Machine Learning: Homework 1

Sen Wang

January 30th, 2019

1 Beta Updating from centored likelihood and MLE for uniform distribution

Part A

The likelihood of obtaining X heads with 5 tosses is,

$$p(X|\theta) = \theta^X(1-\theta)^{5-X} \tag{1}$$

and the prior probability of heads is,

$$p(\theta) = \text{Beta}(\theta|1,1) = 1 \tag{2}$$

Therefore, the posterior required can be simply calculated as,

$$p(\theta|X<3) = p(X<3|\theta)p(\theta) \tag{3}$$

$$= \sum_{x=0}^{2} p(x|\theta)p(\theta) \tag{4}$$

$$= \sum_{x=0}^{2} \theta^x(1-\theta)^{5-x} \tag{5}$$

Part B

a). What is the MLE of a?

The MLE of a provides the maximum likelihood for the given dataset, therefore, the value of a is,

$$a^{\text{MLE}} = \max(|x_n|) \tag{6}$$

b). What probability would the model assign to $\hat{x}_n + 1$ using the MLE estimate of a?

Given the MLE estimate of a the new value is,

$$p(\hat{x}_{n+1}) = \frac{1}{2a^{\text{MLE}}}\mathbb{I}(x \in [-a^{\text{MLE}}, a^{\text{MLE}}]) \tag{7}$$

c). Do you see any problem with the above approach? If yes, briefly suggest a better alternative.

Yes, the problem is that the probability of any data outside the range set by the MLE estimate will be zero. A better alternative is to apply a conjugate prior to the likelihood model and obtain a posterior predictive probability distribution for the new data.

2 Bayesian Analysis for Uniform distribution and the Taxicab problem

2.1 Part A

The posterior $p(\theta|\mathcal{D})$ can be obtained as,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D},\theta)}{p(\mathcal{D}} \tag{8}$$

$$= \begin{cases} \frac{Kb^K}{\theta^{N+K+1}} \frac{(N+K)b^N}{K} & \text{if } m \leq b \\ \frac{Kb^K}{\theta^{N+K+1}} \frac{(N+K)m^N}{K} & \text{if } m > b \end{cases} \tag{9}$$

$$= \frac{(N+K)c^{N+K}}{\theta^{N+K+1}} \tag{10}$$

$$= Pa(\theta|c, N+K) \tag{11}$$

where $c = \max(b, m)$.

2.2 Part B

a). Suppose we see one taxi numbered 100. Using an non-informative prior on θ of the form $p(\theta) = \boldsymbol{Pa}(\theta|0,0) \propto \frac{1}{\theta}$, what is the posterior $p(\theta|\mathcal{D})$

From part A, given that $b = 0, K = 0, N = 1, m = 100$, the posterior probability is simplified as,

$$p(\theta|\mathcal{D}) = Pa(\theta|100, 1) \tag{12}$$

$$= \frac{100}{\theta^2}\mathbb{I}(\theta \geq 100) \tag{13}$$

b). Compute the posterior mean, mode and median number of taxis in the city, if such quantities exist.

The mean of the posterior is computed as,

$$\mathbb{E}(\theta|\mathcal{D}) = \int_{100}^{\infty} \theta \frac{100}{\theta^2} d\theta \tag{14}$$

$$= \int_{100}^{\infty} \frac{100}{\theta} d\theta \tag{15}$$

$$= 100[\ln \theta]_{100}^{\infty} \tag{16}$$

$$= \infty \tag{17}$$

Therefore, the posterior mean does not exist.

For the posterior mode,

$$\text{Mode}(\theta|\mathcal{D}) = \text{argmax}_\theta[p(\theta|\mathcal{D}] \tag{18}$$

$$= 100 \tag{19}$$

which is the the MLE estimate of θ.

For the posterior median, it is calculated by value of θ before and after which the cumulative probability is 0.5, and is calculated as,

$$\int_{100}^{m} \frac{100}{\theta^2} = 0.5 \tag{20}$$

where m is the posterior median, from which we can easily obtain that,

$$m = 200 \tag{21}$$

c). Compute the predictive density over the next taxicab number using Part A of this problem. Then consider the case where $b = K = 0$.

