The multinomial model

Dr. Olanrewaju Michael Akande

Feb 12, 2020

Announcements

- Homework 4 due tomorrow.

Outline

- Categorical data
- Dirichlet distribution
- Conjugacy

Categorical data (univariate)

- Suppose
  - $Y \in \{1, \ldots, d\}$;
  - $\Pr(Y = j) = \theta_j$ for each $j = 1, \ldots, d$; and
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.
- Then the pmf of Y is

$$\Pr[Y = j | \boldsymbol{\theta}] = \prod_{j=1}^{d} \theta_j^{1[Y=j]}.$$

- We say Y has a multinomial distribution with sample size 1, or a categorical distribution.
- Write as $Y|\boldsymbol{\theta} \sim \text{Multinomial}(1, \boldsymbol{\theta})$ or $Y|\boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Clearly, this is just an extension of the Bernoulli distribution.

Dirichlet distribution

- Since the elements of the probability vector $\boldsymbol{\theta}$ must always sum to one, the support is often called a simplex.
- A conjugate prior for categorical/multinomial data is the Dirichlet distribution.
- A random variable $\boldsymbol{\theta}$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, if

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma\left(\sum_{j=1}^{d} \alpha_j\right)}{\prod_{j=1}^{d} \Gamma(\alpha_j)} \prod_{j=1}^{d} \theta_j^{\alpha_j - 1}, \quad \alpha_j > 0 \ \text{ for all } \ j = 1, \ldots, d.$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, and

$$\sum_{j=1}^{d} \theta_j = 1, \ \ \theta_j \geq 0 \ \text{ for all } \ j = 1, \ldots, d.$$

- We write this as $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_d)$.
- The Dirichlet distribution is a multivariate generalization of the beta distribution.

Dirichlet distribution

- Write

$$\alpha_0 = \sum_{j=1}^{d} \alpha_j \quad \text{and} \quad \alpha_j^\star = \frac{\alpha_j}{\alpha_0}.$$

- Then we can re-write the pdf slightly as

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma\left(\alpha_0\right)}{\prod_{j=1}^{d} \Gamma(\alpha_j)} \prod_{j=1}^{d} \theta_j^{\alpha_j - 1}, \quad \alpha_j > 0 \ \text{ for all } \ j = 1, \ldots, d.$$

- Properties:
  - $$\mathbb{E}[\theta_j] = \alpha_j^\star;$$
  - $$\text{Mode}[\theta_j] = \frac{\alpha_j - 1}{\alpha_0 - d};$$
  - $$\mathbb{V}\text{ar}[\theta_j] = \frac{\alpha_j^\star(1 - \alpha_j^\star)}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_j](1 - \mathbb{E}[\theta_j])}{\alpha_0 + 1};$$
  - $$\mathbb{C}\text{ov}[\theta_j, \theta_k] = \frac{\alpha_j^\star \alpha_k^\star}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_j]\mathbb{E}[\theta_k]}{\alpha_0 + 1}.$$

Dirichlet examples

$\text{Dirichlet}(1, 1, 1)$

Dirichlet examples

$\text{Dirichlet}(10, 10, 10)$

Dirichlet examples

$\text{Dirichlet}(10, 10, 10)$

Dirichlet examples

$\text{Dirichlet}(1, 10, 1)$

Dirichlet examples

$\text{Dirichlet}(50, 100, 10)$

Likelihood

- Let $Y_i, \ldots, Y_n | \boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Recall

$$\Pr[Y_i = j | \boldsymbol{\theta}] = \prod_{j=1}^{d} \theta_j^{1[Y_i=j]}.$$

- Then,

$$L[Y; \boldsymbol{\theta}] = \prod_{i=1}^{n} \prod_{j=1}^{d} \theta_j^{1[Y_i=j]} = \prod_{j=1}^{d} \theta_j^{\sum_{i=1}^{n} 1[Y_i=j]} = \prod_{j=1}^{d} \theta_j^{n_j}$$

where n_j is just the number of individuals in category j.

