CS598: Graphical Models, Fall 2016

Lecture 11 : Discrete Latent Variable Models

Lecturer: Sanmi Koyejo *Scribe: Raj Kataria, Amit Das, Sep. 27, 2016*

1 Mixture Models

Mixture models are used for representing data when they appear in clusters. Here, each cluster represents a *mixture* or a *component*. There is one model per component (or cluster) which we will refer to as a *component model*. The collection of all component models leads to the formation of richer model called the *mixture model*.

1.1 Representation

We assume that a mixture model can be represented as a directed graphical model as shown in Figure 1.

Figure 1: Directed Graphical Model representing a Mixture Model

Here, X is the observed random variable representing the data and Z is the hidden (or latent) random variable representing the component. Together, they represent a generative model where the observed data point X is generated from the component (cluster index) Z. In general, X and Z can be continuous or discrete. However, in this lecture, since we are dealing with *discrete* latent variable models, we will assume Z is discrete. The assumption that Z is hidden is also valid since just by observing at X, it is hard to infer which component (or cluster) model generated X.

Assume $\mathbf{X} \in \mathbb{R}^d$ is an observed d-dimensional random vector and $Z \in \{1, 2, \cdots, K\}$ where K represents the number of components (or clusters). Therefore, the mixture model can be mathematically represented as,

$$p(\mathbf{X}) = \sum_{Z=1}^{K} p(\mathbf{X}, Z) = \sum_{Z=1}^{K} p(Z)p(\mathbf{X}|Z), \tag{1}$$

where,

$$\sum_{Z=1}^{K} p(Z) = 1. \tag{2}$$

The equation in (1) is the mixture model equation. It is simply alluding to the fact that we do not know for sure which individual component out of K components generated the data point $\mathbf{X}$ (since Z is hidden). Then the data point can be assumed to be generated by a rich mixture model which is a collection of individual component models $p(\mathbf{X}|Z)$ where each component model contributes to the generation of data point. The contribution weights from the individual component models are determined by $p(Z)$. Thus, if a particular component has a higher weight than all the other $K-1$ components, then it is more likely that the data point was generated by that component.

1.2 Inference and Learning

Two interesting problems associated with mixture models are the following:

- **Inference:** Given $\mathbf{X}$, what is $p(Z|\mathbf{X})$? The term $p(Z|\mathbf{X})$ is the aposteriori probability that the observed data point $\mathbf{X}$ was generated by component Z. This is simply,

$$p(Z|\mathbf{X}) = \frac{p(\mathbf{X}|Z)p(Z)}{p(\mathbf{X})}, \tag{3}$$

  where $p(\mathbf{X}|Z), p(Z), p(\mathbf{X})$ are given in (1) and (2).

- **Learning:** The equation in (3) cannot be solved unless we know (1) and (2). Thus, the learning problem involves finding $p(Z)$ and $p(\mathbf{X}|Z)$. A few approaches for addressing this problem are the following:

  - Theory and statistics.
  - Learning model parameters assuming that the model is part of a family of distributions.
  - Learning without making any distribution assumptions. This falls into the category of learning non-parametric distributions.

  In this lecture, we are interested in the second approach, i.e., learning the model parameters of a family of distributions.

1.3 Learning Mixture Models Using Parametric Distributions

In this section, we formulate the general framework of the learning problem involving parametric distributions.

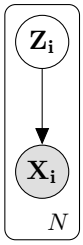

Figure 2: Directed Graphical Model representing a Mixture Model for N i.i.d. data points $\mathbf{X}_i$, where $i = 1, \cdots, N$. The corresponding hidden variables are given by Z_i. The independence of N observations is indicated by the plate.

