CS5590: FOUNDATIONS OF MACHINE LEARNING, Fall 2020

ASSIGNMENT 1

Team Members:

Sai Harsha Kottapalli

CS17BTECH11036

Surya Sai Teja Desu

CS17BTECH11048

1) After addition of gaussian noise ϵ_i to each input

we get, $y'(x,\omega) = \omega_0 + \sum_{i=1}^{D} \omega_i (x_i + \varepsilon_i)$

$$y'(x,\omega) = \omega_0 + \sum_{i=1}^{D} \omega_i x_i + \sum_{i=1}^{D} \omega_i \varepsilon_i$$

$$y'(x,\omega) = y(x,\omega) + \sum_{i=1}^{D} \omega_i \varepsilon_i \quad \ldots (1)$$

Given,

$$E_D(\omega) = \frac{1}{2} \sum_{n=1}^{N} \left\{ y_n(x_n,\omega) - t_n \right\}^2 \ldots (2)$$; sum of squares error function

By applying on new model,

$$\Rightarrow E_D^1(\omega) = \frac{1}{2} \sum_{n=1}^{N} \left\{ y_n'(x_n,\omega) - t_n \right\}^2$$

$$= \frac{1}{2} \sum_{n=1}^{N} \left\{ y(x_n,\omega) + \sum_{i=1}^{D} \omega_i \varepsilon_i - t_n \right\}^2 \boxed{\text{using (1)}}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \left\{ \left((y(x_n,\omega)) - t_n \right)^2 + \left(\sum_{i=1}^{D} \omega_i \varepsilon_i \right)^2 + 2\left(y(x_n,\omega) - t_n\right)\left(\sum_{i=1}^{D} \omega_i \varepsilon_i \right) \right\}$$

$$E_D'(w) = E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\left\{\left(\sum_{i=1}^{D} w_i \varepsilon_i\right)^2 + 2\left(y(x_n,w)-t_n\right)\left(\sum_{i=1}^{D} w_i \varepsilon_i\right)\right\}$$

Applying expectation & linearity of expectation $\left(E[X+Y] = E[X] + E[Y]\right)$

$$E\left[E_D'(w)\right] = E\left[E_D(w)\right] + \frac{1}{2}\sum_{n=1}^{N}\left\{E\left[\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)^2\right] + 2\left(y(x_n,w-t_n)\right) \times \sum_{i=1}^{D} w_i E\left[\varepsilon_i\right]\right\}$$

But, $E\left[\varepsilon_i\right] = 0$

$$\Rightarrow E\left[E_D'(w)\right] = E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\left\{E\left[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \varepsilon_i \varepsilon_j\right] + 0\right\}$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\left\{\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j E\left[\varepsilon_i \varepsilon_j\right]\right\}$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\left\{\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij} \sigma^2\right\}$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\left\{\sum_{i=1}^{D} w_i^2 \sigma^2\right\}$$

$$= E_D(w) + \frac{N}{2}\sum_{i=1}^{D} w_i^2 \sigma^2$$

As we can see, $E\left[E_D'(w)\right]$ is independent of bias parameter w_0.

2.

(2) 1. Given $y = w^T \phi(x)$.

Considering gaussian distribution for each dimension of y.

let $y_i = [y_i^{(1)}\ y_i^{(2)} \dots y_i^{(k)}]$ $1 \le i \le N$.

$\phi(x) = [a_{ij}]_{M \times N}$ where a_{ij} = j^{th} dimension value of x_j.

