# Foundations of Machine Learning Assignment 1

**Arkadipta De**
Indian Institute of Technology
Hyderabad
AI20MTECH14002
ai20mtech14002@iith.ac.in

**Venkatesh E**
Indian Institute of Technology
Hyderabad
AI20MTECH14005
ai20mtech14005@iith.ac.in

**Abstract**

This assignment provides answers to the questions on Linear Regression, Regularized Multi-output Linear Regression. It also provides descriptions and observations for Poisson maximum likelihood estimation and Poisson maximum a-posteriori estimation to predict deaths of Prussian Soldiers due to horse kick. This document also provides description and observation to a Poisson Regression problem.

## 1 Question 1

From the description, after adding Gaussian noise $\epsilon_n$ with zero mean and $\sigma$ variance independently to each of the input variables $x_n$ we get the following linear model,

$$y'(x_n, w) = w_0 + \sum_{d=1}^{D} w_d(x_{nd} + \epsilon_{nd}) \tag{1}$$

$$= w_0 + \sum_{d=1}^{D} w_d x_{nd} + \sum_{d=1}^{D} w_d \epsilon_{nd} \tag{2}$$

$$= y(x_n, w) + \sum_{d=1}^{D} w_d \epsilon_{nd} \tag{3}$$

In equation (3), the noise $\epsilon_{nd}$ is independent across both the data-points ($n$) and features ($d$). Hence new error function becomes,

$$E'_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{y'(x_n, w) - t_n\}^2 \tag{4}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{y(x_n, w) + \sum_{d=1}^{D} w_d \epsilon_{nd} - t_n\right\}^2 \tag{5}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n)\left\{\sum_{d=1}^{D} w_d \epsilon_{nd}\right\} + \left\{\sum_{d=1}^{D} w_d \epsilon_{nd}\right\}^2\right\} \tag{6}$$

Taking expectation in equation (6) and using linearity of expectation we get,

$$\mathbf{E}[E'_D(w)] = \frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n)\left\{\sum_{d=1}^{D} w_d \mathbf{E}[\epsilon_{nd}]\right\} + \mathbf{E}\left[\left\{\sum_{d=1}^{D} w_d \epsilon_{nd}\right\}^2\right]\right\} \tag{7}$$

As, $\mathbf{E}[\epsilon_{nd}] = 0$ hence from the third term in equation (7) we get,

$$\mathbf{E}\left[\left\{\sum_{d=1}^{D} w_d \epsilon_{nd}\right\}^2\right] = \mathbf{E}\left[\sum_{d=1}^{D}\sum_{d'=1}^{D} w_d w_{d'} \epsilon_{nd}\epsilon_{nd'}\right] \tag{8}$$

$$= \sum_{d=1}^{D}\sum_{d'=1}^{D} w_d w_{d'} \mathbf{E}[\epsilon_{nd}\epsilon_{nd'}] \tag{9}$$

$$= \sum_{d=1}^{D}\sum_{d'=1}^{D} w_d w_{d'} \delta_{dd'} \tag{10}$$

$$= \sum_{d=1}^{D} w_d^2 \sigma^2 \tag{11}$$

Using the result from equation (11) in equation (7) we get,

$$\mathbf{E}[E'_D(w)] = \frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n, w) - t_n)^2 + \sum_{d=1}^{D} w_d^2 \sigma^2\right\} \tag{12}$$

$$= E_D(w) + \frac{N\sigma^2}{2}\sum_{d=1}^{D} w_d^2 \tag{13}$$

From equation (13) we observe that, we get $L_2$ regularization term $\lambda = N\sigma^2$, without any bias parameter $w_0$ which is the desired result.

## 2 Question 2

### 2.1 Section 1

The hypothesis is given by,

$$y(x, \mathbf{W}) = \mathbf{W^T}\Phi(x) \tag{1}$$

Where $y$ is a $k \times 1$ dimensional target vector, $\mathbf{W}$ is $M \times k$ parameter matrix and $\Phi(x)$ is an $M \times 1$ vector having elements $\Phi_j(x)$ with $\Phi_0(x) = 1$. Now, we take the conditional distribution of $y$ to be an isotropic Gaussian of the form,

$$p(y|x, \mathbf{W}, \beta) = \mathcal{N}(y|\mathbf{W^T}\Phi(x), \beta^{-1}\mathbf{I}) \tag{2}$$

Where $\beta$ is the precision or inverse variance. Now, if we have a set of $N$ observations $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_N}$, we can combine these into a single matrix $\mathbf{y}$ of dimension $N \times k$ such that the $n$th row is given by $\mathbf{y_n^T}$. Again, we can combine the input vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ into design matrix $\mathbf{X}$ with dimension $M \times k$. Hence, from equation (2), the likelihood function is then given by,

$$p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{W^T}\Phi(x_n), \beta^{-1}\mathbf{I}) \tag{3}$$

Taking log on both side of equation (3) we get,

$$ln\{p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \beta)\} = \sum_{n=1}^{N} ln\{\mathcal{N}(y_n|\mathbf{W^T}\Phi(x_n), \beta^{-1}\mathbf{I})\} \tag{4}$$

$$= \frac{Nk}{2} ln\left\{\frac{\beta}{2\pi}\right\} - \frac{\beta}{2}\sum_{n=1}^{N} ||y_n - \mathbf{W^T}\Phi(x_n)||^2 \tag{5}$$

The equation (3) is the required likelihood equation and the equation (5) is the required log-likelihood equation.
The mean squared error is given by,

$$E_D\{\mathbf{W}\} = \frac{1}{2}\sum_{n=1}^{N}\{y_n - \mathbf{W^T}\Phi(x_n)\}^2 \tag{6}$$

Writing equation (6) in vector notation,

$$E_D\{\mathbf{W}\} = \frac{1}{2}(\mathbf{\Phi W} - \mathbf{Y})^T(\mathbf{\Phi W} - \mathbf{Y}) \tag{7}$$

Taking gradient of equation (7) we get,

$$\nabla_{\mathbf{W}}(E_D\{\mathbf{W}\}) = \frac{1}{2}\nabla_{\mathbf{W}}(\mathbf{W^T\Phi^T\Phi W} - \mathbf{W^T\Phi^T y} - \mathbf{y^T\Phi W} + \mathbf{y^T y}) \tag{8}$$

$$= \frac{1}{2}(2\mathbf{\Phi^T\Phi W} - 2\mathbf{\Phi^T y}) \tag{9}$$

$$= \mathbf{\Phi^T\Phi W} - \mathbf{\Phi^T y} \tag{10}$$

For finding the maximum likelihood estimate of $\mathbf{W}$ we set the gradient to zero and solving for $\mathbf{W}$. Then from equation (10) we get,

$$\mathbf{\Phi^T\Phi W} - \mathbf{\Phi^T y} = 0 \tag{11}$$

$$\mathbf{W_{ML}} = (\mathbf{\Phi^T\Phi})^{-1}\mathbf{\Phi^T y} \tag{12}$$

Here the equation (12) is the maximum likelihood estimate (MLE) of the parameter vector $\mathbf{W}$.
Again if we consider the following error function

$$E(\mathbf{W}) = E_D(\mathbf{W}) + \frac{\lambda}{2}E_W(\mathbf{W}) \tag{13}$$

$$E(\mathbf{W}) = \frac{1}{2}\sum_{n=1}^{N}\{y_n - \mathbf{W^T}\Phi(x_n)\}^2 + \frac{\lambda}{2}||\mathbf{W}||^2 \tag{14}$$

Writing the equation (14) in vector form we get,

$$E(\mathbf{W}) = \frac{1}{2}(\mathbf{\Phi W} - \mathbf{Y})^T(\mathbf{\Phi W} - \mathbf{Y}) + \frac{\lambda}{2}\mathbf{W^T W} \tag{15}$$

Assuming prior on the parameter $\mathbf{W}$ to be a Gaussian with zero mean and $\alpha$ precision (inverse variance) we get the prior as follows,

$$p(\mathbf{W}|\alpha) = \mathcal{N}(W|0, \alpha^{-1}\mathbf{I}) \tag{16}$$

$$= \left\{\frac{\alpha}{2\pi}\right\}^{\frac{N}{2}} exp\left\{-\frac{\alpha}{2}\mathbf{W^T W}\right\} \tag{17}$$

Hence the posterior is given by.

$$p(\mathbf{W}|\mathbf{X}, y, \alpha, \beta) \propto p(y|\mathbf{X}, \mathbf{W}, \beta)p(\mathbf{W}|\alpha) \tag{18}$$

Hence taking the log of posterior and finding out the MAP estimate of the parameter $\mathbf{W}$ from equation (18) we get the MAP estimate as,

$$\text{MAP Estimate} = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{W}) - y_n\}^2 + \frac{\alpha}{2}\mathbf{W^T W} \tag{19}$$

Writing the MAP estimate of parameter $\mathbf{W}$ in vector notation we get,

$$\mathbf{W_{MAP}} = (\lambda\mathbf{I} + \mathbf{\Phi^T\Phi})^{-1}\mathbf{\Phi^T y} \tag{20}$$

Hence, equation (20) is the required maximum a-posteriori estimate (MAP estimate) of the parameters $\mathbf{W}$.