Instead of representing a data point as X as we have been doing so far, we'll represent the i^{th} data point as $\mathbf{X}_i$. Thus, for N i.i.d. observed data points $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N$, we have their corresponding latent variables $Z_1, Z_2, \cdots, Z_N$. The directed graphical model for this case is shown in Figure 2

The joint denisty of the observed and latent variables is given by,

$$p(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N, Z_1, Z_2, \cdots, Z_N) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{N} p(\mathbf{X}_i, Z_i). \tag{4}$$

The mixture model equation for (4) is obtained by marginalizing (4) over $Z_1, Z_2, \cdots, Z_N$. Thus,

$$\begin{aligned} p(\mathbf{X}_1, \cdots, \mathbf{X}_N) &= \sum_{Z_1,\ldots,Z_N} p(\mathbf{X}_1, \cdots, \mathbf{X}_N, Z_1, \cdots, Z_N) \\ &\overset{(4)}{=} \sum_{Z_1,\ldots,Z_N} \prod_{i=1}^{N} p(\mathbf{X}_i, Z_i) \\ &= \prod_{i=1}^{N} \sum_{Z_i} p(\mathbf{X}_i, Z_i) \\ &\overset{(1)}{=} \prod_{i=1}^{N} \sum_{Z_i} p(Z_i) p(\mathbf{X}_i | Z_i) \end{aligned} \tag{5}$$

Since we are dealing with parametric distributions, we will parameterize $p(\mathbf{X}_1, \cdots, \mathbf{X}_N)$ in (5) with a set of parameters $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K\}$, where each $\boldsymbol{\theta}_k$ represents a smaller set of parameters corresponding to component k where $k \in \{1, \cdots, K\}$. Moreover, we define $p(Z_i = k) \overset{\Delta}{=} \pi_k$ as the prior for component k. Thus, $\pi = \{\pi_1, \cdots, \pi_K\}$ represents the prior distribution over the K components. As a result, the entire parameter set consists of $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}$, where $k = 1, \cdots, K$.

Thus, incorporating the model parameters $\boldsymbol{\theta}$ in (5), (5) can be written as,

$$\begin{aligned} p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta}) &= \prod_{i=1}^{N} \sum_{Z_i} p(Z_i = k) p(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k) \\ &= \prod_{i=1}^{N} \sum_{Z_i} \pi_k p(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k) \end{aligned} \tag{6}$$

For the special case of $i = 1$, (6) becomes (1) when parameterized by $\boldsymbol{\theta}$.

The objective of the learning problem is to estimate the parameters $\boldsymbol{\theta}$ such that the likelihood of the mixture model in (6) is maximized. Thus,

$$\begin{aligned} \boldsymbol{\theta}^\star &= \underset{\boldsymbol{\theta}}{\text{argmax}} \; p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\text{argmax}} \; \log p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta}) \end{aligned} \tag{7}$$

1.4 Gaussian Mixture Model

Consider the component model $p(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k)$ in (6). For the special case that each component in $p(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k)$ can be modeled by a Gaussian distribution, the mixture model in (6) becomes a Gaussian Mixture Model (GMM). Thus, for a given component $Z_i = k$, the component model can be represented as,

$$p(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k) = \mathcal{N}(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)), \tag{8}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean and $\boldsymbol{\Sigma}_k$ is the $d \times d$ covariance matrix parameterizing the Gaussian component model in (8). Note that, in the right hand side of (8), the parameter set is represented as $\boldsymbol{\theta}_k$ instead of $\boldsymbol{\theta}$. This is because for a given component $Z_i = k$, the component model is entirely determined by the parameters in $\boldsymbol{\theta}_k$. The parameters of the other components $\boldsymbol{\theta}_j$, $j \neq k$, do not determine the value of the probability in (8). However, there is nothing wrong in using the notation $\mathcal{N}(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta})$ instead of $\mathcal{N}(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k)$. The former notation may be used in some textbooks. When the former notation is used, it must be implicitly understood that only $\boldsymbol{\theta}_k \subseteq \boldsymbol{\theta}$ is used to evaluate (8). Other parameters in $\boldsymbol{\theta}$ are not used to evaluate (8).

