SDS 383D: Modeling II **Spring 2017**

Section 1: Preliminaries

1.1 Exchangeability and de Finetti's theorem

A standard situation in statistics is to be presented with a sequence of observations, and use them to make predictions about future observations. In order to do so, we need to make certain assumptions about the nature of the statistical relationships between the sequence of observations.

A common assumption is that our data are **exchangeable**, meaning that their joint probability is invariant under permutations. More concretely, we say a sequence of N observations is finitely exchangeable if $\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_N \in A_n) = \mathbf{P}(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \ldots, X_{\sigma(N)} \in A_n)$ for any permutation of the integers 1 through N, and that an infinite sequence is infinitely exchangeable if this invariance holds for all values of N.

Exercise 1.1 *Clearly, all iid sequences are exchangeable, but not all exchangeable sequences are iid. Consider an urn, containing r red balls and b blue balls. A sequence of colors is generated by repeatedly sampling a ball from the urn, noting its color, and then returning the ball, plus another ball of the same color, to the urn. Show that the resulting sequence is exchangeable, but not iid.*

Proof:

1.1.1 De Finetti's Theorem

Loosely speaking, de Finetti's Theorem states if a sequence of random variables is infinitely exchangeable, those random variables must be conditionally i.i.d. given some set of parameters. More formally,

Theorem 1.1 (de Finetti) *Let $(X_1, X_2, \ldots)$ be an infinite sequence of random variables in some space $\mathcal{X}$. This sequence is infinitely exchangeable if and only if there exists a probability distribution $\mathbf{Q}_\theta$, parametrized by some random parameter $\theta \sim \nu$, such that the X_i are conditionally iid given $\mathbf{Q}_\theta$ and such that*

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots) = \int_\Theta \prod_{i=1}^{\infty} \mathbf{Q}_\theta(A_i)\nu(d\theta).$$

This means we can imagine that any exchangeable sequence has been generated as a sequence of i.i.d. random variables with some unknown law. This provides a motivation for Bayesian inference: We have a hierarchical model, where data are generated according to some distribution parametrized by a random (in the Bayesian context – i.e. unknown/uncertain) variable θ, and our uncertainty about θ is characterized by some distribution ν.

Let's consider the 0/1 form of de Finetti's theorem, for exchangeable sequences of binary variables:

Theorem 1.2 (de Finetti 0/1) *An infinite sequence $(X_1, X_2, \ldots)$ of binary random variables is exchange-*

able if and only if its distribution can be written as

$$\begin{aligned}\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) &= \int_0^1 \prod_{i=1}^N \left\{\theta^{x_i}(1-\theta)^{1-x_i}\right\} d\nu(\theta) \\ &= \int_0^1 \theta^k (1-\theta)^{N-k} d\nu(\theta)\end{aligned}$$

where $k = \sum_i x_i$.

We will now work through (most of) a proof in the next two exercises.

Exercise 1.2 *We will start off with a finite sequence* $(X_1, \ldots, X_M)$. *For any* $N \leq M$, *show that*

$$\mathbf{P}\left(\sum_{i=1}^N X_i = s \middle| \sum_{i=1}^M X_i = t\right) = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}}$$

Proof: By the assumption of exchangeability and given the event $\sum_{i=1}^M X_i = t$, all possible rearrangements of the s ones in among the N places are equally likely. We can think of an urn containing M items, of which t are ones and $M-t$ are zeros. We pick N terms without replacement. Then the probability of obtaining s ones and $N-s$ zeros of the N items is

$$\frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}} = \binom{N}{s}\frac{(t)_s(M-t)_{N-s}}{(M)_N}$$

where $(x)_y = x(x-1)\ldots(x-y+1)$ and which is the probability function of a hypergeometric distribution $Hyp(M, N, t)$.

We can therefore write

$$\mathbf{P}\left(\sum_{i=1}^N X_i = s\right) = \binom{N}{s}\sum_{t=s}^{M-N+s}\frac{(t)_s(M-t)_{N-s}}{(M)_N}\mathbf{P}\left(\sum_{i=1}^M X_i = t\right), \tag{1.1}$$

.