let $y'_p = [y_1^{(p)}\ y_2^{(p)} \dots y_N^{(p)}]^T$

$\Rightarrow y'_p | x \sim \mathcal{N}(0, \sigma^2) + \phi(x)^T w_p$

$$L = \exp\left(-\frac{1}{2\sigma^2}\left(\phi(x)^T w_p - y'_p\right)^2\right)$$

$$\log L = \frac{-1}{2\sigma^2}\left(\phi(x)^T w_p - y'_p\right)^2$$

find optimal value for w_i

$$\Rightarrow \frac{\partial \log(L)}{\partial w_i} = 0$$

$\Rightarrow$

$$\phi(x) y'_p - \phi(x)\phi(x)^T w_p = 0.$$

$$\Rightarrow w_p = (\phi(x)\phi(x)^T)^{-1}\phi(x) y'_p$$

$$\Rightarrow W = (\phi(x)\phi(x)^T)^{-1}\phi(x) y'$$

$\therefore$ MLE of w

$$w = (\phi(x)\,\phi(x)^T)^{-1}\,\phi(x)\,y.$$

MAP estimate:

Assuming a gaussian prior for w_i with parameter as λ.

$$p(w_i/\lambda) = \left(\frac{\lambda}{2\pi}\right)^{M/2} \exp\left(-\frac{\lambda}{2} w_i^T w\right)$$

from Baye's rule, we have

$$\underbrace{p(W/x,y,\lambda)}_{\text{posterior}} \propto \underbrace{p(x,y/w)}_{L} \cdot \underbrace{p(W/\lambda)}_{\text{prior}}$$

$$\Rightarrow p(W/x,y,\lambda) \propto p(x,y/w) \cdot \prod_{i=1}^{K} p(w_i/\lambda)$$

$$\propto \exp\left(\frac{-1}{2}\sum_{i=1}^{N}\sum_{j=1}^{k}\frac{y_i^{(j)} - w_j^T\phi(x_i)}{\sigma_j^2}\right)\exp\left(-\sum_{j=1}^{k}\frac{\lambda}{2}w_j^Tw_j\right)$$

$$\Rightarrow \log\left(P(w/x,y,\lambda)\right) = \frac{-1}{2}\sum_{i=1}^{N}\sum_{j=1}^{k}\frac{y_i^{(j)} - w_j^T\phi(x_i)}{\sigma_j^2}$$

$$- \sum_{j=1}^{k}\frac{\lambda}{2}w_j^Tw_j + \text{Constant}$$

maximizing log-MAP to find the estimate for w_j.

$$\frac{\partial \log\left(P(w/x,y,\lambda)\right)}{\partial w_j} = 0$$

$$\Rightarrow \quad -\phi(x)y_j' + \phi(x)\phi(x)^Tw_j + \lambda w_j = 0.$$

$$\Rightarrow \quad w_j = \left(\phi(x)\phi(x)^T \cancel{\phi} + \lambda I\right)^{-1}\phi(x)y_j'$$

MAP estimate of w

$$\Rightarrow \quad W = \left(\phi(x)\phi(x)^T + \lambda I\right)^{-1}\phi(x)y.$$

which is same as solution for ridge regression.

Q4 Considering $\phi(0) = (1,0)^T$, $\phi(1) = (0,1)^T$

$$\phi(x) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \qquad y = w^T \phi(x)$$