### 2.2 Section 2

Considering the hypothesis,

$$y(\mathbf{X}, \mathbf{W}) = \mathbf{W^T}\Phi(\mathbf{X}) \tag{1}$$

And, consider we have multiple independent outputs in we can write the equations as follows,

$$\begin{pmatrix} \hat{y_1} \\ \hat{y_2} \end{pmatrix} = \begin{pmatrix} \mathbf{\hat{w_1}^T} \\ \mathbf{\hat{w_2}^T} \end{pmatrix} \begin{pmatrix} \Phi_1(x) \\ \Phi_2(x) \end{pmatrix} \tag{2}$$

Now, putting the given data values we get,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{y_1} = \begin{pmatrix} -1 \\ -1 \\ -2 \\ 1 \\ 1 \\ 2 \end{pmatrix} \qquad \mathbf{y_2} = \begin{pmatrix} -1 \\ -2 \\ -1 \\ 1 \\ 2 \\ 1 \end{pmatrix} \tag{3}$$

Now we get the following,

$$(\mathbf{X^T X})^{-1}\mathbf{X^T} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \tag{4}$$

Hence using maximum likelihood estimate, the parameter $\mathbf{W}$ is given by,

$$\mathbf{\hat{w_1}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y_1} \tag{5}$$

$$\implies \mathbf{\hat{w_1}} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ -2 \\ 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{pmatrix} \tag{6}$$

And,

$$\mathbf{\hat{w_2}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y_2} \tag{7}$$

$$\implies \mathbf{\hat{w_2}} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ -1 \\ 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{pmatrix} \tag{8}$$

Hence from equations (6) and (8) we get the desired parameter matrix $\mathbf{W}$ as follows,

$$\mathbf{W} = \begin{pmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{pmatrix} \tag{9}$$

Hence from equation (9), we obtain the maximum likelihood estimates for parameter $\mathbf{W}$

## 3 Question 3

### 3.1 Dataset

From the given dataset, we split the dataset of 20 years into first 13 years for training and remaining 7 years for testing. We show plot of 2 such corps in Figure 1. The probability density function of a Poisson distribution is expressed as follows,

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1}$$

We now derive estimates of the training dataset for each corps as follows.

### 3.2 Maximum Likelihood Estimate of Parameter

The maximum likelihood estimate of poisson parameter $\lambda$ is expressed as follows,

$$\lambda_{ML} = \arg\max_{\lambda}\{p(D|\lambda)\} \tag{2}$$

We can write the likelihood function as follows,

$$p(D|\lambda) = \prod_{n=1}^{N} p(x_i|\lambda) \tag{3}$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x!} \tag{4}$$

To find the $\lambda$ for which likelihood is maximized, we will first take a logarithm of the likelihood and find its first derivative with respect to $\lambda$, and finally equate it with zero to find the $\lambda$ for which the likelihood is maximized. Hence the log-likelihood is given by,

$$l(D, \lambda) = \left(\sum_{i=1}^{n} x_i\right) ln\lambda - n\lambda - \sum_{i=1}^{n} ln(x!) \tag{5}$$

Taking derivative of the equation (5) with respect to $\lambda$ and equating it to zero we get,

$$\frac{\partial l(D, \lambda)}{\partial \lambda} = 0 \tag{6}$$

$$\implies \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \tag{7}$$

$$\lambda_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

Here equation (8) is the desired MLE of poisson distribution parameter $\lambda$.

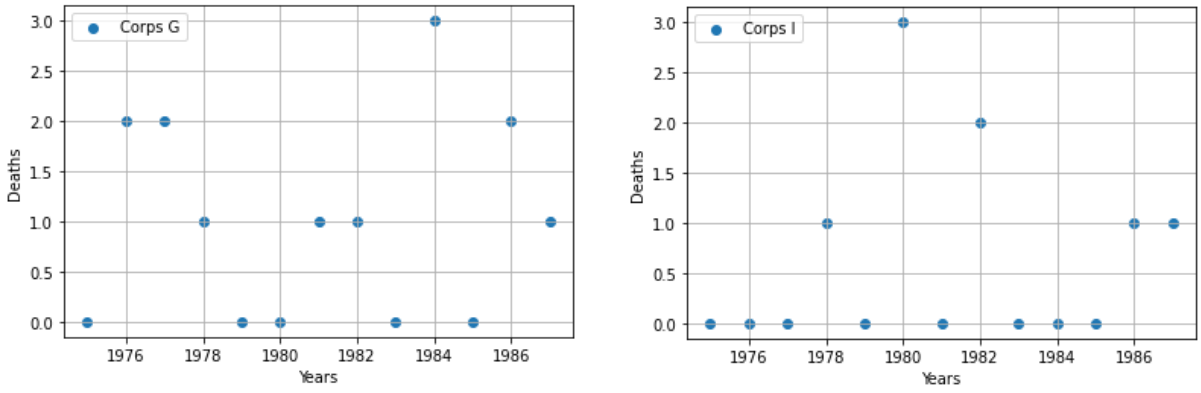


Figure 1: Training Dataset for Corps G and Corps I

### 3.3 Maximum A-posteriori Estimate

The maximum a-posteriori estimate of poisson parameter $\lambda$ is expressed as follows,

$$\lambda_{MAP} = \arg\max_{\lambda}\{p(D|\lambda)p(\lambda)\} \tag{9}$$

Where $p(\lambda)$ is the prior probability over the poisson parameter $\lambda$.

***Choice of Prior for Poisson Likelihood:*** As the Gamma distribution $gamma(k, \theta)$ is the conjugate prior of poisson likelihood and together they constitute a Gamma posterior we choose Gamma distribution $gamma(k, \theta)$ with shape parameter $k > 0$ and scale parameter $\theta > 0$ as our prior over the poisson parameter $\lambda$. Furthermore, Gamma distribution is a natural choice as the support of a Gamma random variable $x$ is $(0, \infty)$. Carefully looking at our training dataset we observe that the frequency of zero deaths is the highest followed by the frequency of one death. The frequency of two deaths is small and the frequency of three death is very small. Hence judging by the observation we choose our parameters $(k, \theta)$ to be $(1, 1.3)$. Hence, the prior $p(\lambda)$ is given by a $gamma(k, \theta)$ distribution as follows,

$$p(\lambda) = \frac{\lambda^{k-1}e^{-\frac{\lambda}{\theta}}}{\theta^k\Gamma(k)} \tag{10}$$

The prior with shape parameter value $k = 1$ and scale parameter value $\theta = 1.3$ is shown in the following Graph 3 where we see that there is high probability for $x = 0$ and decreasing probability of $x = 1$, $x = 2$, $x = 3$ and $x = 4$.