Proof: suppose that $X_1 + X_2 + \ldots + X_N = s \in \{0, 1, 2, \ldots, N\}$. Then exchangeability gives

$$\mathbf{P}\left(\sum_{i=1}^N X_i = s\right) = \binom{N}{s}\mathbf{P}(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(N)})$$

for any permutation σ of $\{1, 2, \ldots, N\}$, i.e.,

$$\mathbf{P}(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(N)}) = \frac{1}{\binom{N}{s}}\mathbf{P}\left(\sum_{i=1}^N X_i = s\right)$$

For arbitrary $M \geq N \geq s \geq 0$, we have by marginalization and Bayes' rule

$$\mathbf{P}\left(\sum_{i=1}^N X_i = s\right) = \sum_{t=s}^{M-(N-s)}\mathbf{P}\left(\sum_{i=1}^N X_i = s \middle| \sum_{i=1}^M X_i = t\right)\mathbf{P}\left(\sum_{i=1}^M X_i = t\right) = \sum_{t=s}^{M-(N-s)}\frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}}\mathbf{P}\left(\sum_{i=1}^M X_i = t\right)$$

Let $F_M(\theta)$ be the distribution function of $\theta = \frac{1}{M}(X_1 + \cdots + X_M)$, i.e. a step function between 0 and 1, with steps of size $\mathbf{P}(\sum_i X_i = t)$ at $t = 0, 1, \ldots, M$. Then we can rewrite Equation 1.1 as

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) = \int_0^1 \frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} dF_M(\theta)$$

Exercise 1.3 *Show that, as $M \to \infty$, we can write*

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) \to \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta)$$

The proof is completed using a result (the Helly Theorem), that shows that any sequence $\{F_M(\theta); M = 1, 2, ...\}$ of probability distributions on [0,1] contains a subsequence that converges to $F(\theta)$.

Proof:

$$\begin{aligned}
\frac{\binom{M\theta}{s}\binom{M(1-\theta)}{N-s}}{\binom{M}{N}} &= \frac{\frac{M\theta!}{s!(M\theta-s)!}\frac{M(1-\theta)!}{(N-s)![M(1-\theta)-(N-s)]!}}{\frac{M!}{N!(M-N)!}} \\
&= \frac{N!}{s!(N-s)!}\frac{M\theta! M(1-\theta)! M!}{(M\theta-s)! M(1-\theta)-(N-s)]! (M-N)!} \\
&\approx \frac{N!}{s!(N-s)!}\frac{(M\theta)^s [M(1-\theta)]^{N-s}}{M^N} \\
&= \binom{N}{s}\theta^s (1-\theta)^{N-s}
\end{aligned}$$

which indicates that

$$\frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} = \frac{\binom{M\theta}{s}\binom{M(1-\theta)}{N-s}}{\binom{M}{N}\binom{N}{s}} = \theta^s (1-\theta)^{N-s}$$

Therefore, we prove that

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) \to \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta)$$

1.2 The exponential family of distributions

De Finetti's theorem can be seen as a motivation for Bayesian inference. If our data are exchangeable, we know that they are iid according to some unknown probability distribution $F_\theta(X)$, which we can think of as a **likelihood function**, and that they can be represented using an mixture of such iid sequences. As we saw from the 0/1 case, the distribution over probabilities is given by the limit of the empirical distribution function. When not working in this limit, we may choose to model this distribution over the parameters of our likelihood function using a **prior** distribution $\pi(\theta)$ – ideally one that both assigns probability mass to where we expect the empirical distribution might concentrate, and for which $\int_\Theta F_\theta(X)\pi(d\theta)$ is tractable.

The exponential family of probability distributions is the class of distributions parametrized by θ whose density can be written as

$$p(x|\theta) = h(x)\exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}$$

where

- $\eta(\theta)$ (sometimes just written as η), is a transformation of θ that is often referred to as the **natural or canonical parameter**.
- $T(X)$ is known as a **sufficient statistic** of X. We see that $p(x|\theta)$ depends only on X through $T(X)$, implying that $T(X)$ contains all the relevant information about X.
- $A(\eta(\theta)$ (or $A(\eta)$) is known as the **cumulant function** or the **log partition function** (remember, a partition function provides a normalizing constant).

Example 1.1 (The Bernoulli distribution) *A Bernoulli random variable X takes the value $X = 1$ with probability π and $X = 0$ with probability $1 - \pi$; it's density can be written:*

$$\begin{aligned} p(x|\pi) =& \pi^x(1-\pi)^{1-x} \\ =& \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \end{aligned}$$

By rewriting in this exponential family form, we see that

- $\eta = \log\left(\frac{\pi}{1-\pi}\right)$
- $T(x) = x$
- $A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$
- $h(x) = 1$

Exercise 1.4 *The Poisson random variable has PDF*

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Re-write the density of the Poisson random variable in exponential family form. What are η, $T(x)$, $A(\eta)$ and $h(x)$? What about if we have n independent samples $x_1, \ldots, x_n$?