$$y = \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$$

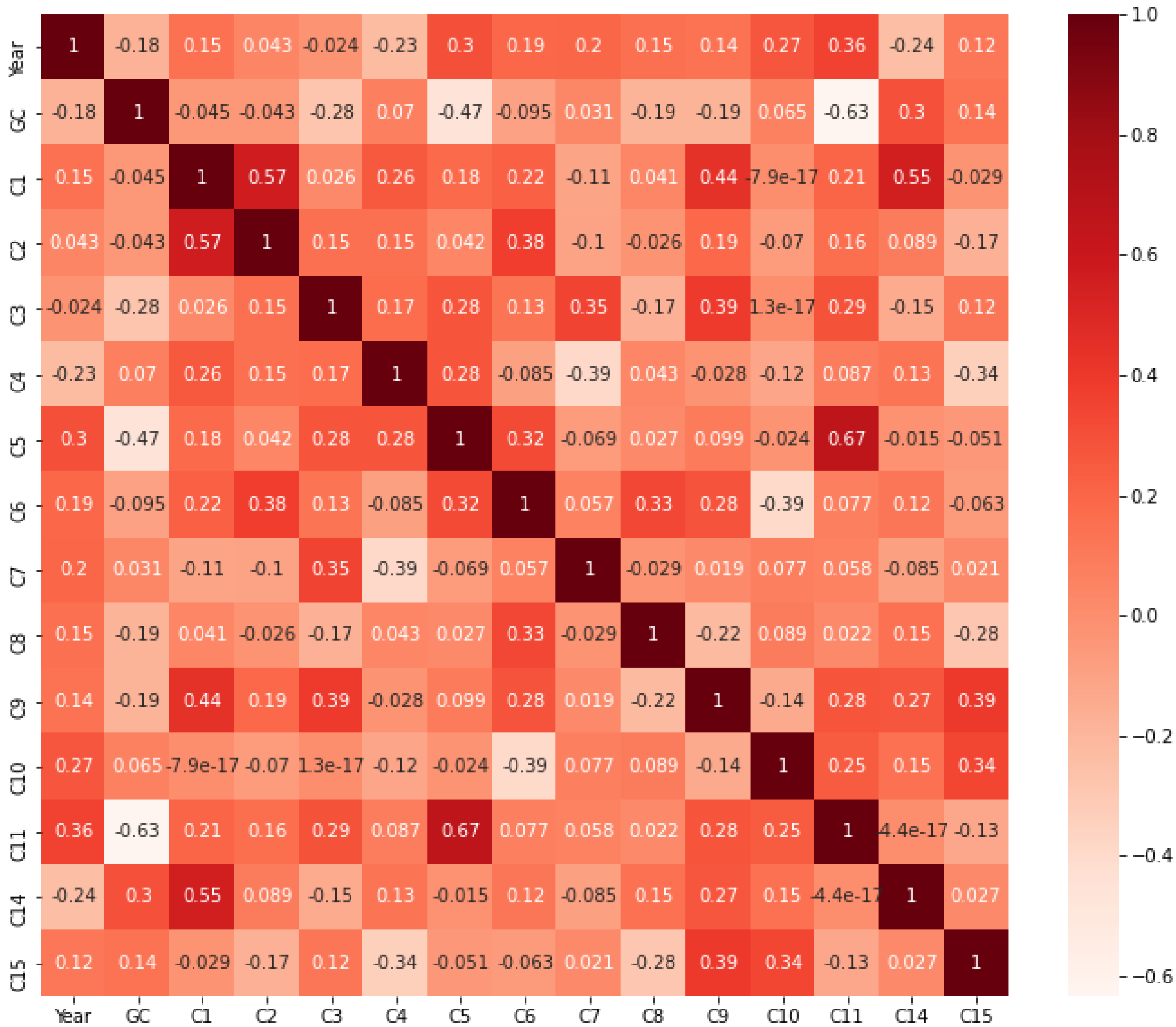
$$\text{MLE of } w = (\phi(x)\phi^T(x))^{-1}\phi(x)y$$

$$= \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} -4 & -4 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}.$$

$$\therefore \quad w = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}.$$

3. Since, the number of deaths due to horse kick is not correlated with year number and the input has no other features, we will take expectation of poisson distribution as the predicted value for any input.



For proving that the feature year is not correlated we will first plot the Pearson correlation heatmap and see there isn't much correlation of corps with year. Hence, we use expectation as the predicted value.

For maximum likelihood estimation:

Given observations $y_1, y_2, \dots y_n$ for input $x_1, x_2 \dots x_n$.

$L(\theta) = f(x_1, x_2, \dots, x_n | \theta)$ if θ is true val of param. the probability that we observe $x_1, x_2 \dots x_n$.

for MLE, we maximize $L(\theta)$.

