EQ2415 – Machine Learning and Data Science
HT22

Tutorial 8 A. Honoré, A. Ghosh

1 Graphical models – EM

Chapter 9 - EM (Bishop).

1.1 K-means

Suppose $N \in \mathbb{N}_*$ data points $\mathbf{x}_n \in \mathbb{R}^d$ for $n = 1, \cdots, N$. Suppose we can measure distances in $\mathbb{R}^d$ with a bivariate function d, e.g. $d(\mathbf{x}_n, \mathbf{x}_{n'}) = ||\mathbf{x}_n - \mathbf{x}_{n'}||_2$. We want to assign all our data points to one of $K \in \mathbb{N}_*$ clusters, characterized by their means $\boldsymbol{\mu}_k \in \mathbb{R}^d$ for $k = 1, \cdots, K$. We use the notation $r_{nk} \in \{0, 1\}$, where $r_{nk} = 1$ and $r_{nk'} = 0$ for $k' \neq k$ if point n is assigned to the k-th cluster.

The goal of K-means clustering is to

1. learn the means of the each cluster and
2. assign every point in the data set to one of the clusters.

Question 1 Write the cost function to minimize for the K-means algorithm, in terms of the $\mathbf{x}_n$, r_{nk}, $\boldsymbol{\mu}_k$ and the distance function.
Solution: Find r_{nk} and $\boldsymbol{\mu}_k$ which minimize

$$J(r_{nk}, \boldsymbol{\mu}_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2, \tag{1}$$

■

K-means is an iterative algorithm. This means that at iteration i, the algorithm uses information from the previous iteration $i-1$ to minimize a cost function J.

Assume that at iteration i, you have an estimate for $\boldsymbol{\mu}_k$: $\boldsymbol{\mu}_k^{(i-1)}$.

Question 2 What are the optimal values for $r_{nk}^{(i)}$ according to the current estimate $\boldsymbol{\mu}_k$?

Solution:

$$r_{nk}^{(i)} = \begin{cases} 1 & \text{if } k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j^{(i-1)}||_2^2 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

■

Question 3 How can you optimize $\boldsymbol{\mu}_k^{(i)}$ based on the new estimates for $r_{nk}^{(i)}$?
Solution: Derive and set to 0:

$$\begin{aligned} &\left. \frac{\partial J(r_{nk}^{(i)}, \boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k} \right|_{\boldsymbol{\mu}_k^{(i)}} = 0 \\ \implies &2 \sum_{n=1}^{N} r_{nk}^{(i)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(i)}) = 0 \\ &\boldsymbol{\mu}_k^{(i)} = \frac{\sum_n r_{nk}^{(i)} \mathbf{x}_n}{\sum_n r_{nk}^{(i)}} \end{aligned} \tag{3}$$

■

Question 4 Assume that you repeat the two steps above until the cost function J stops decreasing. What convergence guarantees do you have with such an algorithm ?

Solution: Convergence guarantees to a local minima because the two steps decrease the objective function. Not convergence guarantees to converge to a global minima because the means are initialized randomly.