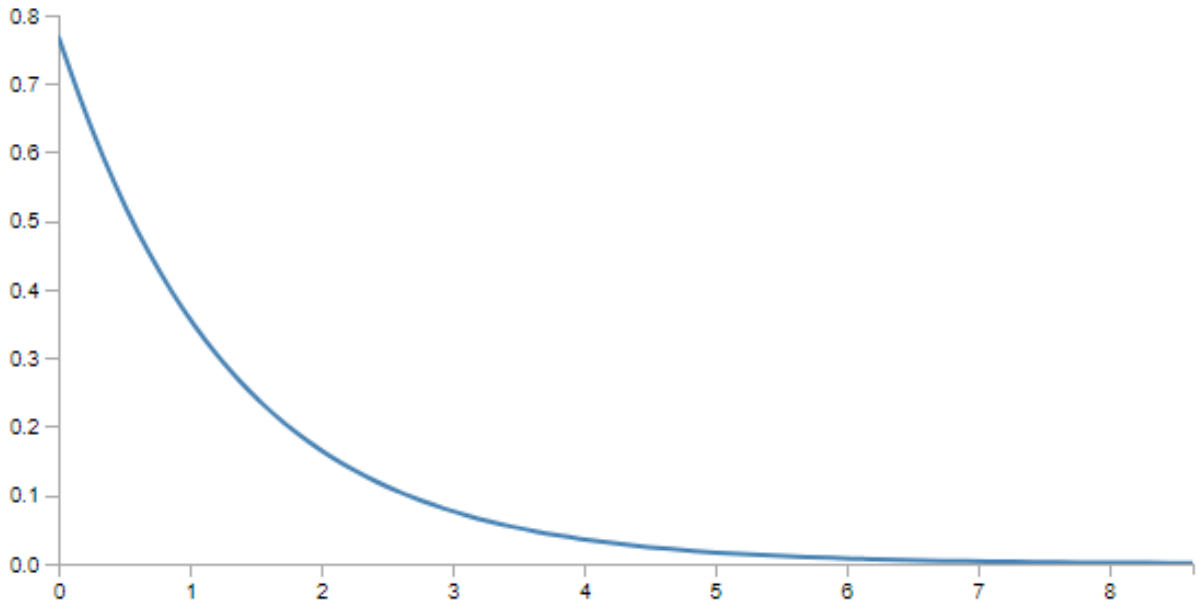


Figure 2: Gamma Prior with shape parameter $k = 1$ and scale parameter $\theta = 1.3$

Now, we can maximize the logarithm of the posterior distribution $p(\lambda|D)$ using the following,

$$lnp(\lambda|D) \propto lnp(D|\lambda) + lnp(\lambda) \tag{11}$$

$$= ln\lambda(k - 1\sum_{i=1}^{n} x_i) - \lambda\left(n + \frac{1}{\theta}\right) - \sum_{i=1}^{n} ln(x_i!) - kln\theta - ln\Gamma(k) \tag{12}$$

Taking derivative of the equation (12) with respect to $\lambda$ and equating it to zero we get,

$$\lambda_{MAP} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}} \tag{13}$$

| Corps | $\lambda_{MLE}$ | $\lambda_{MAP}$ |
|---|---|---|
| **G** | 1.00 | 0.9441 |
| **I** | 0.6153 | 0.5810 |
| **II** | 0.6153 | 0.5810 |
| **III** | 0.6153 | 0.5810 |
| **IV** | 0.4615 | 0.4357 |
| **V** | 0.3846 | 0.3631 |
| **VI** | 0.8461 | 0.7988 |
| **VII** | 0.5384 | 0.5083 |
| **VIII** | 0.3076 | 0.2905 |
| **IX** | 0.6923 | 0.6536 |
| **X** | 0.5384 | 0.5083 |
| **XI** | 1.00 | 0.9441 |
| **XIV** | 1.4615 | 1.3798 |
| **XV** | 0.3076 | 0.2905 |

Table 1: MLE and MAP values of the poisson parameter $\lambda$ for each corps

Here equation (13) is the desired MAP estimate of the poisson parameter $\lambda$.
In the Table 1 we show the estimated MLE and MAP values of poisson parameter $\lambda$ for each corps computed using the equations (8) and (13).
Below is the graph of MLE and MAP estimates for two of the corps.

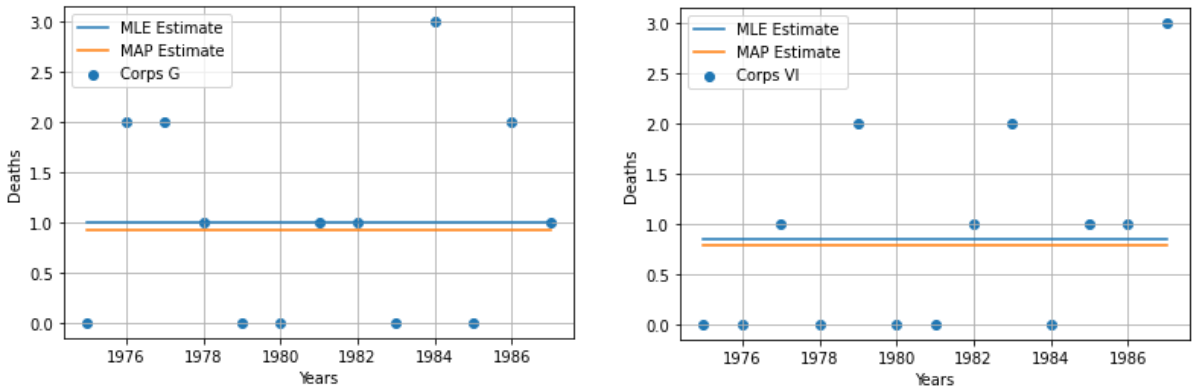


Figure 3: MLE and MAP values of Poisson parameter $\lambda$ for corps G and corps VI

### 3.4 Loss and Prediction

Using the maximum likelihood estimate ($\lambda_{MLE}$) and the maximum a-posteriori estimate ($\lambda_{MAP}$) of the poisson parameter we make predictions on the the testing data (i.e on the remaining 7 years of death counts in each corps) and compute the Root Mean Squared Error (RMSE) for each of the corps. The RMSE is defined as follows,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{14}$$

Where $y_i$ is the true death count for a specific corps in the $i$th year and $\hat{y}_i$ is the predicted death count for that specific corps in the $i$th year. Here $N$ is the total number of predictions or total numbers of years.
We show the RMSE for each of the corps where the predictions have been made using both the maximum likelihood estimate ($\lambda_{MLE}$) and the maximum a-posteriori estimate ($\lambda_{MAP}$) of the poisson parameter in the Figure 4.

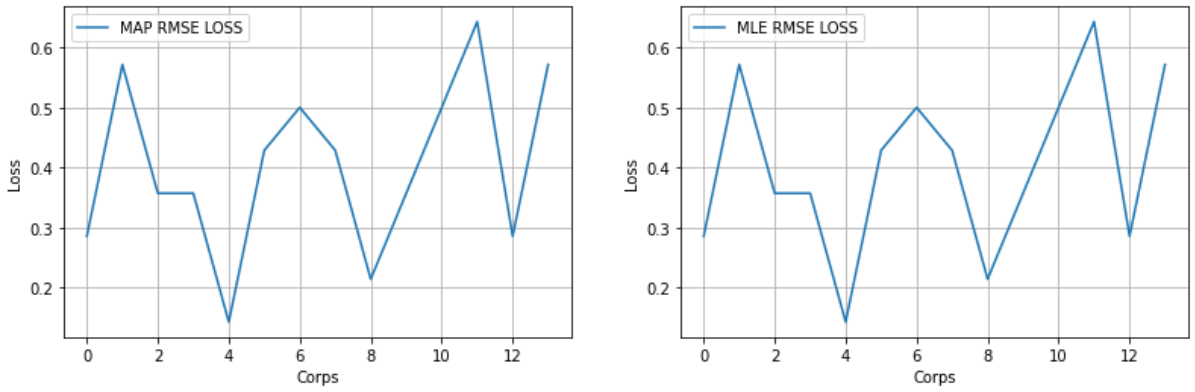


Figure 4: RMSE for predictions using MLE (Maximum Likelihood estimate) and MAE (Maximum A-posterior Estimate) of Poisson parameter $\lambda$

