Stamatics Mini Project - II

May 2021

Concepts Required: Maximum Likelihood Estimation, Linear Regression, Standard Normal Distribution, Bernoulli Distribution.

1 Submission Directions

PDF

1. The parts (1), (2), and (3a) of the problem statement have to be submitted as a PDF outlining all the mathematical details of the algorithm.
2. Mention only the answers of part (3b) in the PDF obtained from the code. Do not paste the code in the PDF.

Code

1. Submit a well-commented R/Python script for part (3b).
2. Use of external packages for direct implementation of the model is not allowed.

2 Problem Statement

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

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

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 $\mathbf{p}$. 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 $\mathbf{p}$. Suppose we assume a Dirichlet prior on $\mathbf{p}$, so that $\mathbf{p} \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$, where $\alpha_i > 0$ for $i = 1, \ldots, K$, with probability density function

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

1. What is the posterior distribution of $\mathbf{p}$ having observed the data $\mathbf{x}$? Write all the steps.

2. What is the posterior mean of $\mathbf{p}$? Write this posterior mean as a convex combination of the prior mean and the MLE. What happens to the posterior mean of $\mathbf{p}$ 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 $1 - 10$, 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 (a) and (b). What values of α_i have been chosen here?

Write out all the 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.)