CSE512 Fall 2019 - Machine Learning - Homework 2

September 21, 2019

Name: Anmol Shukla

Solar ID: 112551470

NetID: anmshukla

Email address: Anmol.shukla@stonybrook.edu

Names of people whom I have discussed the homework with: None

1 Question 1 — Parameter Estimation

1.1 MLE

Part 1: Give the log-likelihood function of X given λ

We know that the discrete random variable X follows a Poisson distribution with parameter λ -

$$P(X = k|\lambda) = \frac{\lambda}{k!}e^{-\lambda}$$

The likelihood function of X given λ is -

$$P(X_1, X_2..., X_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda}{k!}e^{-\lambda}$$

Taking log on both the sides, we have the log-likelihood function-

$$\begin{aligned}
\ln P(X_1, X_2..., X_n|\lambda) &= \ln\left(\prod_{i=1}^{n} \frac{\lambda}{k!}e^{-\lambda}\right) && (1)\\
&= \sum_{i=1}^{n} \ln\left(\frac{\lambda}{k!}e^{-\lambda}\right) && (2)\\
&= \sum_{i=1}^{n} \left(-\lambda \ln(e) + \ln(\frac{1}{x_i!}) + x_i \ln(\lambda)\right) && (3)\\
&= \sum_{i=1}^{n} (-\lambda - \ln(x_i) + x_i \ln(\lambda)) && (4)\\
&= -n\lambda - \sum_{i=1}^{n} (\ln(x_i) - x_i \ln(\lambda)) && (5)
\end{aligned}$$

Part 2: Compute the MLE for λ in the general case

We can find λ_{MLE} by differentiating w.r.t. λ

$$\frac{dP}{d\lambda} = 0 \tag{6}$$

$$\frac{d}{d\lambda_{MLE}}\left(-n\lambda_{MLE} - \sum_{i=1}^{n}\left(\ln(x_i) - x_i\ln(\lambda_{MLE})\right)\right) = 0 \tag{7}$$

$$-n + \frac{1}{\lambda_{MLE}}\sum_{i=1}^{n} x_i = 0 \tag{8}$$

$$\lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{9}$$

Part 2: Compute the MLE for λ using the observed case

Putting the values for X (wait time for calling Uber car) in eq(9), we have -

$$\lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{4+12+3+5+6+9+10}{7} = 7$$

1.2 MAP

Part 1: Computer the posterior distribution over λ

$$
\begin{aligned}
P(\lambda \mid D) =& P(\lambda) \cdot P(D \mid \lambda) \\
=& \left(\prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
=& e^{-n\lambda} \lambda^{\sum_{i=1}^{n} X_i} \left(\prod_{i=1}^{n} \frac{1}{X_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
=& e^{-n\lambda} e^{-\beta\lambda} \lambda^{\sum_{i=1}^{n} X_i} \lambda^{\alpha-1} \left(\prod_{i=1}^{n} \frac{1}{X_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \\
=& e^{-n\lambda-\beta\lambda} \lambda^{\sum_{i=1}^{n} X_i+\alpha-1} \left(\prod_{i=1}^{n} \frac{1}{X_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \\
=& \lambda^{\sum_{i=1}^{n} X_i+\alpha-1} e^{-\lambda n-\beta\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \prod_{i=1}^{n} \frac{1}{X_i!} \\
=& \lambda^{\sum_{i=1}^{n} X_i+\alpha-1} e^{-\lambda n-\beta\lambda} \\
=& \lambda^{\sum_{i=1}^{n} X_i+\alpha-1} e^{-\lambda(n+\beta)}
\end{aligned}
$$

Thus the posterior distribution over λ is the Gamma distribution - $p(\lambda; \sum_i X_i + \alpha, n+\beta)$