The posterior predictive density, given $N = 1$, is calculated as,

$$p(x|\mathcal{D}, \alpha) = \begin{cases} \int_c^\infty p(x|\theta)p(\theta|\mathcal{D}, \alpha)d\theta & \text{if } x \leq c \\ \int_x^\infty p(x|\theta)p(\theta|\mathcal{D}, \alpha)d\theta & \text{if } x > c \end{cases} \tag{22}$$

$$= \begin{cases} \frac{(1+K)}{(2+K)c} & \text{if } x \leq c \\ \frac{(1+K)c^{K+1)}}{(2+K)x^{K+2}} & \text{if } x > c \end{cases} \tag{23}$$

If non-informative prior is used, where $b = K = 0$, then we have,

$$p(x|\mathcal{D}, \alpha) = \begin{cases} \frac{1}{2m} & \text{if } x \leq c \\ \frac{m}{2x^2} & \text{if } x > c \end{cases} \tag{24}$$

d). Use the predictive density formula to compute the probability that the next taxi you will see has number 100, 50, or 150.

Suppose, as in part a), the taxi number we saw is 100, then the probability to see the taxi numbers are,

$$p(x = 100|\mathcal{D}, \alpha) = \frac{1}{200} \tag{25}$$

$$p(x = 50|\mathcal{D}, \alpha) = \frac{1}{100} \tag{26}$$

$$p(x = 150|\mathcal{D}, \alpha) = \frac{100}{2 \times 150^2} = \frac{1}{450} \tag{27}$$

e). Briefly describe some ways we might make the model more accurate at prediction.

First, we can apply an informative prior based on our initial knowledge about the total number of taxi in the city. Secondly, the model can also be improved by applying a parameter prior and then use empirical Bayes method.

3 Naive Bayes

Part A

A Naive Bayes model (multivariate Bernoulli version) for the spam classification problem can be presented as follows.

First, the Naive Bayes assumption requires that,

$$p(\mathbf{x}|y=c,\theta)=\prod_j p(x_j|y=c,\theta_{jc}) \tag{28}$$

where in this case, x_j is individual word in the vocabulary, y is the class state (spam or non-spam), and θ_{jc} is the parameter associating the probability of x_j given the state $y=c$.

Then, the multivariate Bernoulli distribution applies to the individual probability distribution. From which we are able to obtain the joint probability for the features and the class. And the joint log-likelihood is then given by,

$$\log p(\mathcal{D}|\theta)=\sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D}\sum_{c=1}^{C}\sum_{i:y_i=c} \log p(x_{ij}|\theta jc) \tag{29}$$

where N_c is the number of times class label c occurs and π_c is the class prior for class label c.

The MLE can then be calculated through taking derivatives with respect to the intended variables.

The class prior MLE is,

$$\hat{\pi_c}=\frac{N_c}{N} \tag{30}$$

and in our case, all features are binary ($x_j \in (0,1)$), therefore the MLE for the model parameters are,

$$\hat{\theta_{jc}}=\frac{N_{jc}}{N_c} \tag{31}$$

where N_{jc} corresponds to the number of occurrences of class label c where feature x_j is on.

Now, to answer the original question, the parameters required can be calculated by counting the occurrences, and the results are shown below.

$$\theta_{\text{spam}}=\frac{N_{\text{spam}}}{N}=\frac{3}{7} \tag{32}$$

$$\theta_{\text{secret|spam}}=\frac{N_{\text{secret|spam}}}{N_{\text{spam}}}=\frac{2}{3} \tag{33}$$

$$\theta_{\text{secret|non-spam}}=\frac{N_{\text{secret|non-spam}}}{N_{\text{non-spam}}}=\frac{1}{4} \tag{34}$$

$$\theta_{\text{sports|non-spam}}=\frac{N_{\text{sports|non-spam}}}{N_{\text{non-spam}}}=\frac{1}{2} \tag{35}$$

$$\theta_{\text{dollar|spam}}=\frac{N_{\text{dollar|spam}}}{N_{\text{spam}}}=\frac{1}{3} \tag{36}$$

$$\tag{37}$$

3.1 Part B

The training procedure involves calculating the MLE parameters, including class prior and the feature probability, as shown in Part A. The posterior predictive probability for classification is simplified to the product the MLE estimates for the prior and feature probability, as shown below,

$$p(y_i=c|\mathbf{x}_i,\mathcal{D})=p(y_i=c|\pi^{\text{MLE}})P(\mathbf{x}_i|y_i=c,\theta^{\text{MLE}}) \tag{38}$$

After the computation (script as shown in the Zip file), the misclassification rate for both datasets are,

$$\text{err}_1=19.4\% \tag{39}$$

$$\text{err}_2=33.3\% \tag{40}$$

The accuracy can be further improved if the parameters are not estimated with MLE but using MAP estimates, or obtain the classification probability through marginalisation of the posterior probability and the prior.