- Maximum likelihood estimate of θ_j is

$$\hat{\theta}_j = \frac{n_j}{n}, \;\; j = 1, \ldots, d$$

Posterior

- Set $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_d)$.

$$\begin{aligned}
\pi(\boldsymbol{\theta}|Y) &\propto L[Y;\boldsymbol{\theta}]\pi[\boldsymbol{\theta}] \\
&\propto \prod_{j=1}^{d} \theta_j^{n_j} \prod_{j=1}^{d} \theta_j^{\alpha_j - 1} \\
&\propto \prod_{j=1}^{d} \theta_j^{\alpha_j + n_j - 1} \\
&= \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_d + n_d)
\end{aligned}$$

- Posterior expectation:

$$\mathbb{E}[\theta_j|Y] = \frac{\alpha_j + n_j}{\sum_{l=1}^{d}(\alpha_l + n_l)}.$$

Combining information

- For the prior, we have

$$\mathbb{E}[\theta_j] = \frac{\alpha_j}{\sum_{j=1}^{d} \alpha_j}$$

- We can think of
  - $\theta_{0j} = \mathbb{E}[\theta_j]$ as being our **"prior guess"** about θ_j, and
  - $n_0 = \sum_{j=1}^{d} \alpha_j$ as being our **"prior sample size"**.
- We can then rewrite the prior as $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(n_0\theta_{01}, \ldots, n_0\theta_{0d})$.

Combining information

- We can write the posterior expectation as:

$$
\begin{aligned}
\mathbb{E}[\theta_j|Y] &= \frac{\alpha_j + n_j}{\sum_{l=1}^d (\alpha_l + n_l)} \\
&= \frac{\alpha_j}{\sum_{l=1}^d \alpha_l + \sum_{l=1}^d n_l} + \frac{n_j}{\sum_{l=1}^d \alpha_l + \sum_{l=1}^d n_l} \\
&= \frac{n_0 \theta_{0j}}{n_0 + n} + \frac{n \hat{\theta}_j}{n_0 + n} \\
&= \frac{n_0}{n_0 + n} \theta_{0j} + \frac{n}{n_0 + n} \hat{\theta}_j.
\end{aligned}
$$

since MLE is

$$
\hat{\theta}_j = \frac{n_j}{n}
$$

- Once again, we can express our posterior expectation as a weighted average of the prior expectation and MLE.
- We can also extend the Dirichlet-multinomial model to more variables (contingency tables).