As our data points are iid,

maximize $L(\theta) = f(x_1|\theta) . f(x_2|\theta) \dots f(x_n|\theta)$

For likelihood on poisson distribution,

$$P(Y|\lambda) = f(y_1|\lambda) . f(y_2|\lambda) \dots f(y_n|\lambda)$$

$$\Rightarrow P(Y|\lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \; ; \left(\because f(y_i|\lambda) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right)$$

& $Y = y_1, y_2, \dots y_n$.

$$\lambda_{ML} = \arg\max_{\lambda} \log(P(Y|\lambda))$$

$$\Rightarrow \frac{d}{d\lambda}\left(-n\lambda + \sum_{i=1}^{n} y_i \log\lambda + (-1)\log\left(\prod_{i=1}^{n} y_i!\right)\right) = 0$$

$$\Rightarrow \lambda_{ML} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Refer **Table 1** below for the poisson parameters (ML) and rmse values.

For maximum aposteriori estimation:

The gamma distribution is the conjugate prior for the likelihood function - poisson. Hence, we choose gamma distribution for prior distribution. It has two hyper parameters: (alpha. beta) which can be found via grid search over training set.

gamma distribution with parameters $\rightarrow (\alpha, \beta)$:

$$P(\lambda | \alpha, \beta) = \frac{\beta^{\alpha} \lambda^{\alpha - 1} e^{-\beta \lambda}}{\Gamma(\alpha)}$$

$$P(\lambda | y, \alpha, \beta) \propto P(y | \lambda) \cdot P(\lambda | \alpha, \beta)$$

$$\propto \frac{e^{-(\beta + n)\lambda} \, \lambda^{\left(\sum_{i=1}^{n} y_i + \alpha - 1\right)}}{\Gamma(\alpha) \prod_{i=1}^{n} y_i !}$$

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}} \log P(\lambda | y, \alpha, \beta)$$

$$\Rightarrow \frac{d}{d\lambda}\left(-(\beta + n)\lambda + \left(\sum_{i=1}^{n} y_i + \alpha - 1\right) \log \lambda - \log\left(\Gamma(\alpha) \prod_{i=1}^{n} y_i !\right)\right) = 0$$

$$\Rightarrow -(\beta + n) + \frac{\sum_{i=1}^{n} y_i + \alpha - 1}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{n} y_i + \alpha - 1}{n + \beta}$$

After doing a grid search over possible values of (alpha, beta) from {1..10} and taking Σ(rmse) as the cost metric over all corps on the training set, we find that alpha = 2 and beta = 1 is the best fit for it.

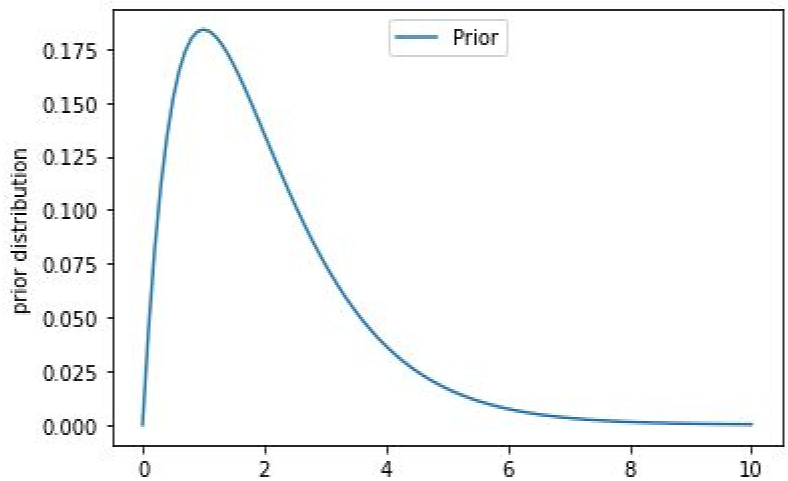
Finally **Table 1** below shows the MAP parameters and rmse values.