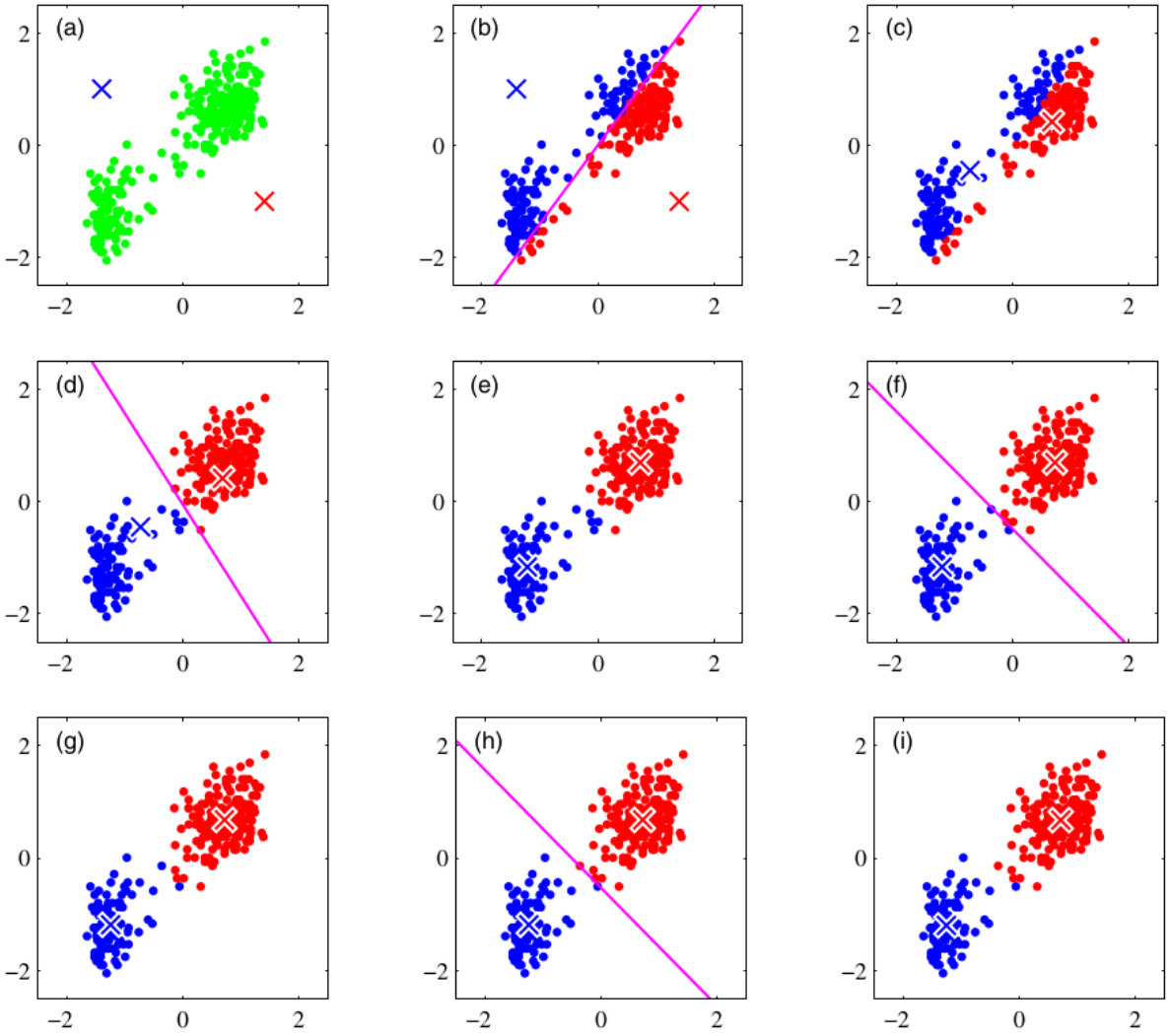

Figure 1: Example of running K-means on a 2-dimensional dataset. Bishop 2006 Figure 9.1

■

Applying K-means to a dataset is a way to find underlying clusters in the data. Clusters are structures in data that can be useful for inference tasks e.g. classification, if the clusters turn out to correspond to certain classes of interest. K-means, although very powerful because simple to implement, has a number of limitations.

1. It assigns every point in the dataset to one and only one cluster, this kind of hard assignment might not be adequate to points lying midway to two (or more) cluster centers.
2. The cluster mean is not robust to outliers
3. Euclidean distance is not appropriate for categorical variables.
4. The shape of the clusters are limited.

An underlying structure in data is also called a latent structure, and it is often better to use a framework that is more general than hard cluster assignment to represent and infer the structure. We will see that a probabilistic formulation leads to more general solution for latent structure inference.