### 3.5 Likelihood, Prior and Posterior

We plot the prior calculated using equation (10) and likelihood calculated using equation (4). We also plot the posterior for each of the corps in the following Figure **??** which we obtain by multiplying the prior and the likelihood. Below are the plots of priors, likelihoods and posteriors of **Corps II**, **IV** and **VI**

#### 3.5.1 Corps II

The plots for **Corps II** are following,

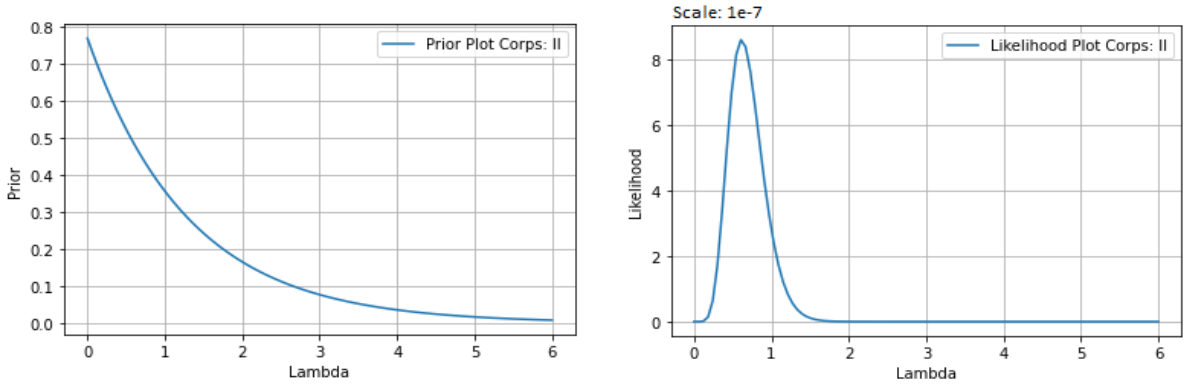


Figure 5: Plot of Prior and Likelihood for **Corps II**

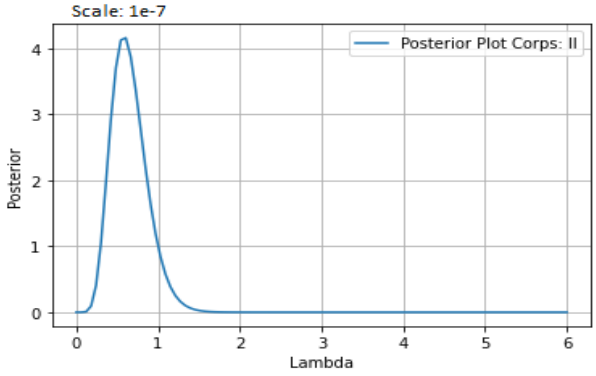


Figure 6: Plot of gamma posterior for **Corps II**

### 3.5.2 Corps IV

The plots for **Corps IV** are following,

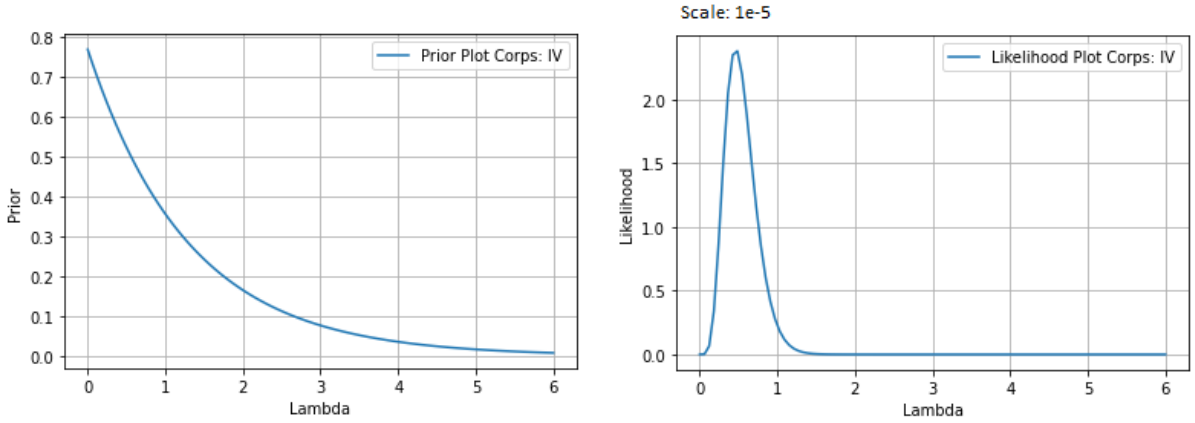


Figure 7: Plot of Prior and Likelihood for **Corps IV**

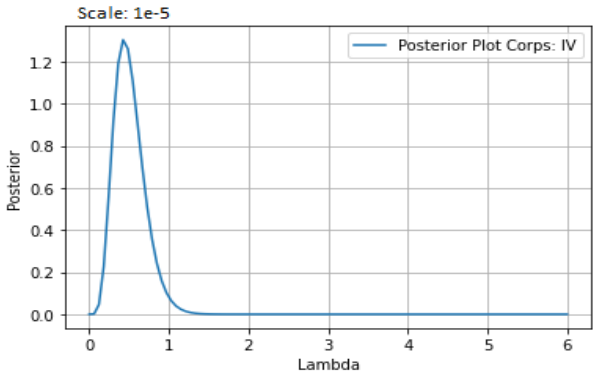


Figure 8: Plot of gamma posterior for **Corps IV**

### 3.5.3 Corps VI

The plots for **Corps VI** are following,

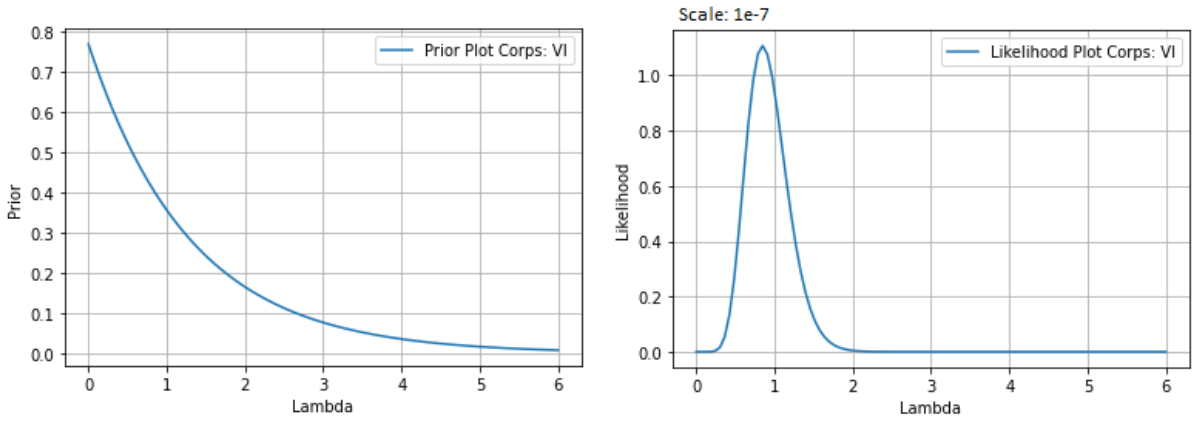


Figure 9: Plot of Prior and Likelihood for **Corps VI**

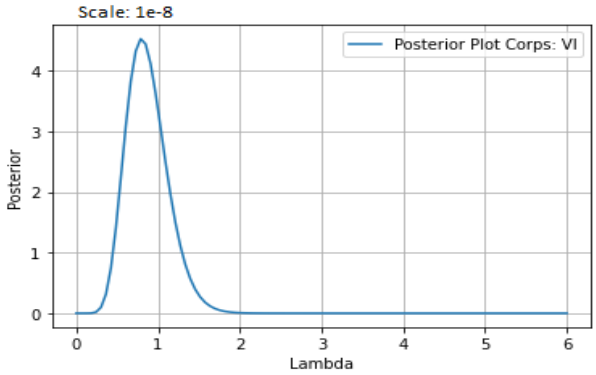


Figure 10: Plot of gamma posterior for **Corps VI**

### 3.6 Observation on Modes of Distribution

#### 3.6.1 Observations on Mode of Likelihood and Posterior

We briefly discuss the observations on mode of likelihood and posterior for **Corps II**, **Corps IV** and **Corps VI**.

