} = 1$  and  $r_{nj} = 0$  for  $j \neq k$ . This is known as the 1-of- $K$  coding scheme. We can define an objective function, sometimes called a *distortion measure*, given by

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

Our goal is to find values for the  $r_{nk}$  and the  $\boldsymbol{\mu}_k$  so as to minimize  $J$ .

- 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 the  $r_{nk}$  fixed. This two-stage optimization is then repeated until convergence.
- Consider first the determination of the  $r_{nk}$ . We can simply assign the  $n^{th}$  data point to the closest cluster centre.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- Consider the optimization of the  $\boldsymbol{\mu}_k$  with the  $r_{nk}$  held fixed. The objective function can be minimized by setting its derivative with respect to  $\boldsymbol{\mu}_k$  to zero giving

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \quad (3)$$

- This  $K$ -means algorithm may converge to a local rather than global minimum of  $J$ .
- In practice, a better initialization procedure would be to choose the cluster centres  $\boldsymbol{\mu}_k$  to be equal to a random subset of  $K$  data points.
- It is worth noting that the  $K$ -means algorithm itself is often used to initialize the parameters in a Gaussian mixture model before applying the EM algorithm.
- The  $K$ -means algorithm can be generalized by introducing a more general dissimilarity measure  $\nu(\mathbf{x}, \mathbf{x}')$  instead of the Euclidean distance, which gives the  $K$ -medoids algorithm.

## 2 Mixtures of Gaussians

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

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

Let us introduce a  $K$ -dimensional binary random variable  $\mathbf{z}$  to represent the state of the Gaussian distribution using 1-of- $K$  code.

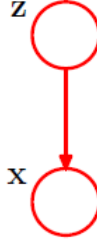


Figure 1: Graphical representation of a mixture model

- 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  $\mathbf{z}$  is specified in terms of the mixing coefficients  $\pi_k$ , such that

$$p(z_k = 1) = \pi_k \quad (5)$$

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

$$p(\mathbf{x} | z_k = 1) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (6)$$

So the joint distribution is given by

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (7)$$

- We use  $\gamma(z_k)$  to denote  $p(z_k = 1 | \mathbf{x})$ , whose value can be found using Bayes' theorem

$$\gamma(z_k) \equiv p(z_k = 1 | \mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (8)$$

the  $\gamma(z_k)$  can be viewed as the *responsibility* that component  $k$  takes for explaining the observation  $\mathbf{x}$ .  $\gamma(z_{nk}) \equiv p(z_k = 1 | \mathbf{x}_n)$ .

- Suppose we have a data set of observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ . The log of the likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \quad (9)$$

It is worth emphasizing that there is a significant problem associated with the maximum likelihood framework applied to Gaussian mixture models, due to the presence of singularities.

- These singularities provide another example of the severe over-fitting that can occur in a maximum likelihood approach. If we adopt a Bayesian approach, this difficulty does not occur.
- *EM algorithm.* The expectation-maximization algorithm is an elegant and powerful method for finding maximum likelihood solutions for models with latent variables. The EM algorithm can be generalized to obtain the variational inference framework.
- Setting the derivatives of  $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  in equation (9) with respect to the  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}_k$  to zero, we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \quad (10)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \quad (11)$$

where we have defined

$$N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (12)$$

- If we want to maximize  $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  with respect to the  $\pi_k$ , we must take account of the constraint  $\sum_{k=1}^K \pi_k = 1$ . This can be achieved using a Lagrange multiplier and maximize the following quantity

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right) \quad (13)$$

which gives

$$0 = \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \quad (14)$$

If we multiply both sides by  $\pi_k$  and sum over  $k$  making use of the constraint  $\sum_{k=1}^K \pi_k = 1$ , we can find  $\lambda = -N$  and we can obtain

$$\pi_k = \frac{N_k}{N} \quad (15)$$

- For EM algorithm, in the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibility given by equation (8); in the maximization step, or M step, we use the probabilities to re-estimate the means, covariances and mixing coefficients using the equation (10), (11), (15).

### 3 An Alternative View of EM

- The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables. We denote the set of all observed data by  $\mathbf{X}$ , in which the  $n^{th}$  row represents  $\mathbf{x}_n^T$ , and similarly we denote the set of all latent variables by  $\mathbf{Z}$ , with a corresponding row  $\mathbf{z}_n^T$ . The set of all model parameters is denoted by  $\boldsymbol{\theta}$ , and so the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\} \quad (16)$$

- The presence of the sum prevents the logarithm from acting directly on the joint distribution, resulting in complicated expressions for the maximum likelihood solution.
- *The General EM Algorithm.*

1. Choose an initial setting for the parameters  $\boldsymbol{\theta}^{old}$ .
2. **E Step.** Evaluate  $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ .
3. **M Step.** Evaluate  $\boldsymbol{\theta}^{new}$  given by

$$\boldsymbol{\theta}^{new} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

where

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

4. Check for convergence of either the log likelihood or the parameters values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

and return to step 2.

- The EM algorithm can also be used to find MAP solutions for models in which a prior  $p(\boldsymbol{\theta})$  is defined over the parameters. In this case the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + \ln p(\boldsymbol{\theta})$ . Suitable choices for the prior will remove the singularities.
- Consider the problem of maximizing the likelihood for the complete data set  $\{\mathbf{X}, \mathbf{Z}\}$ . This likelihood function takes the form

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \quad (17)$$

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

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \} \quad (18)$$

Comparison with the log likelihood function (9) for the incomplete data shows that summation over  $k$  and the logarithm have been interchanged.

## 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  $\boldsymbol{\theta}$ .
- We shall suppose that direct optimization of  $p(\mathbf{X}|\boldsymbol{\theta})$  is difficult, but that optimization of the complete-data likelihood function  $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$  is significantly easier.
- For any choice of  $q(\mathbf{Z})$ , the following decomposition holds

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \quad (19)$$

where we have defined

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \quad (20)$$

$$KL(q||p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \quad (21)$$

Because the Kullback-Leibler divergence satisfies  $KL(q||p) \geq 0$ , we see that the quantity  $\mathcal{L}(q, \boldsymbol{\theta})$  is a lower bound on the log likelihood function  $\ln p(\mathbf{X}|\boldsymbol{\theta})$ .

- The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions.
  1. In the **E** step, the lower bound  $\mathcal{L}(q, \boldsymbol{\theta}^{old})$  is maximized with respect to  $q(\mathbf{Z})$  while holding  $\boldsymbol{\theta}^{old}$  fixed when  $q(\mathbf{Z})$  is equals to  $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ .
  2. 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  $\boldsymbol{\theta}^{new}$ .

## 5 Appendix