1.2 Probabilistic K-means

Let us denote $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ the set of observed training data (see previous section).

The probabilistic interpretation of K-means arises by defining the clusters in terms of distributions rather than simply by there means. We will aim at maximizing the likelihood of the dataset $\mathbf{X}$ wrt to a mixture of Gaussian model:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right], \tag{4}$$

Note:

1. It is possible to reach infinite likelihood if a mean = a datapoint
2. Identifiability problem: $K!$ combinations of parameters lead to the same likelihood.

Question 1. What condition on $\boldsymbol{\mu}_k, \pi_k$ and $\boldsymbol{\Sigma}_k$ must be satisfied at a maximum likelihood ?

Solution: Derivatives wrt $\boldsymbol{\mu}_k$ must be 0:

$$\begin{aligned}
&\left. \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_k} \right|_{\boldsymbol{\mu}_k^\star} = 0 \\
\implies 0 &= \sum_{n=1}^{N} \frac{\frac{\partial \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \text{ where we derived ln wrt } \boldsymbol{\mu}_k
\end{aligned}$$

Next, we derive the numerator (using equation (86)) from the Matrix Cookbook)

$$\begin{aligned}
&\frac{\partial \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} = \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k), \\
&\text{Denoting } \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \in \mathbb{R}, \text{ we get:} \\
&0 = \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)
\end{aligned} \tag{5}$$

Finally :

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

Derivatives wrt $\boldsymbol{\Sigma}_k$ must be 0:[1]

$$\begin{aligned}
&\left. \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k} \right|_{\boldsymbol{\Sigma}_k^\star} = 0 \\
\boldsymbol{\Sigma}_k^\star &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned} \tag{6}$$

The mixture parameter $\boldsymbol{\pi}_k$ has the additional constraint that $\sum_{k=1}^{K} \pi_k = 1$, thus we introduce a Lagrangian multiplier and maximize wrt π_k and λ the quantity l:

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

This gives

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \tag{7}$$

Multiplying by π_k summing over k and using the constraint :

$$\lambda = -N, \pi_k = \frac{N_k}{N},$$

[1] Details for the derivation at `https://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall07/reading/gauss.pdf`

where $N_k = \sum_n \gamma(z_{nk})$ and $N = \sum_k N_k$. ■

Question 2. Using the equations derived above, find an iterative algorithm to maximize the log-likehood of data wrt to the parameters of a Gaussian Mixture model.
Solution: The EM for Gaussian Mixtures p438 Bishop 2006:

1. Initialize the parameters
2. Evaluate the responsibilities $\gamma(z_{nk})$
3. Re-estimate the parameters using the current responsibilities
4. Check convergence

■

Question 3. (Relation with K-means)
What theoretical assumption on the mixture model do we have to make for the algorithm we derived to reduce to K-means ?
Solution: We assume $\mathbf{\Sigma}_k = \epsilon \mathbf{I}$, this leads to

$$\gamma(z_{nk}) = \frac{\pi_k \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_k||^2/2\epsilon)}{\sum_j \pi_j \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_j||^2/2\epsilon)}. \tag{8}$$

If we denote $k^\star$ the cluster mean that is closer to point n, then for $k \neq k^\star, \gamma(z_{nk}) \to 0$ and $\gamma(z_{nk^\star}) \to 1$ when $\epsilon \to 0$, in turn leading to a hard assignment to cluster $k^\star$ for point n. ■

1.3 A graph for Gaussian mixture models

In a Gaussian mixture model, we assume that input data $\mathbf{x} \in \mathbb{R}^d$ are emitted by a Gaussian distribution, chosen at random among $K \in \mathbb{N}_*$ possible distributions. The parameters of the distributions can be collected in sets $\boldsymbol{\mu}$, $\mathbf{\Sigma}$, $\boldsymbol{\pi}$ where, e.g. $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ with appropriate dimensions.

The choice of the emission distribution can be modeled with a categorical unobserved (i.e. latent) variable $\mathbf{z} \in \{0, 1\}^K$.

