Cross Validated

Stamatics, IIT Kanpur

Alok Kumar (190099), Tarun Kanodia (190902), Nilay Beniwal (190555)

Mini-Project 2

Problem Statement

Suppose we observe multinomial data. That is, let n be a positive integer, and let **p** $= (p_1, \dots, p_K)$ be probabilities so that $\sum_{i=1}^{K} p_i = 1$. Let **x** $= (x_1, .., x_K) \sim$ Multi$(p_1, ..., p_K)$, where **x** has probability mass function

$$Pr(\mathbf{x} = (x_1, ..., x_K)|\mathbf{p}) = \frac{n!}{x_1!..x_k!} \prod_{i=1}^{K} p_i^{x_K}, \quad x_i \in \{0, .., n\} and \sum_{i=1}^{K} x_i = n$$

Intuitively, the multinomial distribution models the number of instances of an ith event out of n trials (where K are the total possible events) and p_i represents the probability of observing an event i.

Typically, we are interested in estimating the parameter **p**. The MLE of p_i can be shown to be

$$\hat{p_i} = x_i n$$

However, we want to use a Bayesian method to estimate **p**. Suppose we assume a Dirichlet prior on **p**, so that **p** $\sim$ Dir$(\alpha_1, ..., \alpha_K)$, where $\alpha_i > 0$ for $i = 1, ..., K$, with probability density function

$$f(p_1, ..p_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \quad p_i \in (0, 1) and \sum_{i=1}^{K} p_i = 1$$

1. What is the posterior distribution of **p** having observed the data **x**? Write all the steps.

2. What is the posterior mean of **p**? Write this posterior mean as a convex combi-

nation of the prior mean and the MLE. What happens to the posterior mean of **p** as n increases?

3. The popular website "IMDB" has a database of movies, and a summary of their respective ratings. Users rate different movies on the website on a scale of 110, 1 being bad and 10 being great.

   Suppose n users rate a given movie. Let x_i be the observed number of people who gave rating i. Let R be the average rating the movie has received.

   Now, IMDB has a popular "Top 250" movies of all time list. However, due to varying number of votes for different movies from different eras/countries, IMDB uses the following "Bayesian average" to obtain a rating:

   $$Rating \; = nn + mR + mn + mC,$$

   where
   $R =$ actual average rating of the movie
   $n =$ number of votes for the movie
   $m =$ minimum votes required to be listed in the Top Rated list (2500)
   $C = 5.5$

   (a) Explain how the above rating can be obtained from the model presented in parts 1 and 2. What values of α_i have been chosen here? Write out all mathematical steps.

   (b) We will use this system to rank Bollywood movies. Load the dataset of movies using the line below in R:
   data <- read.csv("bollywood.csv")
   Note: You must have the dataset saved in the same folder as R/Python script to load the dataset. Here imdb id = ID of the movie on IMDB. For example if the id is $tt4934950$, you can access the movie page on $https : //www.imdb.com/title/tt4934950/$.
   IMDB rating = the rating of the movie on IMDB.
   IMDB votes = the number of votes given to the movie.
   Q: Generate a "Top 10" list according to the IMDB ranking system. Write

down the IMDB id of these 10 movies. (Remember to share code for this part.)

Solution

1. The likelihood is:

$$L(p,x) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} p_i^{x_i}$$
$$\propto \prod_{i=1}^{K} p_i^{x_i}$$

The prior is a Dirichlet distribution which has a pdf proportional to

$$\prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

Therefore, the posterior pdf is proportional to:

$$\prod_{i=1}^{K} p_i^{x_i} \times \prod_{i=1}^{K} p_i^{\alpha_i - 1} = \prod_{i=1}^{K} p_i^{\alpha_i + x_i - 1}$$

This is proportional to pdf of a Dirichlet distribution with parameters $\alpha_1+x_1, \alpha_2+x_2, ...\alpha_K + x_K$.

2. The mean of Dirichlet distribution is

$$\begin{aligned}
E(p_j) &= \int ... \int p_j \frac{\Gamma(\alpha)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} dp_1..dp_{K-1} \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} \frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \int .. \int \frac{\Gamma(\alpha+1)}{\prod_{i=1}^{K} \Gamma(\alpha'_i)} \prod_{i=1}^{K} p_i^{\alpha'_i - 1} dp_1...dp_{K-1} \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} \frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \\
&= \frac{\alpha_j}{\alpha}
\end{aligned}$$

where $\alpha'_i = \alpha_i$ when $i \neq j$ and $\alpha'_j = \alpha_j + 1$, and $\alpha = \sum_{i=1}^{K} \alpha_i$.

Thus, the posterior mean is given by:

$$\begin{aligned} E[p_i|x,\alpha] &= \frac{\alpha_i + x_i}{n + \sum_{l=1}^{k} \alpha_l} \\ &= \kappa \frac{\alpha_i}{\sum_{l=1}^{K} \alpha_l} + (1-\kappa)\hat{p}_i \end{aligned}$$

where $\hat{p_i} = \frac{x_i}{n}$ is the maximum likelihood estimator of p_i and $\kappa = \frac{\sum_l \alpha_l}{n + \sum_l \alpha_l} \in (0,1)$. We can see that the posterior mean is a convex combination of prior mean and the maximum likelihood estimate of p_i. Also, as $n \to \infty$, it can be seen that the posterior mean concentrates around the maximum likelihood estimate, because $\kappa \to 0$.

3. (a)

   (b) IMDB IDs of Top 10 movies are:

      i. tt2338151
      ii. tt5074352
      iii. tt1188996
      iv. tt1954470
      v. tt1954470
      vi. tt2082197
      vii. tt3863552
      viii. tt1562872
      ix. tt4430212
      x. tt2574698