Part 2: Derive an analytic expression for MAP of λ

From part 1, we know that the posterior distribution for λ is a gamma distribution with parameters $(\sum_i X_i + \alpha, n+\beta)$. We have also been given in the question that the mode for a gamma distribution is $\frac{(\alpha-1)}{\beta}$. For any probability distribution, we know that it will have the highest probability at its mode. Substituting the value of $\alpha = \sum_{i=1}^{n} X_i + \alpha$ and $\beta = n + \beta$ in the mode of a gamma distribution, we have -

$$
\text{MAP of } \lambda = \frac{\sum_{i=1}^{n} X_i + \alpha - 1}{n+\beta}
$$

1.3 Estimator Bias

Part 1 We have been given that $\eta = e^{-2\lambda}$ where λ comes from Poisson distribution. Taking ln on both the sides we have -

$$\begin{aligned}\ln(\eta) =& -2\lambda\ln(e)\\ \lambda =& -\frac{\ln(\eta)}{2}\end{aligned}$$

We know that $X \sim Poisson(\lambda)$, therefore we can substitute λ in the log-likelihood function of X given λ in eq(5)

$$\begin{aligned}P(X|\lambda) =& -n\lambda - \sum_{i=1}^{n}(\ln(x_i) - x_i\ln(\lambda))\\ P(X|\eta) =& -n(-\frac{\ln(\eta)}{2}) - \sum_{i=1}^{n}(\ln(x_i) - x_i\ln(-\frac{\ln(\eta)}{2}))\end{aligned}$$

Since we have a single observation X -

$$P(X|\eta) = \frac{\ln(\eta)}{2} - \ln(X) + X\ln(-\frac{\ln(\eta)}{2})$$

Differentiating w.r.t. η and setting it to 0, we can get η_{MLE}

$$\begin{aligned}\frac{d}{d\eta}P(X|\lambda) =& 0\\ \frac{d}{d\eta}P(X|\lambda) =& \frac{1}{2\eta} + \frac{X}{\eta\ln\eta}\\ \frac{1}{2\eta} + \frac{X}{\eta\ln\eta} =& 0\\ \frac{X}{\eta\ln\eta} =& -\frac{1}{2\eta}\\ \ln(\eta) =& -2X\\ \eta_{MLE} =& e^{-2X}\end{aligned}$$

Hence, $\hat{\eta} = e^{-2X}$ is the MLE for η

Part 2 For bias of $\hat{\eta}$ we first need $E[\eta]$

$$
\begin{aligned}
E[\hat{\eta}] &= \sum_{X=0}^{\infty} \hat{\eta} P(X) \\
&= \sum_{X=0}^{\infty} e^{-2X} P(X) \\
&= \sum_{X=0}^{\infty} e^{-2X} \frac{\lambda^X e^{-\lambda}}{X!} \\
&= e^{-\lambda} \sum_{X=0}^{\infty} e^{-2X} \frac{\lambda^X}{X!} \\
&= e^{-\lambda} \sum_{X=0}^{\infty} \frac{\frac{\lambda^X}{e^{2X}}}{X!} \\
&= e^{-\lambda} \sum_{X=0}^{\infty} \frac{(\frac{\lambda}{e^2})^X}{X!}
\end{aligned}
$$

We can use the Taylor expansion to solve the above equation $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$$
\begin{aligned}
E[\hat{\eta}] &= e^{-\lambda} \sum_{X=0}^{\infty} \frac{(\frac{\lambda}{e^2})^X}{X!} \\
&= e^{-\lambda} e^{\lambda/e^2} \\
&= e^{-\lambda+\lambda/e^2} \\
&= e^{-\lambda+\lambda/e^2} \\
&= e^{-\lambda(1-1/e^2)}
\end{aligned}
$$

Therefore, the bias of $\eta = e^{-2\lambda}$ is $e^{-\lambda(1-1/e^2)} - e^{-2\lambda}$

Part 3 Let us say that $\hat{\eta} = (-1)^X$. Now, let us find $E[\hat{\eta}]$.

$$\begin{aligned}
E[\hat{\eta}] &= \sum_{X=0}^{\infty} (-1)^X P(X) \\
&= \sum_{X=0}^{\infty} (-1)^X \frac{\lambda^X e^{-\lambda}}{X!} \\
&= e^{-\lambda} \sum_{X=0}^{\infty} (-1)^X \frac{\lambda^X}{X!} \\
&= e^{-\lambda} \sum_{X=0}^{\infty} \frac{(-\lambda)^X}{X!} \\
&= e^{-\lambda} e^{-\lambda} \quad \text{(using Taylor expansion)} \\
&= e^{-2\lambda}
\end{aligned}$$

Therefore, $\eta - E[\hat{\eta}] = e^{-2\lambda} - e^{-2\lambda} = 0$. Thus, the estimator $(-1)^X$ is unbiased. But even though it is unbiased, it is not a good estimator mainly because it changes with data frequently i.e. it is sensitive to the data. Thus the only unbiased estimator estimates $e^{-2\lambda}$ to be 1 if X is even, -1 if X is odd.

