SML Recitation Notes Week 5: Bayesian Nonparametrics

Min Xu

February 22, 2011

Frequentist Gaussian Mixture Model:

1. Let K be number of mixture components (clusters), let N be number of samples
2. Let $\mu_1, ...\mu_K \in \mathbb{R}$ be **mixture centers**, let σ be a fixed **mixture variance**
3. Let $0 < p_1, ...p_K \leq 1$ be **mixture weights**, let $p_1 + ... + p_K = 1$
4. Let one-dimensional data $X_1, \ldots, X_N \sim \sum_{k=1}^{K} p_k N(\mu_k, \sigma^2)$. More specifically, for each sample $i = 1, ..., N$:
   Randomly generate **mixture indicator** Z_i by letting $Z_i = k$ with probability p_k (we will denote this $Multi(p_1, \ldots, p_K)$)
   Draw $X_i \sim N(\mu_{Z_i}, \sigma^2)$

The parameters of the Gaussian Mixture Model are K the number of clusters, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ the mixture centers, σ the mixture variance, $\mathbf{p} = \{p_1, \ldots, p_K\}$ the mixture weights.

To summarize our model:

$$Z_i \sim Multi(p_1, \ldots, p_K)$$
$$X_i | Z_i \sim N(\mu_{Z_i}, \sigma^2)$$

Suppose we have data $X_1, ...X_N$, we can fit the Gaussian Mixture Model to the data by inferring all the parameters through maximum likelihood estimation (or some other measurement of goodness of fit). Note that $Z_1, ...Z_N$ are not parameters; they are latent variables: there are different sets of Z_i's for different collection of samples. The latent variables are artifacts of our model construction and also used to interpret the model and more easily perform inference.

1 Bayesian Finite-Mixture Model

In Bayesian statistics, we treat the parameters as latent random variables. Consequently, we do longer care about maximum likelihood estimates but rather, we care about the **posterior distribution of the parameters** given the data. We need a prior distribution on the parameters however to compute the posterior.

In hand-wavy mathematical notation (unrelated to our GMM notations), if data X is generated from $p(X|\theta)$ with parameter θ, then the posterior

$$\underbrace{p(\theta|X)}_{\text{posterior}} \propto \underbrace{p(X|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

Remark 1. The central object of interest in Bayesian analysis is the **generative model**. The generative model ultimately just specifies the distribution from which the data is generated. However, it also encodes underlying meaning behind the way the data is generated through latent variables and fixed (hyper)parameters. The prior is just a part of the generative model.

In Bayesian generative model, there is no such thing as "non-random unknown parameters". All parameters are either pre-set to some value a priori or they are latent variables.

Given the data, Bayesian data analysis would perform **inference** to find posterior distribution of the latent variables conditioned on the data. We would then get the desired information about the data from these posterior distributions.

And so for Bayesian finite-mixture modeling, we define the following priors (and hence a generative model)

1. For each $k = 1, ...K$, let $\mu_k \sim N(0, A)$ where A is a hyperparameter (usually set to something large)
2. Let the collection $(p_1, ..., p_K) \sim$ Dirichlet$(\alpha_1, ..., \alpha_K)$ where $\alpha_1, ..., \alpha_K$ are hyperparameters
3. For simplicity, we will say σ is a constant, say $\sigma = 1$
4. For finite mixture model, we will say that K is also a constant

To summarize our generative model: We have fixed parameters $A, \sigma, \alpha_1, \ldots, \alpha_K$

$$\begin{aligned} \mu_k &\sim N(0, A) \\ (p_1, \ldots, p_K) &\sim Dir(\alpha_1, \ldots, \alpha_K) \\ Z_i|\mathbf{p} &\sim Multi(p_1, \ldots, p_K) \\ X_i|Z_i, \boldsymbol{\mu}, \sigma &\sim N(\mu_{Z_i}, \sigma^2) \end{aligned}$$

We can represent this via the **Plate Notation**:

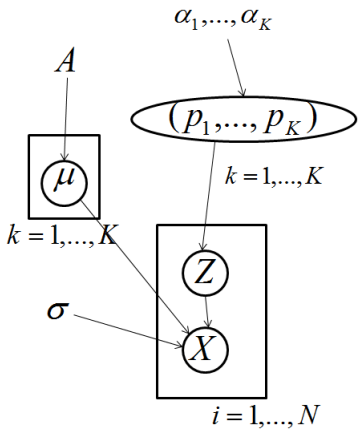

Figure 1: Plate notation for Finite Mixture Model

Here is how we interpret the plate notation:

- Circled symbols are random variables, uncircled symbols are pre-set parameters/hyperparameters

- Plates indicate that there are many instances of the variables inside
- The arrows indicate "dependency"; distribution of a random variable is completely specified by only the symbols that point to it

It is important to note that we did not choose the distribution of the priors $(p_1, \ldots, p_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$ and $\mu_k \sim N(0, A)$ arbitrarily. Dirichlet and Gaussian are conjugate priors to Multinomial and fixed-variance Gaussian distributions respectively.

2 Dirichlet-Process Mixture Model

Now, we would like to have arbitrary number of clusters.

There are several equivalent views of Dirichlet-Process Mixture, that is, there are several seemingly different generative processes that all define the same distributions on the data. These seemingly different generative processes are based on equivalent definitions of a Dirichlet Process.

2.1 View 1: Direct Dirichlet Process

Definition 1. A Dirichlet Process is a distribution over all infinite discrete distributions with the following property. Let $A_1, ...A_n$ be a partition of $\mathbb{R}$, let $G \sim DP(\alpha, G_0)$, then $(G(A_1), \ldots, G(A_n)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_n))$

Here is the generative model:
We have fixed parameters α, σ, A

$$\begin{aligned} G &\sim DP(\alpha, N(0, A)) \\ (\mu_i', Z_i)|G &\sim G \text{ for } i = 1, \ldots, N \\ X_i|\mu_i, Z_i &\sim N(\mu_i', \sigma^2) \text{ for } i = 1, \ldots, N \end{aligned}$$

Several things have changed from the finite mixture model:

- Before, we had K distinct μ_k's, one for each mixture whereas now, we have a μ_i' for every sample. Many of the μ_i''s will repeat and from the reptitions, we implicitly get the mixture indicators Z_i's; that is, $Z_i = Z_j$ iff $\mu_i' = \mu_j'$
- G is a infinite discrete distribution. **Infinite** because a draw from G can take on possibly infinite number of values (these are the infinite number of potential cluster centers); **discrete** because two different draws from G can take on the same value. G is really the infinite version of the $\{(\mu_k, p_k)\}_{k=1,\ldots K}$ parameters from the finite mixture model.
- Here, G_0, the base distribution for Dirichlet Process, is just $N(0, A)$

Remark 2. Often, people will write G as $G(x) = \sum_{k=1}^{\infty} p_k \delta_{\mu_k}(x)$. This is the probabililty mass function of G, that is, $G(x)$ is the probability that a draw from G is equal to x.

This generative model is hard to interpret because we don't know what G really looks like.

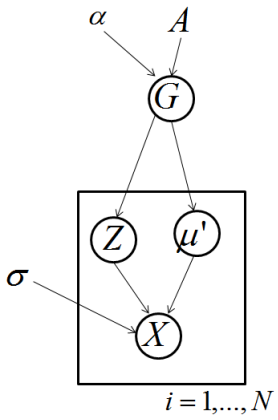

Figure 2: Plate notation for Dirichlet Process Mixture view 1

2.2 View 2: Stick Breaking Prior

Definition 2. We can draw an infinite discrete distribution G from $DP(\alpha, G_0)$ as such:

1. Draw $\mu_1, \mu_2, \ldots$ independently from G_0
2. Draw $V_1, V_2, ... \sim Beta(1, \alpha)$
3. Let $p_1 = V_1$, $p_2 = V_2(1 - V_1)$, $p_3 = V_3(1 - V_2)(1 - V_1)$, etc.
4. Let G be a discrete distribution that puts mass p_j at μ_j

We can now interpret $\{\mu_k\}_{k=1,\ldots,\infty}$ as the mixture centers and $\{p_k\}_{k=1,\ldots,\infty}$ as the mixture weights for the infinite number of clusters.