- **Corps II :** From the Figure 5 and Figure 6 we observe that, mode of the prior is 0 which is as per assumption. The mode of the likelihood is is $\lambda_{MLE} = 0.6153$ and mode of posterior is $\lambda_{MAP} = 0.5810$
- **Corps IV :** From the Figure 7 and Figure 8 we observe that, mode of the prior is 0 which is as per assumption. The mode of the likelihood is is $\lambda_{MLE} = 0.4615$ and mode of posterior is $\lambda_{MAP} = 0.4357$
- **Corps VI :** From the Figure 9 and Figure 10 we observe that, mode of the prior is 0 which is as per assumption. The mode of the likelihood is is $\lambda_{MLE} = 0.8461$ and mode of posterior is $\lambda_{MAP} = 0.7988$
- All the modes of likelihoods and posteriors can be verified in the code which separately computes the mode of likelihood and posterior, and it turns out that model of likelihood is indeed $\lambda_{MLE}$ and mode of posterior is indeed $\lambda_{MAP}$

#### 3.6.2 Observations on Mode of Distribution of Deaths

We plot the mode for the complete dataset (13 years of training data and 7 years of testing data) as well as mode using the predicted dataset (13 years of training data and next 7 years of predicted testing data using MAP estimate) for the corps **II**, **IV** and **VI**.

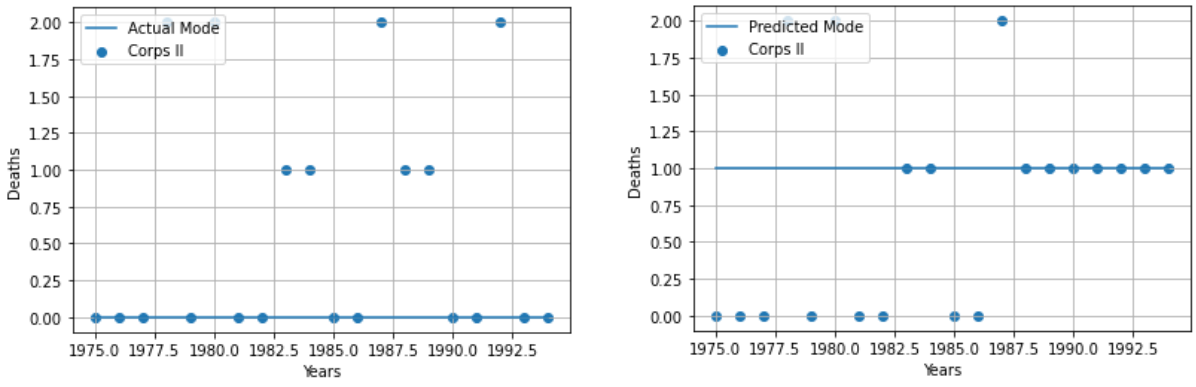


Figure 11: Observation on Mode of Corps **II** on Complete dataset

In the Figure 11 we observe that the actual mode for the given dataset of 20 years is 0 whereas, the mode for the training dataset (13 years) and the predicted dataset (next 7 years) has turned out to be 1.

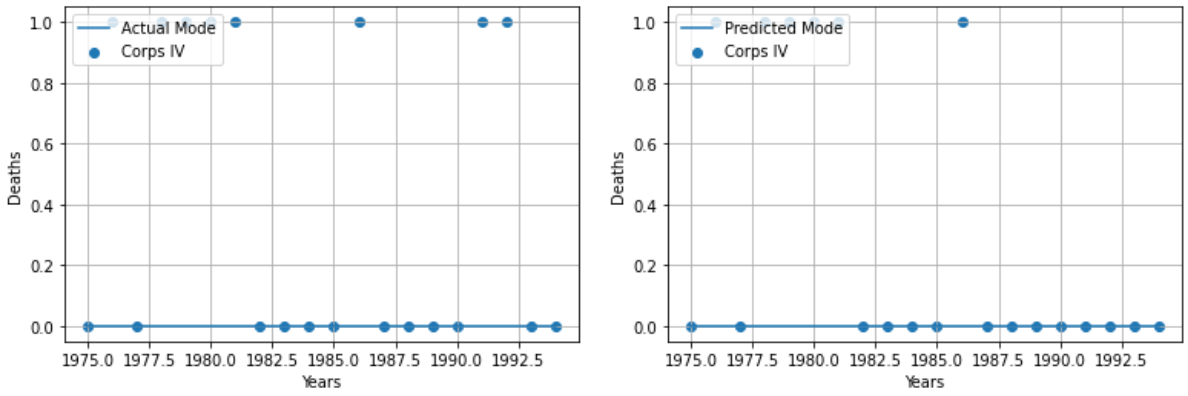


Figure 12: Observation on Mode of Corps **IV** on Complete dataset

In the Figure 12 we observe that the actual mode for the given dataset of 20 years is 0 and also the mode for the training dataset (13 years) and the predicted dataset (next 7 years) has turned out to be 0.

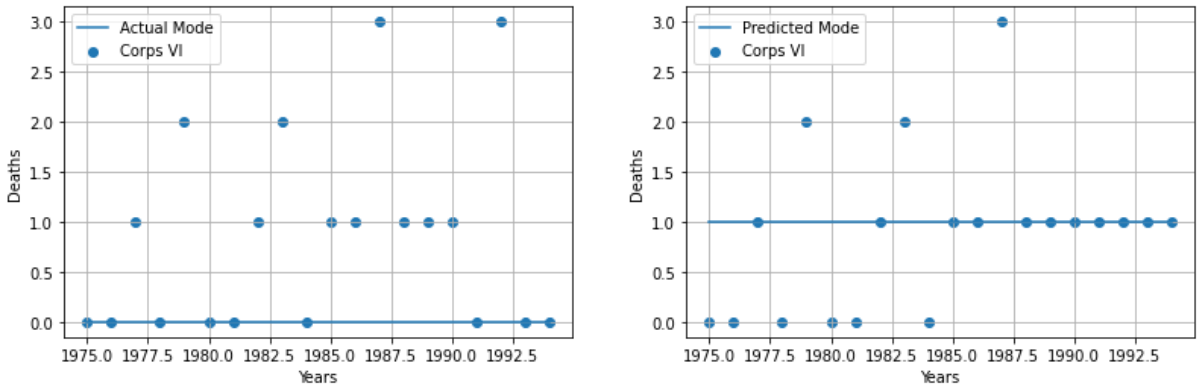


Figure 13: Observation on Mode of Corps **VI** on Complete dataset

In the Figure 13 we plot the mode of corps **VI**. We observe that the actual mode for the given dataset of 20 years is 0 whereas the mode for the training dataset (13 years) and the predicted dataset (next 7 years) has turned out to be 1.

### 3.7 Codes

Codes for the problem are attached in the zip file inside folder **Q3**. The folder contains the main python script and corresponding Readme file. A jupyter notebook and HTML notebook containing the results is also attached.

## 4 Question 4

In this question as we see in the below graphs that the response variable count follows Poisson distribution. The plot for the response variable following Poisson distribution is given below,
As clearly observed from Figure 14, the count response variable follows Poisson Distribution, hence use Poisson regression to predict counts. Below is discussion, observations and results of the Poisson Regression model used.