Corp	ML		MAP	
	λ	RMSE	λ	RMSE
G	1	0.755929	1	0.755929
I	0.692308	1.11244	0.714286	1.10657
II	0.615385	0.729756	0.642857	0.731925
III	0.615385	0.729756	0.642857	0.731925
IV	0.461538	0.484764	0.5	0.5
V	0.384615	0.587989	0.428571	0.553283
VI	0.846154	0.989804	0.857143	0.989743
VII	0.538462	0.898011	0.571429	0.892143
VIII	0.307692	0.509421	0.357143	0.5
IX	0.692308	0.738393	0.714286	0.742307
X	0.538462	1.15969	0.571429	1.14286
XI	1	1.13389	1	1.13389
XIV	1.46154	1.02381	1.42857	1
XV	0.307692	0.941214	0.357143	0.928571

Table 1

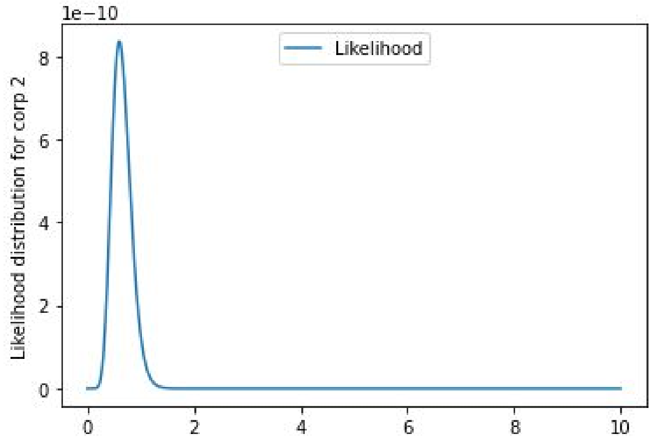
Graphs

Prior

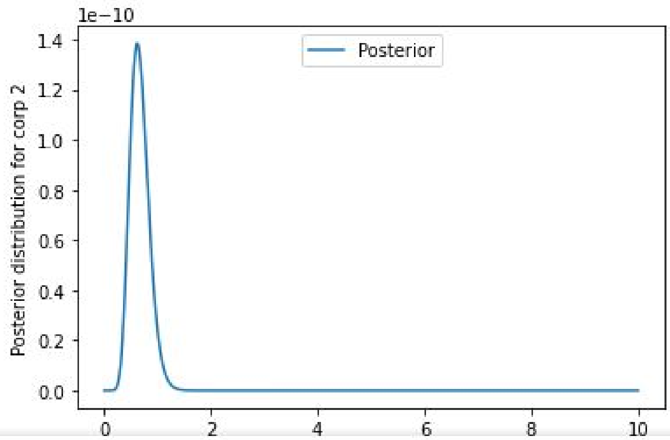



```
(Prior)Maxima at:  1.0