1.4 3.2

Question 1

Following are the values for RMSE (the values have been rounded-off to 4 decimal places) -

Lambda	**RMSE Training**	**RMSE Validation**	**RMSE LOOCV (training)**
0.01	1.1776	2.4874	2.5160
0.1	1.2646	2.1351	2.1704
1	1.5942	1.9953	2.0094
10	2.1926	2.3482	2.3196
100	2.9713	3.0173	2.9965
1000	3.3317	3.3454	3.3353

Following is the plot for train, validation and leave-one-out-cross-validation RMSE:

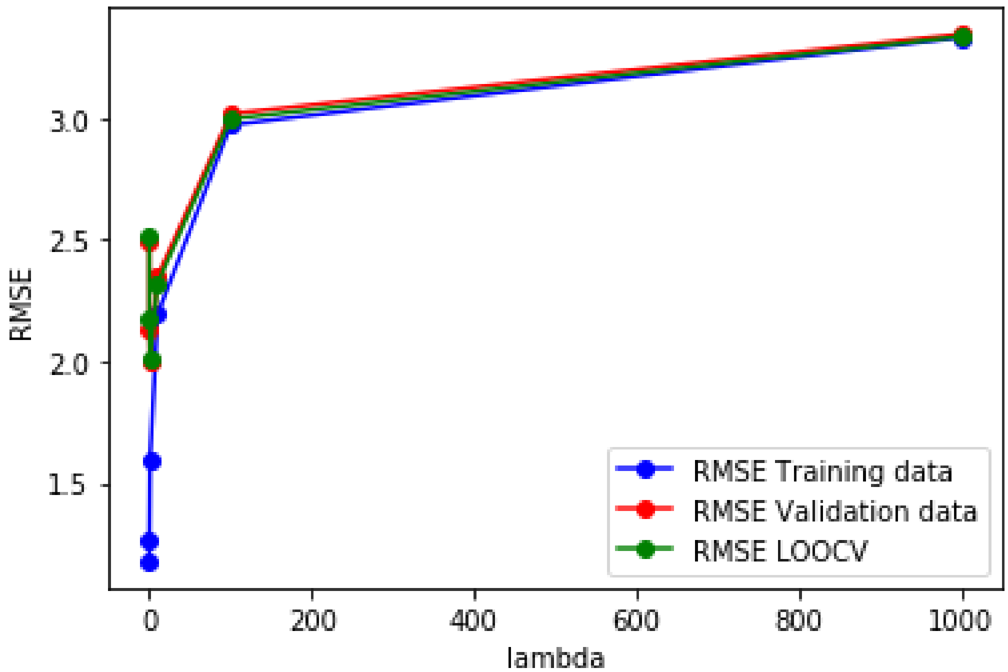

Figure 1: **Step 3** for obtaining Z_{11}. We subtract P_{22} from previous result.

Question 2

The λ that achieves the best LOOCV performance is $\lambda = 1$.
The objective value is defined as $\lambda||w||^2 + \sum_{i=1}^{n}(w^T x_i + b - y_i)^2$ which is equal to 24463.0632. The value of regularization term ($\lambda||w||^2$) is 4656.71487 and the sum of square errors is 12450.5013

Question 3

The weights for the top 10 most important features have the highest weights - flavors nice, soft, low alcohol, cuts, yeast, spices, relatively, love, appealing, cherries plums. It intuitively makes sense because there are the properties of a wine that would be rated high/low.

The weights for the leat 10 important features have weights close to 0 - price dry, cocktail, currant cola, future, new french, little heavy, sweet black, red, pineapple orange, infused. Most of these features do not correlate well with the goodness/badness of a wine.

Question 4

Please see the attached predTestLabels.csv and the python code file.