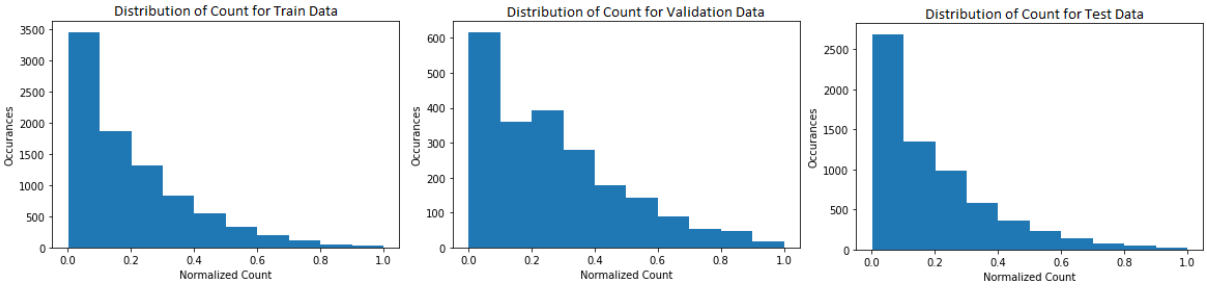


Figure 14: Plots of Normalized Count Data for Training, Validation and Testing Data following Poisson Distribution

### 4.1 Maximum Likelihood Estimate and Loss Function for Poisson Regression

Given a set of parameters $\theta$ and an input vector $x$, the mean of the predicted Poisson distribution is given by,

$$\lambda = E(Y|x) = e^{\theta^T x} \tag{1}$$

and the Poisson distribution's probability mass function is given by,

$$p(y|x;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{2}$$

$$= \frac{e^{y\theta^T x} e^{-e^{\theta^T x}}}{y!} \tag{3}$$

We are given a dataset consisting of $m$ input feature vectors $x_i \in R^{n+1}, i = 1, \ldots, m$ along with a set of $m$ response values $y_i \in R, i = 1, \ldots, m$. Then, for a given set of parameters $\theta$, the probability of attaining this particular set of data is given by

$$p(y_1, y_2, \ldots, y_m|x_1, x_2, \ldots, x_m; \theta) = \prod_{i=1}^{m} \frac{e^{y_i\theta^T x_i} e^{-e^{\theta^T x_i}}}{y_i!} \tag{4}$$

Rewriting equation (4) in the form of the likelihood function in terms of $\theta$ we get,

$$L(\theta|X;Y) = \prod_{i=1}^{m} \frac{e^{y_i\theta^T x_i} e^{-e^{\theta^T x_i}}}{y_i!} \tag{5}$$

Taking logarithm on both sides of the equation (5) we get the log-likelihood as follows,

$$l(\theta|X;Y) = \sum_{i=1}^{m} \{y_i\theta^T x_i - e^{\theta^T x_i} - log(y_i!)\} \tag{6}$$

In equation (6) the parameters $\theta$ only appears in the first two terms of each term in the summation. Therefore, given that we are only interested in finding the best value for $\theta$ we may drop the third term and rewrite the log-likelihood equation as follows,

$$l(\theta|X;Y) = \sum_{i=1}^{m} \{y_i\theta^T x_i - e^{\theta^T x_i}\} \tag{7}$$

Here the equation (7) is the log-likelihood function that we need to maximize in order to get the Maximum Likelihood Estimate (MLE) value of parameters $\theta$, which is equivalent to minimizing the negative

log-likelihood and finding the MLE value of $\theta$. Hence, the **loss function** $E(\theta)$ **for Poisson Regression** is the negative log-likelihood given by,

$$E(\theta) = -\sum_{i=1}^{m} \{y_i \theta^T x_i - e^{\theta^T x_i}\} \tag{8}$$

To find the MLE of parameter $\theta$ we need to differentiate the error function (8) with respect to $\theta$ and set it to zero i.e,

$$\frac{\partial(E(\theta))}{\partial\theta} = 0 \tag{9}$$

The equation (9) has no closed form solution for $\theta$ but the negative log-likelihood in equation (8) is a convex function and can be solved using standard convex optimization techniques such as gradient descent. For the $i$th coefficient of the parameter vector $\theta$ the gradient descent update rule is as follows,

$$\theta_i := \theta_i - \alpha(e^{\theta^T x_i} - y_i)x_i \tag{10}$$

Where $x_i$ is the $i$th feature of the feature vector $x$ and $\alpha$ is the learning rate.

### 4.2 Discussion on Statistics of Dataset

We plot the mean values of response variable count against **Holiday and Working day**, **Hour**, **Month** in the following graphs.
We also plot the mean values of response variable count against **Days of the week** and **Year**.
Each graph contains the **mean values of the response variable count against corresponding features**.

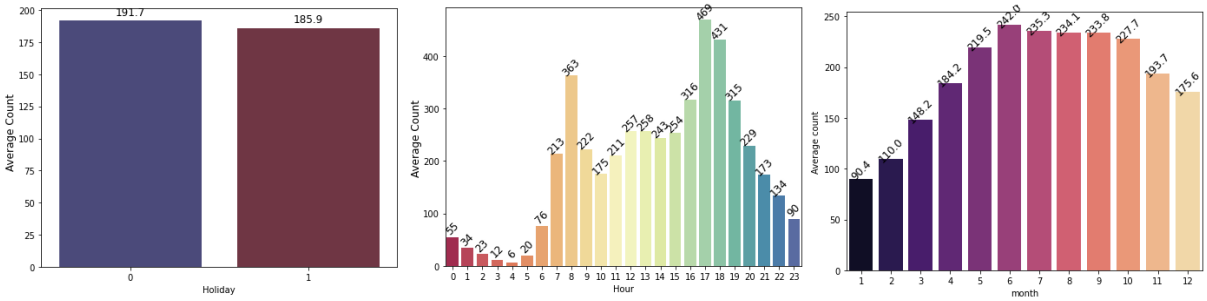


Figure 15: Plots of Mean Count Data against Holiday and Working day, Hours, and Months

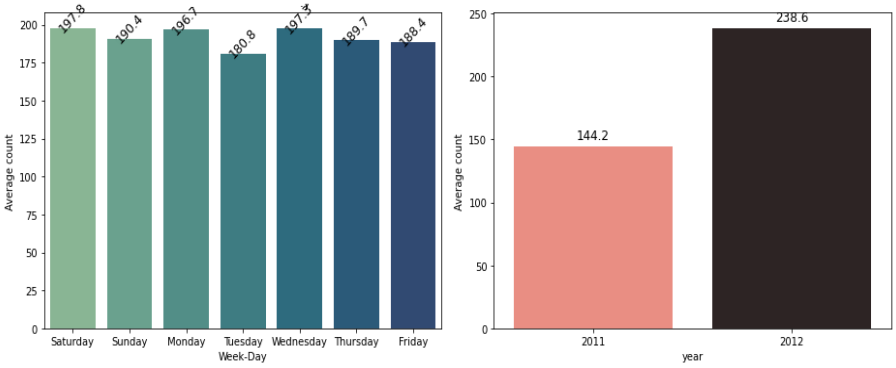


Figure 16: Plots of Mean Count Data against Days of Week and Year

We plot the median values of response variable count against **Holiday and Working day**, **Hour**, **Month**

in the following graphs.
We also plot the median values of response variable count against **Days of the week** and **Year**.
Each graph contains the **median values of the response variable count against corresponding features**.