We can now define a new generative model: we have fixed parameters α, σ, A

$$\begin{aligned} \mu_k &\sim N(0, A) \text{ for } k = 1, \ldots, \infty \\ \{p_k\}_{k=1,\ldots,\infty} &\sim \text{ Stick-Breaking}(\alpha) \\ Z_i|\boldsymbol{p} &\sim \text{ Infinite-Multi}(\{p_k\}_{k=1,\ldots,\infty}) \\ X_i|Z_i, \boldsymbol{\mu}, \sigma &\sim N(\mu_{Z_i}, \sigma^2) \end{aligned}$$

To be more precise, for all i, we set $Z_i = k$ with probability p_k where k can range from 1 to infinity.

What we have really done here is to explicitly describe the Infinite Discrete Distribution G in term of $\{\mu_k\}_{k=1,\ldots,\infty}$ and $\{p_k\}_{k=1,\ldots,\infty}$. The previous $\mu_i' = \mu_{Z_i}$ here.

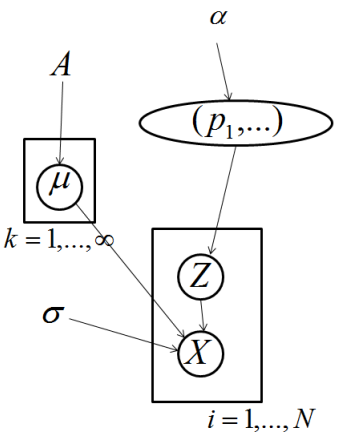

Figure 3: Plate notation for Dirichlet Process Mixture view 2

2.3 View 3: Chinese Restaurant Process

From View 1, we know we generated $\mu'_1, \dots \mu'_N$ by first drawing G from a DP and then drawing from G. We can also marginalize out G and describe just the distribution of the μ'_i's with the Chinese Restaurant Process.

Definition 3. Let G_0, α be parameters to a Dirichlet Process. Let $G \sim DP(\alpha, G_0)$ and let $\mu'_1, \dots, \mu'_N \sim G$. Then the marginal distributions of $(\mu'_1, \dots, \mu'_N) \sim CRP(\alpha, G_0)$ is the Chinese Restaurant Process:

1. Draw $\mu'_1 \sim G_0$
2. Draw $\mu'_2 = \begin{cases} \mu'_1 & \text{w.p.} \frac{1}{2+\alpha-1} \\ \sim G_0 & \text{w.p.} \frac{\alpha}{2+\alpha-1} \end{cases}$
3. ... Draw $\mu'_n = \begin{cases} \mu'_j & \text{w.p.} \frac{1}{n+\alpha-1} \text{ for } j = 1, \dots, n-1 \\ \sim G_0 & \text{w.p.} \frac{\alpha}{n+\alpha-1} \end{cases}$

The intuitive description:

- Every sample μ'_i is a new customer
- We say two customers μ'_i and μ'_j sit at same table if $\mu'_i = \mu'_j$
- The customers come in one by one. A new customer can either randomly sit by an old customer or demand that the restaurant brings out a new table.
- After all N customers come in, the number of tables are the number of *unique* values among $\{\mu'_i\}_{i=1,\dots,N}$

Clusters formed by the CRP has the "rich gets richer" phenomenon. Notice that we also implicitly draw the cluster indicator Z_i's from CRP.

With the CRP, we can define our third generative model:

$$(\mu'_i, Z_i) \sim CRP(\alpha, G_0)$$
$$X_i | Z_i, \mu'_i, \sigma \sim N(\mu'_i, \sigma^2)$$

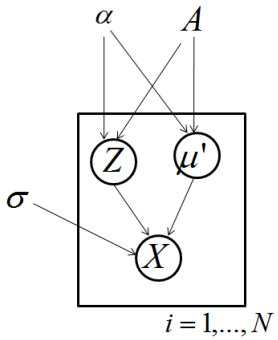

Figure 4: Plate notation for Dirichlet Process Mixture view 3