PRML Note
C09 Mixture Models and EM

Yang Zhao

Department of Automation, Tsinghua University

1 K-means Clustering

- Suppose we have a data set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ 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 $k = 1, \cdots, K$, 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 $k = 1, \cdots, K$ describing which of K clusters the data point $\boldsymbol{x}_n$ is designed to, so that if data point $\boldsymbol{x}_n$ 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

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

  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 detemination of the r_{nk}. We can simply assign the n^{th} data point to the closest cluster centre.

$$r_{nk} = \begin{cases} 1 & if\ k = argmin_j \|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & otherwise \end{cases} \tag{2}$$

- 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

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}} \tag{3}$$

- 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 $\nu(\boldsymbol{x}, \boldsymbol{x}')$ 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

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

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

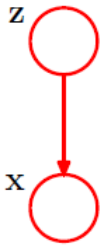

Figure 1: Graphical representation of a mixture model

- The joint distribution is expressed in the form $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$. The marginal distribution over $\boldsymbol{z}$ is specified in terms of the mixing coefficients π_k, such that

$$p(z_k = 1) = \pi_k \tag{5}$$

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

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

  So the joint distribution is given by

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

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

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

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

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

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

  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 $lnp(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ in equation (9) with respect to the $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to zero, we obtain

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

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

  where we have defined

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

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

$$lnp(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\Big(\sum_{k=1}^{K} \pi_k - 1\Big) \tag{13}$$

  which gives

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

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

$$\pi_k = \frac{N_k}{N} \tag{15}$$

- 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, corvariances 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 $\boldsymbol{x}_n^T$, and similarly we denote the set of all latent variables by $\mathbf{Z}$, with a corresponding row $\boldsymbol{z}_n^T$. The set of all model parameters is denoted by $\boldsymbol{\theta}$, and so the log likelihood function is given by

$$lnp(\mathbf{X}|\boldsymbol{\theta}) = ln\Big\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\Big\} \tag{16}$$