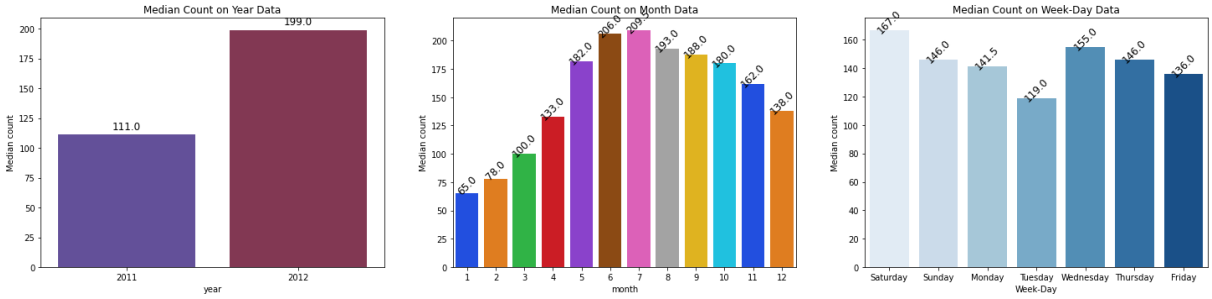


Figure 17: Plots of Median Count Data against Year, Months and Days of Week

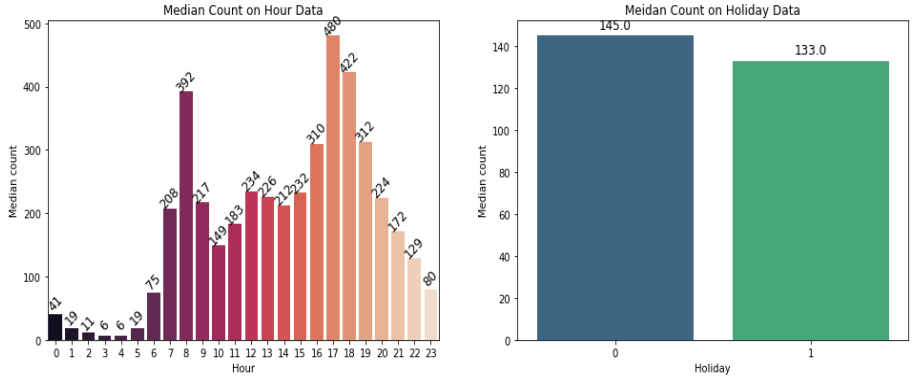


Figure 18: Plots of Median Count Data against Hour and Weekday and Holiday

Below we tabulate the average count statistics and median count statistics with respect to different features as seen from the plots in Figure 15, 16, 17 and 18.

| Hour (HRS) | Mean Count | Median Count | Hour (HRS) | Mean Count | Median Count |
|---|---|---|---|---|---|
| **0 HRS** | 55 | 41 | **12 HRS** | 257 | 234 |
| **1 HRS** | 34 | 19 | **13 HRS** | 258 | 226 |
| **2 HRS** | 23 | 11 | **14 HRS** | 243 | 212 |
| **3 HRS** | 12 | 6 | **15 HRS** | 254 | 232 |
| **4 HRS** | 6 | 6 | **16 HRS** | 316 | 310 |
| **5 HRS** | 20 | 19 | **17 HRS** | **469** | **480** |
| **6 HRS** | 76 | 75 | **18 HRS** | 431 | 422 |
| **7 HRS** | 213 | 208 | **19 HRS** | 315 | 312 |
| **8 HRS** | 363 | 392 | **20 HRS** | 229 | 224 |
| **9 HRS** | 222 | 217 | **21 HRS** | 173 | 172 |
| **10 HRS** | 175 | 149 | **22 HRS** | 134 | 129 |
| **11 HRS** | 211 | 183 | **23 HRS** | 90 | 80 |

Table 2: Hour-wise Average Count

| Days | Mean Count | Median Count |
|---|---|---|
| **Monday** | **197.8** | **167** |
| **Tuesday** | 190.4 | 146 |
| **Wednesday** | 196.7 | 141.5 |
| **Thursday** | 180.8 | 119 |
| **Friday** | 197.3 | 155 |
| **Saturday** | 189.7 | 146 |
| **Sunday** | 188.4 | 136 |
| **Working Day** | **191.7** | **145** |
| **Holiday** | 185.9 | 133 |
| **Year** | **Mean Count** | **Median Count** |
| **2011** | 144.2 | 111 |
| **2012** | **238.6** | **199** |

Table 3: Day and Year-wise Average Count

| Month | Mean Count | Median Count |
|---|---|---|
| **January** | 90.4 | 65 |
| **February** | 110.0 | 78 |
| **March** | 148.2 | 100 |
| **April** | 184.2 | 133 |
| **May** | 219.5 | 182 |
| **June** | **242.0** | 206 |
| **July** | 235.3 | **209.5** |
| **August** | 234.1 | 193 |
| **September** | 233.8 | 188 |
| **October** | 227.7 | 180 |
| **November** | 193.7 | 162 |
| **December** | 175.6 | 138 |

Table 4: Month-wise Average Count

### 4.3 Plots of Count against Features

We plot the response variable Count against six features i.e **Humidity (Relative Humidity)**, **atemp ("feels like" temperature in Celsius)**, **Wind-speed**, **Average Count by Hour across Season**, **Average Count by Hour across Days** and **Average Count by Hour across Different Weathers** as follows,

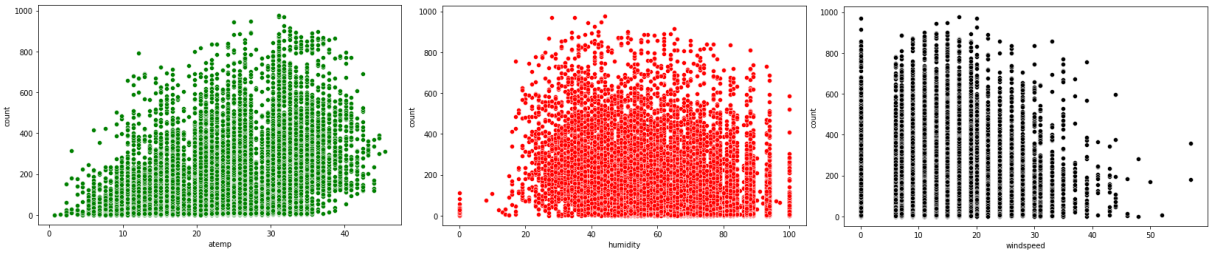


Figure 19: Plots of Count Data against atemp, Humidity and Wind-speed

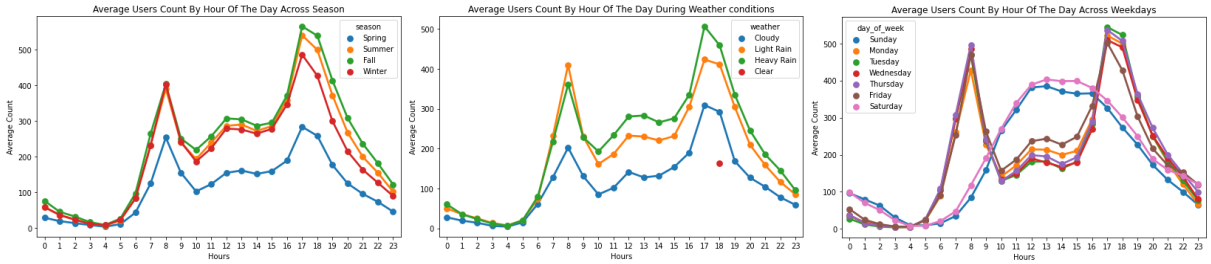


Figure 20: Plots of Count Data against Hours across Seasons, Days and Weather

From the Figure 19 and 20 we point out some observations as follows,

- We see that behaviour of response variable **Count** shows similar behaviour in the seasons of Summer, Fall and Winter whereas it shows somewhat different behaviour in the Spring season.
- We see from the graph that in the Light Rain, Heavy Rain and Clear weather the behaviour of the **Count** variable is similar. Again in cloudy weather the behaviour is different.
- As expected, the behaviour of **Count** variable is similar on weekdays and different on weekends.

- Across all the graphs, we see that there are two regions of hours when most of the count variable value occurs viz. around 7:00 HRS to 9:00 HRS and around 17:00 HRS to 20:00 HRS.

### 4.4 Summary of Hyper-parameter Optimization on Validation Data for L1, L2 and No Regularization

At first, we remove correlated features from the dataset, we visualize the correlation heat-map and decide which features to exclude. The correlation heat-map of features before and after removal of correlated features are shown in the below figure.

