Bayesian Nonparametrics Notes

Xi Tan (xtan3.1415926@gmail.com)

March 23, 2018

Contents

1 Introduction

This note is based on Peter Orbanz's BNP notes:

`http://stat.columbia.edu/~porbanz/npb-tutorial.html`

2 Notation

Bold upper case letters represent matrices, e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\Theta}$. Bold lower case letters represent vector-valued random variables and their realizations (we do not distinguish between the two), e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}$. Curly upper case letters represent spaces (i.e., possible values) of random variables, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \Theta$.

3 Terminology

3.1 Parametric and nonparametric models

In a set of probability spaces $\{(\mathcal{Y}, \mathcal{F}, \mathcal{P}_\Theta)\}$, a *statistical model* $\mathcal{M}$ on a sample space $\mathcal{Y}$ is a set of probability measures $\mathcal{P}_\Theta$ on $\mathcal{Y}$. If we write $PM(\mathcal{Y})$ for the space of all probability measure on $\mathcal{Y}$, a model is a subset $\mathcal{M} \subset PM(\mathcal{Y})$. Every element of $\mathcal{M}$ has a one-to-one mapping (hence the model is *identifiable*) with its parameter $\boldsymbol{\theta}$ with values in a parameter space Θ, that is,

$$\mathcal{M}(\mathbf{y}) = \{P_{\boldsymbol{\theta}}(\mathbf{y}) | \boldsymbol{\theta} \in \Theta\}, \quad \mathbf{y} \in \mathcal{Y}. \tag{1}$$

We call a model *parametric* if Θ has finite dimension, and *nonparametric* if Θ has infinite dimension.

To formulate statistical problems, we assume that n observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ with values in $\mathcal{Y}$ are observed, which are drawn i.i.d. from a measure $P_{\boldsymbol{\theta}}$ in the model, i.e.,

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \sim_{iid} P_{\boldsymbol{\theta}} \qquad \text{for some} \quad \boldsymbol{\theta} \in \Theta \tag{2}$$

The objective of statistical *inference* is then to draw conclusions about the value of $\boldsymbol{\theta}$ (and hence about the distribution $P_{\boldsymbol{\theta}}$ of the data) from the observations.

3.2 Bayesian and Bayesian nonparametric models

In Bayesian statistics, all parameters are considered as random variables. Hence under a Bayesian model, data are generated in two stages, i.e.,

$$\boldsymbol{\theta} \sim P(\boldsymbol{\theta}) \tag{3}$$

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\theta} \sim_{iid} P_{\boldsymbol{\theta}}(\mathbf{y}) \tag{4}$$

The objective is then to determine the *posterior distribution* – the conditional distribution of $\boldsymbol{\theta}$ given the observed data,

$$\pi(\boldsymbol{\theta} | \mathbf{y}_1, \ldots, \mathbf{y}_n) \tag{5}$$

A *Bayesian nonparametric* model is a Bayesian model whose parameter space Θ has infinite dimension. To define a Bayesian nonparametric model, we have to define a prior π on an infinite-dimensional space, which is a stochastic process with paths (i.e. realizations) in Θ.

4 Clustering and the Dirichlet process

4.1 Finite mixture models

The basic assumption of a clustering problem is that each observation $\mathbf{y}_i$ belongs to a single cluster $k \in \{1, \cdots, K\}$, which has a cluster distribution

$$P_k(\mathbf{y}_i | z_i = k) \tag{6}$$

where we have defined a latent variable z_i, indicating the cluster assignment of observation $\mathbf{y}_i$. Note that under the Bayesian framework, the latent variable z_i itself has a distribution

$$p_k^i \equiv P(z_i = k) \tag{7}$$

The marginal distribution of the observation $\mathbf{y}_i$ is then

$$P(\mathbf{y}_i) = \sum_{k=1}^{K} P(z_i = k) P_k(\mathbf{y}_i | z_i = k) \tag{8}$$

A model of this form is called a *finite mixture model.*

4.2 Bayesian mixture models

Suppose we know there are K clusters, we first sample the cluster parameters from some base measure:

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim_{iid} G(\boldsymbol{\beta}) \tag{9}$$

We then independently sample the latent cluster assignment vectors and the actual observations:

$$(p_1^i, \ldots, p_K^i) \sim \text{Dirichlet}_K(\alpha) \tag{10}$$

$$z_i \sim \text{Categorical}(p_1^i, \ldots, p_K^i) \tag{11}$$

$$\mathbf{y}_i \sim P_k(\mathbf{y}_i | \boldsymbol{\theta}_k, z_i = k) \tag{12}$$

4.3 Dirichlet Process

Definition 4.1 If $\alpha > 0$ and if G is a probability measure on Ω_ϕ, the random discrete probability measure Θ generated by

$$V_1, V_2, \cdots \sim_{iid} \text{Beta}(1, \alpha) \tag{13}$$

$$C_k = V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{14}$$

$$\Phi_1, \Phi_2, \cdots \sim_{iid} G \tag{15}$$

is called a *Dirichlet process (DP)* with base measure G and concentration α, and denote its law by $\text{DP}(\alpha, G)$.

5 Latent features and the Indian buffet process