4 Empirical Bayes

Given the likelihood model and the prior model, the posterior can be expressed as,

$$p(\theta|\mathcal{D},\eta) \propto \prod_{i}^{N} p(x_i|\theta_i)p(\theta_i|\eta) \tag{41}$$

where $\eta = (\alpha, \beta)$. The hyperparameters can be estimated using the fixed point method proposed by Minka and the values are,

$$\alpha = 0.8269 \tag{42}$$
$$\beta = 607.5 \tag{43}$$

Now, the posterior mean for each city i can be calculated as,

$$\theta_i = \lambda_i m_{1i} + (1 - \lambda_i)\hat{\theta}^{\text{MLE}} \tag{44}$$

where,

$$\lambda_i = \frac{\alpha_0}{\alpha_0 + N_i} \tag{45}$$
$$\alpha_0 = \alpha + \beta \tag{46}$$
$$m_{1i} = \frac{\alpha}{\alpha_0} \tag{47}$$

After converting above equations into computation, the results for the posterior means, MLE estimate and MAP estimate are shown below. The MATLAB code for the computation is in the Zip File.

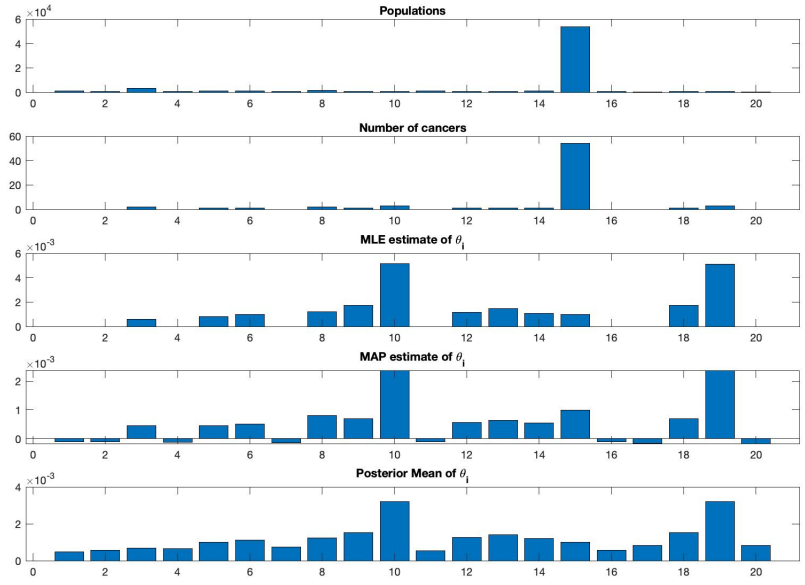

Figure 1: Bars plots of population, cancer cases, MLE estimate, MAP estimate, and posterior mean of θ_i

5 Reject option in classifiers and Newsvendor problem

Part A

a). Show the condition for the minimum risk.

The minimum risk requires two conditions. Firstly, the probability for classifying option j is larger than any other option, which translates mathematically to,

$$p(Y = j|x) \geq p(Y = k|x) \ \forall k \tag{48}$$

Secondly, the loss of the reject option is larger than the weighted loss that j is the wrong choice, which translates mathematically to,

$$\lambda_r \geq \lambda_s(1 - p(Y = j|x)) \tag{49}$$

$$p(Y = j|x) \geq 1 - \frac{\lambda_r}{\lambda_s} \tag{50}$$

5.0.1 b). Describe qualitatively what happends as λ_r/λ_s is increased from 0 to 1.

As the relative cost of rejection increases, or approaches to 1, assume the classification model has the same performance, the likelihood to choose the reject option decreases.

Part B

The expected profit, as specified in the question statement, is,

$$\mathbb{E}_\pi = \int_Q^\infty (P - C)Qf(D)dD + \int_0^Q (P - C)Df(D)dD - \int_0^Q C(Q - D)f(D)dD \tag{51}$$

The extremum of the expected profit can be obtained by equating its first derivative respect to Q to zero, using the Leibniz rule, which is that for the function $y(x)$ of the form

$$y(x) = int_{a(x)}^{b(x)} f(x, t)dt \tag{52}$$

its derivative with respect to x is,

$$\frac{dy(x)}{dx} = f(x, b(x))\frac{db(x)}{dx} - f(x, a(x))\frac{da(x)}{dx} + \int_{a(x)}^{b(x)} \frac{\partial f(x, t)}{\partial x}dt \tag{53}$$

Therefore, we can obtain,

$$\frac{\partial \mathbb{E}_\pi}{\partial Q} = [-(P - C)f(Q)Q + (P - C)(1 - F(Q))] + [(P - C)f(Q)Q - CF(Q)] \tag{54}$$

$$= (P - C) - PF(Q) \tag{55}$$

$$= 0 \tag{56}$$

Therefore, we obtain the final result that the optimal quantity Q^* satisfies,

$$F(Q^*) = \frac{P - C}{P} \tag{57}$$