```

Corp 2

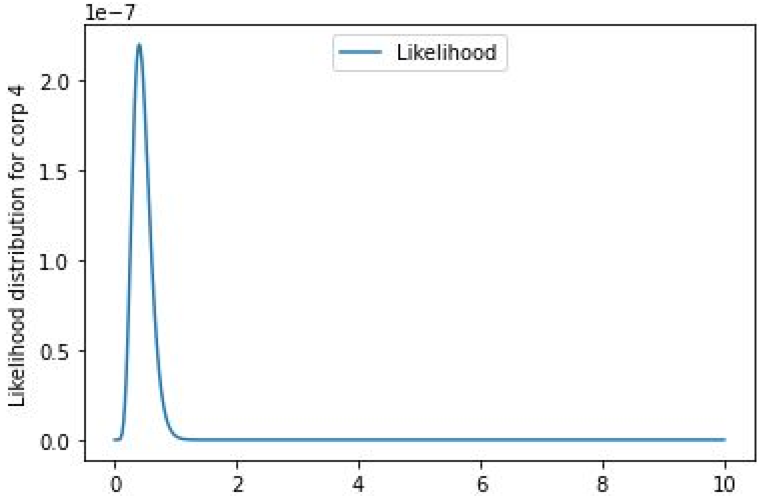



```
(Likelihood)Maxima at:  0.6
```

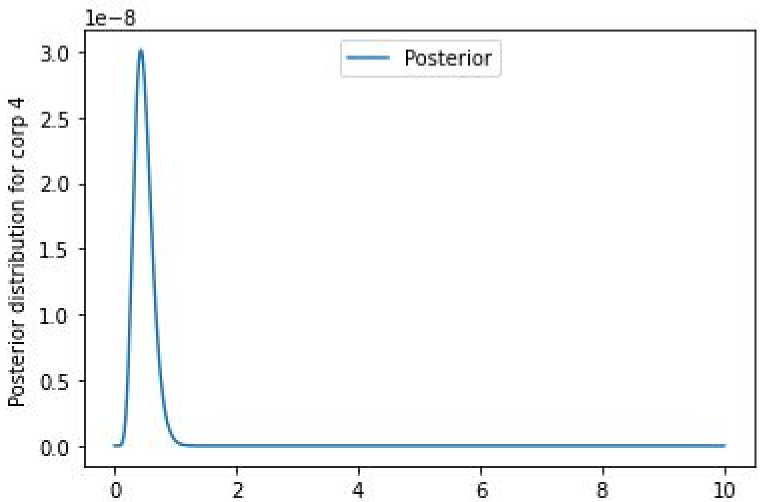



```
(Posterior)Maxima at:  0.62
```

Corp 4

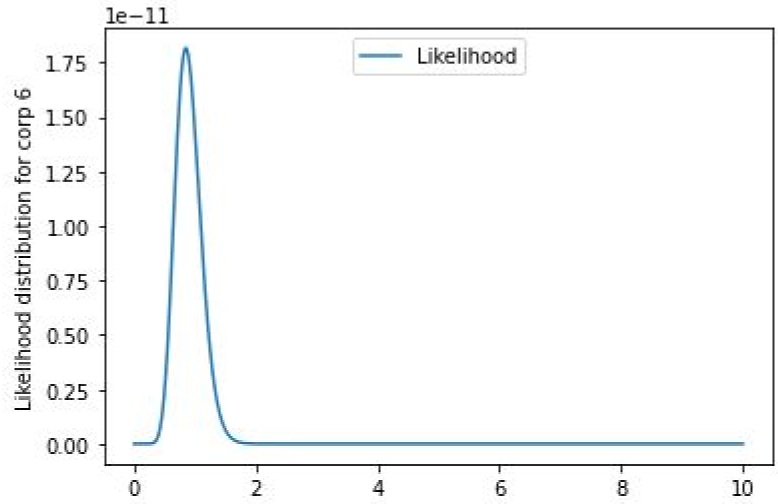



```
(Likelihood)Maxima at:  0.4
```

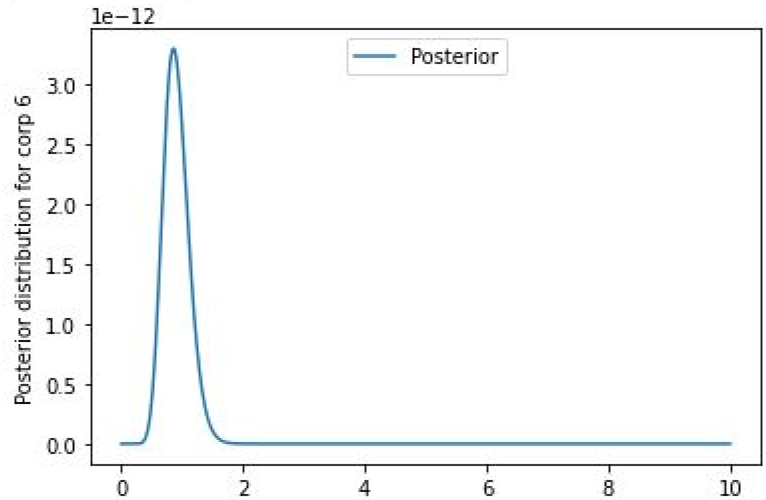



```
(Posterior)Maxima at:  0.43
```

Corp 6



(Likelihood)Maxima at: 0.85



(Posterior)Maxima at: 0.86

The above maximas are obtained after considering the whole dataset for Likelihood and Prior graphs. The graph has been plotted against lambda values with precision 0.1, Maximas for which have been written below the corresponding ones.

Q4

1.