The other term $p(Z_i = k) = \pi_k$ in (6) is the prior probability of component k. Thus, the GMM for the data point $\mathbf{X}_i$, can be written as,

$$\begin{aligned} p(\mathbf{X}_i; \boldsymbol{\theta}) &= \sum_{Z_i} \pi_k \mathcal{N}(\mathbf{X}_i | Z_i = k; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k . \boldsymbol{\Sigma}_k)) \\ &= \sum_{Z_i} \pi_k \mathcal{N}(\mathbf{X}_i; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k . \boldsymbol{\Sigma}_k)) \end{aligned} \tag{9}$$

where,

$$\mathcal{N}(\mathbf{X}_i; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k.\boldsymbol{\Sigma}_k)) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right). \tag{10}$$

The learning problem in GMM is to determine the parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, where $k = 1, \cdots, K$. This can be obtained by substituting the individual Gaussian component model from (9) in (6) and the resulting equation in (7). Thus, the GMM learning problem becomes,

$$\begin{aligned} \boldsymbol{\theta}^\star &= \underset{\boldsymbol{\theta}}{\text{argmax}} \ \log p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{X}_i; \boldsymbol{\theta}_k)\right) \end{aligned} \tag{11}$$

The total number parameters to be estimated in (11) is $K + Kd + K\frac{d(d+1)}{2}$. Due to the symmetric nature of covariance matrices, the maximum number of parameters per covariance matrix is $\frac{d(d+1)}{2}$. Finding the optimal $\boldsymbol{\theta}^\star$ in (11) is a d-dimensional non-concave optimization problem. It is hard to find a globally optimum solution $\boldsymbol{\theta}^\star$ directly. In the next class, we will see how a sub-optimal solution can be obtained using the Expectation-Maximization (EM) algorithm Dempster et al. (1977). It is sub-optimal because it is guaranteed to find a locally optimal value of $\boldsymbol{\theta}$ but not guaranteed to find the global optimum $\boldsymbol{\theta}^\star$.

Here, we'll introduce the term τ_k^i which will be used during EM. For an observed data point $\mathbf{X}_i$, τ_k^i is the aposteriori probability that the data point $\mathbf{X}_i$ was generated by component k. Thus,

$$\begin{aligned} \tau_k^i &\triangleq p(Z_i = k|\mathbf{X}_i; \boldsymbol{\theta}) \\ &= \frac{p(\mathbf{X}_i|Z_i = k; \boldsymbol{\theta}_k)p(Z_i = k)}{p(\mathbf{X}_i; \boldsymbol{\theta})} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{X}_i; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k.\boldsymbol{\Sigma}_k))}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{X}_i; \boldsymbol{\theta}_j = (\boldsymbol{\mu}_j.\boldsymbol{\Sigma}_j))} \end{aligned} \tag{12}$$

1.5 Determining the number of components (K) in a Mixture Model

There are a number of ways to determine the number of components K.

- With increasing values of K, the log likelihood score, $\log p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta})$ in (7) increases. However, after a certain value of K, the increase in score is marginal. Thus, if the *increase* in score becomes less than some heuristically determined threshold, then we stop increasing K.
- The model complexity is a function of K and can be determined by AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). Both AIC and BIC increase when K

increases. Thus, by using a modified cost function where the original cost function is penalized by AIC or BIC, the model complexity can be controlled. This is given as,

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \Big\{ \; - \log \, p(\mathbf{X}_1, \cdots, \mathbf{X}_N; \boldsymbol{\theta}) + C(\boldsymbol{\theta}) \Big\} \tag{13}$$

- Another way to control K is by using the out-of-sample fit method. Here, for each $\boldsymbol{\theta}_k$, we evaluate the likelihood of data points from a cross-validation set that is not present during training. If the likelihood of the cross-validation set increases more than a heuristically set threshold, we increase K. Else, we do not increase K any further.

Bibliography

Dempster, A. P., Laird, N. M. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat Society B.* **39** 1–38.