Proof:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\exp(x\log\lambda - \lambda)$$

By rewriting in this exponential family form, we see that

- $\eta = \log\lambda$
- $T(x) = x$
- $A(\eta) = \lambda = e^\eta$
- $h(x) = \frac{1}{x!}$

The joint density of i.i.d. Poisson random variables $x_1, \ldots, x_n$ is

$$p(x_1, \ldots, x_n|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} = \frac{1}{\prod_{i=1}^n x_i!}\exp\left(\log\lambda\sum_{i=1}^n x_i - n\lambda\right)$$

So we have

- $\eta = \log \lambda$
- $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$
- $A(\eta) = n\lambda = ne^\eta$
- $h(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$

Exercise 1.5 *The gamma random variable has PDF*

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

What are the natural parameters and sufficient statistics for the gamma distribution, given n observations $x_1, \ldots, x_n$?

Proof:

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \frac{1}{x} \exp\{\alpha \log x - \beta x + \alpha \log \beta - \log \Gamma(\alpha)\}$$

By rewriting in this exponential family form, we see that

- $\eta_1 = \alpha, \eta_2 = -\beta$
- $T_1(x) = \log x, T_2(x) = x$
- $A(\eta_1, \eta_2) = \log \Gamma(\alpha) - \alpha \log \beta = \log \Gamma(\eta_1) - \eta_1 \log(-\eta_2)$
- $h(x) = \frac{1}{x}$

The joint density of i.i.d. Gamma random variables $x_1, \ldots, x_n$ is

$$\begin{aligned} p(x_1, \ldots, x_n|\alpha, \beta) =& \frac{1}{\prod_{i=1}^n x_i} \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^n x_i \right)^\alpha e^{-\beta \sum_{i=1}^n x_i} \\ =& \frac{1}{\prod_{i=1}^n x_i} \exp\left(\alpha \log \prod_{i=1}^n x_i - \beta \sum_{i=1}^n x_i - n \log \Gamma(\alpha) + n\alpha \log \beta \right) \end{aligned}$$

So we have

- $\eta_1 = \alpha, \eta_2 = -\beta$
- $T_1(x_1, \ldots, x_n) = \sum_{i=1}^n (\log x_i), T_2(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$
- $A(\eta_1, \eta_2) = n \log \Gamma(\alpha) - n\alpha \log \beta = n \log \Gamma(\eta_1) - n\eta_1 \log(-\eta_2)$
- $h(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i}$

1.2.1 Cumulants and moments of exponential families

We are probably most familiar with using the PDF or the CDF of a random variable to describe its distribution, but there are other representations that can be useful. The **moment generating function** $M_X(s) = \mathbb{E}[\exp(s^T x)] = \int_{\mathcal{X}} e^{s^T x} p_X(x) dx$ is the Laplace transform of the PDF $p_X(x)$. As the name suggests, we can use the moment-generating function to generate the (uncentered) moments of a random variable; the nth moment is given by

$$m_n = \frac{d^n M_X}{ds^n}\Big|_{s=0}$$

Exercise 1.6 *For exponential family random variables, we know that the sufficient statistic $T(X)$ contains all the information about X, so (for univariate X) we can write the moment generating function of the sufficient statistic as $\mathbb{E}[e^{sT(x)}|\eta]$. Show that the moment generating function for the sufficient statistic of an arbitrary exponential family random variable with natural parameter η can be written as*

$$M_{T(X)}(s) = \exp A(\eta + s) - A(\eta)$$

A related representation is the **cumulant generating function** $C_X(s) = \log \mathbb{E}[e^{s^T x}] = \log(M_X(s))$. Clearly, for exponential families this takes the form $C_{T(X)}(s) = A(\eta+s) - A(\eta)$. This explains why $A(\eta)$ is sometimes called the cumulant function! The cumulant function can be used to generate the cumulants of a distribution as

$$\kappa_n = \frac{d^n C_X}{ds^n}\Big|_{s=0}$$

The first three cumulants are the same as the first three central moments of the distribution – meaning, the cumulant generative function is a useful tool for calculating mean, variance and the third central moment.

Exercise 1.7 *It is usually easier to calculate mean and variance using the cumulant generating function rather than the moment generating function. Starting from the exponential family representation of the Poisson distribution from Exercise 1.4, calculate the mean and variance of the Poisson using a) the moment generating function, and b) the cumulant generating function.*