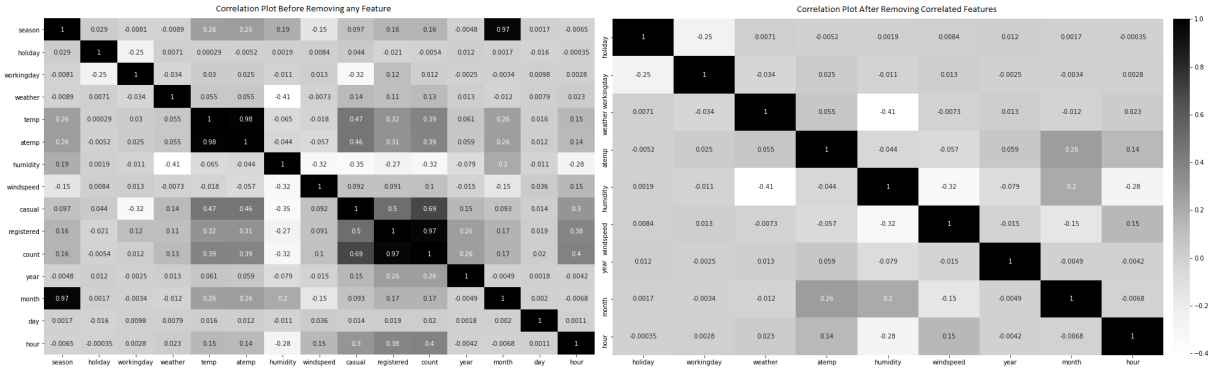


Figure 21: Correlation Heat-Map of Input Features before and after removal of correlated features

After that we train the model using 80% of the training data as training data and 20% of the remaining training data as validation data which we use for hyper-parameter tuning. We use several settings of hyper-parameters and observe validation data loss in terms of negative log-likelihood and choose the best hyper-parameters which gives the lowest validation loss. We find out the best setting of hyper-parameters for Un-regularized, L1 regularized and L2 regularized setting. The set of all hyper-parameters and corresponding validation loss that we have experimented is given in **two separate excel files** for L1 and L2 regularization each. In the below table we report the best hyper-parameter settings for the problem.

| Regularization Mode | Hyper-parameters | Values |
|---|---|---|
| **Un-regularized** | Weight Initialization | Zero Initialization |
| | Learning Rate | 0.0001 |
| | Epochs | 200 |
| | Regularization Constant | 0.0001 |
| **L1 Regularized** | Weight Initialization | Random Initialization |
| | Learning Rate | 0.0001 |
| | Epochs | 700 |
| | Regularization Constant | 0.0001 |
| **L2 Regularized** | Weight Initialization | Random Initialization |
| | Learning Rate | 0.0001 |
| | Epochs | 700 |
| | Regularization Constant | 0.0001 |

Table 5: Best Hyper-parameter values

### 4.5 Report on Test Data using L1, L2 and No Regularization

We train the model on training data and test on the validation data and finally test on the test data under all the settings (i.e Un-regularized, L1 regularized, L2 regularized). While training we use the hyper-parameters from the Table 5. Below we show the training loss curve for all the settings.
We report the negative-likelihood loss, Root Mean Squared Logarithmic Error (RMSLE) and Root

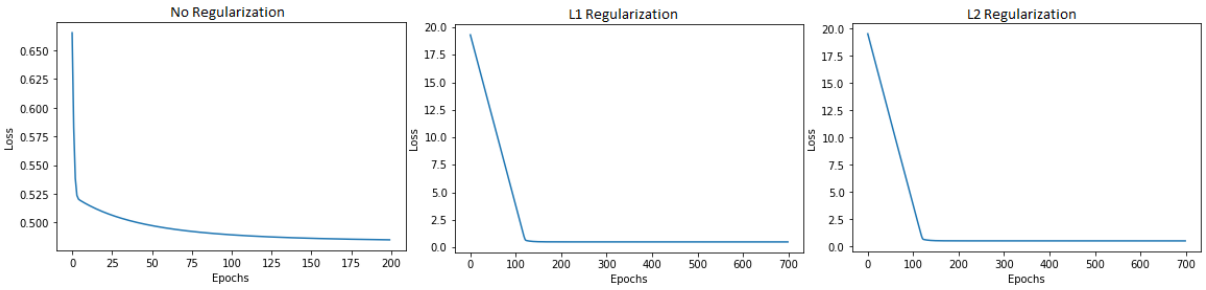


Figure 22: Training Loss for Un-regularized, L1 regularized and L2 regularized experiment

Mean Squared Error (RMSE) for validation data and test data for all the experimental setups (i.e Un-regularized, L1 regularized, L2 regularized). The **negative log-likelihood** is defined by the following equation,

$$E(\theta) = -\sum_{i=1}^{m} \{y_i \theta^T x_i - e^{\theta^T x_i}\} \tag{11}$$

The **Root Mean Squared Logarithmic Error (RMSLE)** is defined as follows,

$$RMSLE(\theta) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[log(\hat{y}_i + 1) - log(y_i + 1)]^2} \tag{12}$$

$$= \sqrt{\frac{1}{m}\sum_{i=1}^{m}[log(e^{\theta^T x_i} + 1) - log(y_i + 1)]^2} \tag{13}$$

The **Root Mean Squared Error (RMSE)** is defined as follows,

$$RMSE(\theta) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[\hat{y}_i - y_i]^2} \tag{14}$$

$$= \sqrt{\frac{1}{m}\sum_{i=1}^{m}[e^{\theta^T x_i} - y_i]^2} \tag{15}$$

In the table below we report the validation and test loss for all the experiment settings.

| Mode | Dataset | Negative Log Likelihood | Normalized RMSLE | Normalized RMSE | RMSLE | RMSE |
|---|---|---|---|---|---|---|
| **Un-regularized** | Validation | 0.5957 | 0.1433 | 0.1980 | 1.1182 | 201.2788 |
| | Test | 0.4781 | 0.1191 | 0.1564 | 1.2377 | 155.0540 |
| **L1 Regularized** | Validation | 0.5976 | 0.1438 | 0.1981 | 1.0958 | 201.4554 |
| | Test | 0.4781 | 0.1191 | 0.1564 | 1.2377 | 155.0540 |
| **L2 Regularized** | Validation | 0.5976 | 0.1438 | 0.1981 | 1.0958 | 201.4554 |
| | Test | 0.4763 | 0.1175 | 0.1551 | 1.1816 | 153.8378 |

Table 6: Summary of Validation and Test Loss for Poisson Regression (Normalized indicates loss computed on normalized counts)

### 4.6 Determining Most Important Features

After training the Poisson Regression model on best hyper-parameters for L1, L2 and No Regularization setting, we use the absolute values of the final weights which act as coefficients of the corresponding features and plot the absolute values for each of the Regularization setting and find the most important features. The plot of absolute weights for each feature in L1, L2 and No Regularization setting.

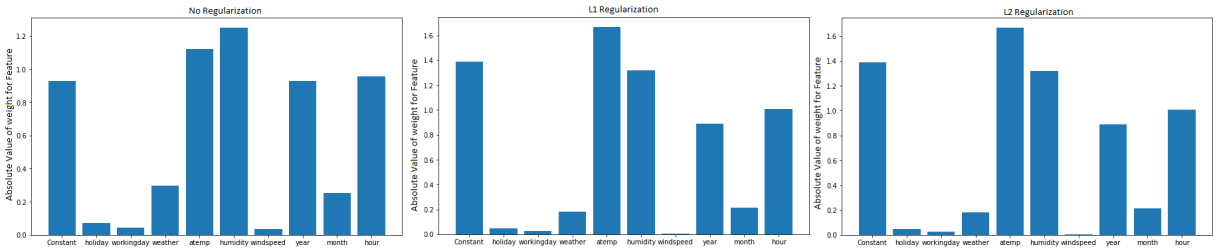


Figure 23: Plots of Absolute values of final weights against corresponding features

From Figure 23 we determine the most important features contributing to the response variable of Count are as follows,

- atemp (”Feels Like” Temperature in Celsius)
- Humidity (relative humidity)
- Hour
- Year

### 4.7 Codes

Codes for the problem are attached in the zip file inside folder **Q4**. The folder contains the main python script and corresponding Readme file. A jupyter notebook and HTML notebook containing the results is also attached. It also contains the excel files containing results on different hyper-parameters for L1 and L2 regularization. It is recommended to run the Jupyter Notebook in Google Colaboratory.