Question 1. Draw a directed graphical model representing a Gaussian mixture model.

Solution:

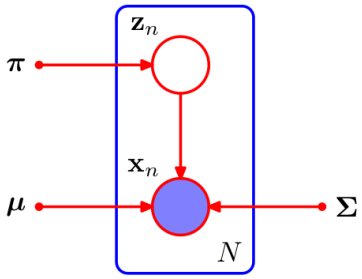

Figure 2: Directed graph representation of a Gaussian mixture model (Bishop Fig 9.6).

■ Consider the directed graph for a Gaussian mixture model shown in Figure 2. Suppose that you get N data points $\mathbf{X} = \{\mathbf{x}_n\}$ and you want to infer the corresponding values for the latent variables $\mathbf{Z} = \{\mathbf{z}_n\}$.

Question 2. By making use of the d-separation criterion discussed in Section 8.2 (Bishop, 2006), write the factorization for the joint distribution $p(\mathbf{x}_n, \mathbf{z}_n)$ (omitting the parameters for simplicity).
Solution: $p(\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)$ ■

Question 3. Show that the posterior distribution of the latent variables $p(\mathbf{Z}|\mathbf{X})$ factorizes as a product of posterior distributions for the different data points.

Solution: The directed graph indicates that there is only a link from $\mathbf{z}$ to $\mathbf{x}$. The samples are i.i.d., thus we have that $p(\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{z}_n)$ and $p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n)$.

Using Bayes' Rule:

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X}) &= \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})} \\ &= \frac{\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n) \prod_{n=1}^{N} p(\mathbf{z}_n)}{\prod_{n=1}^{N} p(\mathbf{x}_n)} \\ &= \prod_{n=1}^{N} \frac{p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{x}_n)} \\ &= \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{x}_n) \end{aligned} \tag{9}$$

■

1.4 Expectation-Maximization (EM)

(Bishop Ex. 9.4) Suppose we wish to use the EM algorithm to maximize the posterior distribution over parameters $p(\theta|\mathbf{X})$ for a model containing latent variables $\mathbf{Z}$, where $\mathbf{X}$ is the observed data set. Show that 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

$$Q'(\theta, \theta^{old}) = Q(\theta, \theta^{old}) + \ln p(\theta) \tag{10}$$

where $Q(\theta, \theta^{old})$ is defined in (11).

Recall, read the details in section 9.3 (Bishop, 2006): In the classical formulation of EM, we aim at finding a maximum likelihood solution for the parameter θ of a statistical model, i.e. we aim at maximizing $\ln p(\mathbf{X}|\theta)$ wrt θ.

For this we use an iterative approach (1) first we evaluate the posterior of the latent variable wrt some estimate of the parameters θ^{old}, (2) we maximize, wrt θ, a function

$$Q(\theta, \theta^{old}) = E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\theta)\right]. \tag{11}$$

Solution: We write

$$\begin{aligned} \ln p(\theta|\mathbf{X}) &\propto \ln p(\mathbf{X}|\theta)p(\theta) \\ &= \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)p(\theta)\right] \end{aligned} \tag{12}$$

Then we follow the same development as the classical EM, but replace the term for the "complete data log-likelihood" $p(\mathbf{X}, \mathbf{Z}|\theta)$ with $p(\mathbf{X}, \mathbf{Z}|\theta)p(\theta)$ in the expectation.

$$\begin{aligned} Q'(\theta, \theta^{old}) &= E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\theta)p(\theta)\right] \\ &= E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\theta) + \ln p(\theta)\right] \\ &= E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\theta)\right] + E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\theta)\right] \\ &= E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\theta)\right] + \ln p(\theta) E_{\mathbf{Z}|\mathbf{X},\theta^{old}} \left[1\right] \\ &= Q(\theta, \theta^{old}) + \ln p(\theta) \end{aligned} \tag{13}$$

■