Bayesian Nonparametrics: Bayesian treatments of nonparametric problems

Traditional Bayes approach:

1. Posit a model with parameters $\theta$, data $X$, prior $p(\theta)$, and likelihood $p(X \mid \theta)$
2. Compute the posterior $p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)}$ using Bayes' rule, where $p(X) = \int p(X \mid \theta)p(\theta) d\theta$.
3. Examine results, and repeat if necessary.

Aside: The maximum a posteriori approach is not as informative as the Bayes approach. It does not give the entire posterior distribution, only the mode.

Aside: There is a Bayes interpretation of the LASSO solution as the posterior with the prior on $\beta$ as the Laplacian distribution. However, as the Laplacian distribution assigns zero mass to 0, so does the posterior, which is inconsistent with the LASSO.

Clustering: Suppose we have some data like the figure below, and we would like to cluster it. We can fit a Gaussian mixture model, with priors on the parameters for each cluster. More specifically, if there are $K$ clusters, the priors are:

- the mixture probabilities $\theta \sim \text{Dirichlet}(\alpha)$, a distribution on the $K$-dimensional simplex
• the cluster means $\mu_k \sim N(0, \sigma_0^2 I)$

• the cluster covariance matrices $\Sigma_k \sim$ inverse Wishart

To sample from the model, we first choose the cluster: $Z_i \sim \text{categorical}(\theta)$ and then let $X_i \sim N(\mu_{Z_i}, \Sigma_{Z_i})$. We may represent this graphically as:

![Diagram of Gaussian mixture model]

**Figure 2: Source: Wikipedia article on mixture models**

Remark: We may not know *a priori* how many clusters there are. We can put a prior on $K$, but this may be computationally difficult. We can also use AIC/BIC or some other model selection criterion. Lastly, instead of using a Dirichlet prior on $\theta$, we use a Dirichlet process as a ‘prior’ for $\theta$ and $K$.

Aside: In one dimension, instead of fitting a Gaussian mixture model, we can
also do kernel density estimation. That is, we plot a histogram of the data, put Gaussian kernels on each 'bump', and add up the kernels to estimate the density.

**Dirichlet distribution**

- parameter $\alpha$, a $K$-dimensional vector
  \[
  \text{Dir}(\theta, \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} = \exp(\sum_i (\alpha_i - 1)\log(\theta_i) - \sum_i \log(\Gamma(\alpha_i)) + \log(\sum_i \alpha_i))
  \]

  where $\theta$ is in the $K$-dimensional simplex

- If $\tau_i \sim \text{Gamma}(\alpha_i, 1)$ and $\theta_i = \frac{\tau_i}{\sum_i \tau_i}$, then $\theta \sim \text{Dir}(\alpha)$. Also, $\sum_i \tau_i \sim \text{Gamma}(\sum_i \alpha_i, 1)$ and $\theta$ is independent of $\sum_i \tau_i$.

- $(\theta_1, 1 - \theta_1) \sim \text{Dir}(\alpha_1, \sum_i \alpha_i - \alpha_i) = \text{Beta}(\alpha_i, \sum_i \alpha_i - \alpha_1)$

- $E(\theta) = \frac{\alpha}{\sum_i \alpha_i}$

- As $\sum_i \alpha_i \to \infty$, $\theta \xrightarrow{P} E(\theta)$, which is essentially a delta function. As $\sum_i \alpha_i \to 0$, $\theta$ gets sparser and $P(0, \ldots, 0, 1, 0, \ldots, 0) \approx \frac{\alpha_i}{\sum_i \alpha_i}$. 

3
List of topics

- Dirichlet Processes
- MCMC for Dirichlet Processes
- Hierarchical Dirichlet Processes
- Infinite Hidden Markov Models
- Gaussian Processes
- Beta processes

One example: Chinese Restaurant Process

If we have a (Chinese) restaurant which has infinite number of tables and infinite number of seats at each table. Assume there is cluster effect, i.e. people tend to sit together. The probability of an incoming customer choosing a new table is proportional to \( \alpha \) and the probability of sitting at an existing table is proportional to the number of people at that table. We have a flow of customers, \( X_1, X_2, \ldots \) and the imagine the tables as the clusters, the number of clusters and the number of people at each cluster are called Chinese Restaurant Process. It is a characterization of Dirichlet Process.

We now discuss some properties of CPR (Chinese Restaurant Process):

- Exchangability: The partition of the customers at each table does not depend on the order the customers come in.
- Limiting behavior: as \( j \) is big, expected number of tables is of order \( \alpha \log j \).

Proof:

\[
\mathbb{E}[\# tables] = \mathbb{E}\left[ \sum_{j} I_j \right], \quad \text{where } I_j \text{ denotes person } j \text{ sitting at a new table}
\]

\[
= \sum_{j=0}^{J-1} \frac{\alpha}{\alpha + j}
\]

\[
\approx \alpha \log J
\]

Now we have a new interpretation of the mixture model: instead of thinking of the cluster centers \( Z_j \) as draws from a parameter \( \theta \) that comes from a Dirichlet distribution as in Figure , we can think of the points as the customers that according to the rule stated above forming the clusters, and each cluster is specified by a set of parameters \( \phi_i \) iid from some prior distribution as in Figure.
Mixture model from Dirichlet process perspective