Example: pre-election polling

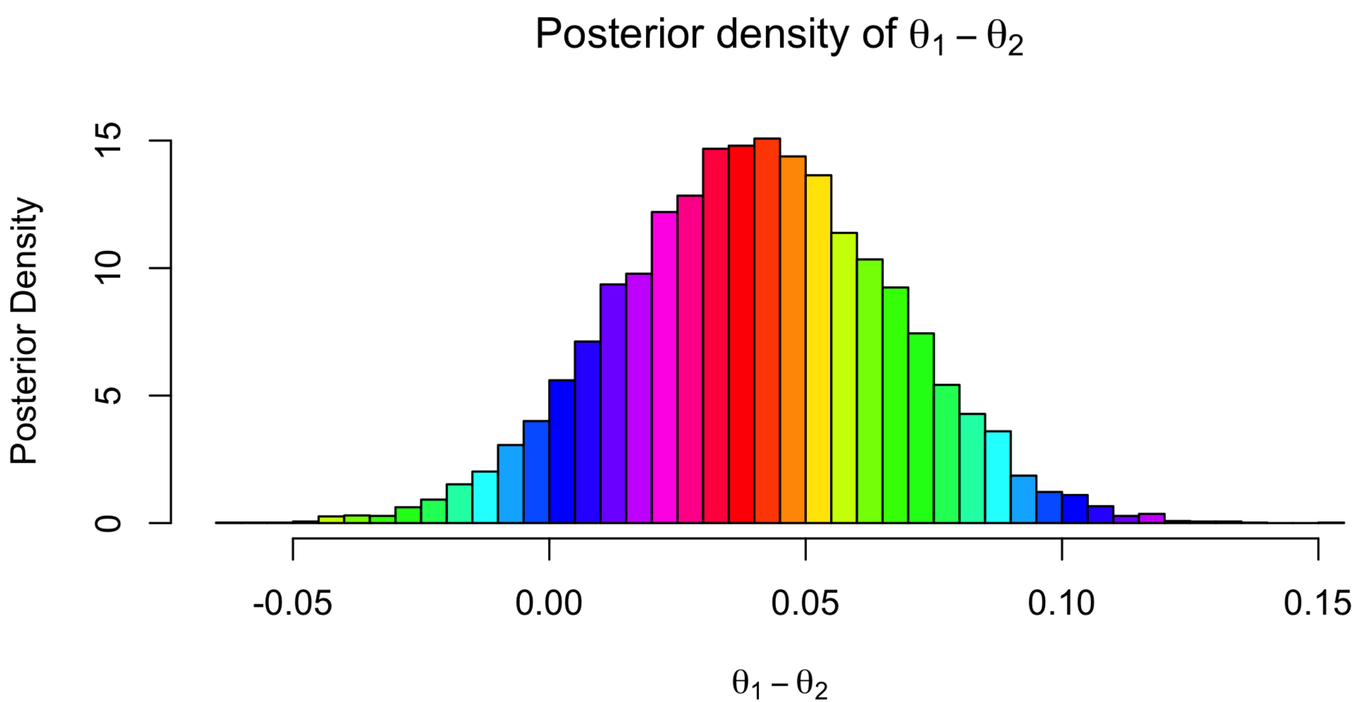
- Fox News Nov 3-6 pre-election survey of 1295 likely voters for the 2016 election.
- For those interested, FiveThirtyEight is an interesting source for pre-election polls.
- Out of 1295 respondents, 622 indicated support for Clinton, 570 for Trump, and the remaining 103 for other candidates or no opinion.
- Drawing inference from pre-election polls is way more complicated and nuanced that this. We only use the data here for this simple illustration.
- Assuming no other information on the respondents, we can assume simple random sampling and use a multinomial distribution with parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, the proportion, in the survey population, of Clinton supporters, Trump supporters and other candidates or no opinion.

Example: pre-election polling

- With a noninformative uniform prior, we have $\pi(\boldsymbol{\theta}) = \mathrm{Dirichlet}(1, 1, 1)$.
- The resulting posterior is $\mathrm{Dirichlet}(1 + n_1, 1 + n_2, 1 + n_3) = \mathrm{Dirichlet}(623, 571, 104)$.
- Suppose we wish to compare the proportion of people who would vote for Trump versus Clinton, we could examine the posterior distribution of $\theta_1 - \theta_2$.
- We can even compute the probability $\Pr(\theta_1 > \theta_2 | Y)$.

Example: pre-election polling

```r
#library(gtools)
PostSamples <- rdirichlet(10000, alpha=c(623,571,104))
#dim(PostSamples)
hist((PostSamples[,1] - PostSamples[,2]),col=rainbow(20),xlab=expression(theta[1]-theta[2])
     ylab="Posterior Density",freq=F,breaks=50,
     main=expression(paste("Posterior density of ",theta[1]-theta[2])))
```



Example: pre-election polling

- Posterior probability that Clinton had more support than Trump in the survey population, that is, $\Pr(\theta_1 > \theta_2 | Y)$, is

```
#library(gtools)
mean(PostSamples[,1] > PostSamples[,2])
```

```
## [1] 0.9345
```

- Once again, this is just a simple illustration with a very small subset of the 2016 pre-election polling data.

- Inference for pre-election polls is way more complex and nuanced that this.