For poisson regression, the likelihood function can be written as

$$p(X, Y, \theta) = \Pi_{i=1}^{N} \frac{e^{y_i \theta x_i} . e^{-\theta x_i}}{y!}$$

That implies log likelihood as

$$log(p(\theta|X, Y) = \sum_{i=1}^{N} (y_i \theta x_i - e^{\theta x_i} - log(y!)) = \sum_{i=1}^{N} (y_i \theta x_i - e^{\theta x_i}) + constant$$

MLE estimate of θ

$$\mathbf{argmax}_\theta \log(p(\theta|X, Y))$$

i.e., theta that maximizes the log-likelihood

Therefore, the loss function that needs to be minimized is

$$-log(p(\theta|X, Y) = \sum_{i=1}^{N} (e^{\theta x_i} - y_i \theta x_i)$$

2. To understand the statistics of the count of bikes in the given Bike Sharing Demand dataset, mean count of bikes used per hour, per month, per year is calculated. The below table compiles the findings

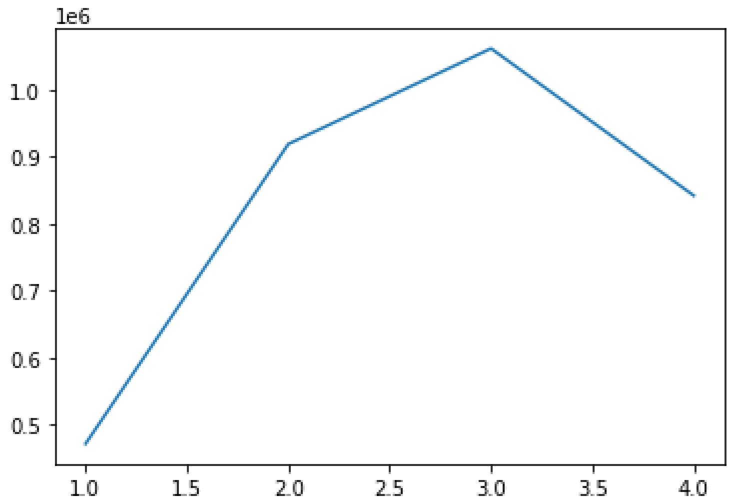
	Mean value
Per year	1646339.5
Per month	182926.6111
Per hour	189.463

3. The dataset consists of 16 dimensions out of which the date as string is removed, to ensure the data is all numerical. Alternatively, we can parse the string, but since the details are present at the level of an hour for each day, parsing a date will only cause redundancy

Thus the dataset used for our experiments consists of 15 dimensions.

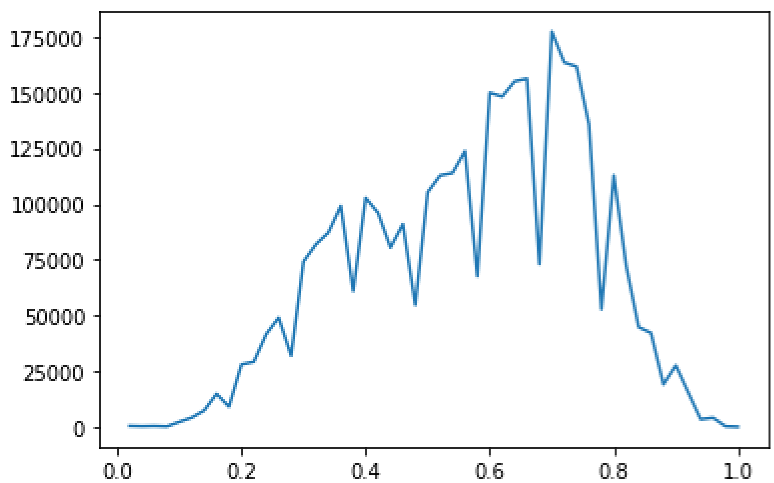
We tried understand the behaviour of 5 selected features (`season, working day, temp, windspeed, casual`) out of 15 and plotted count against each of these features

Below are the findings and inference we can draw from the graphs
season vs count



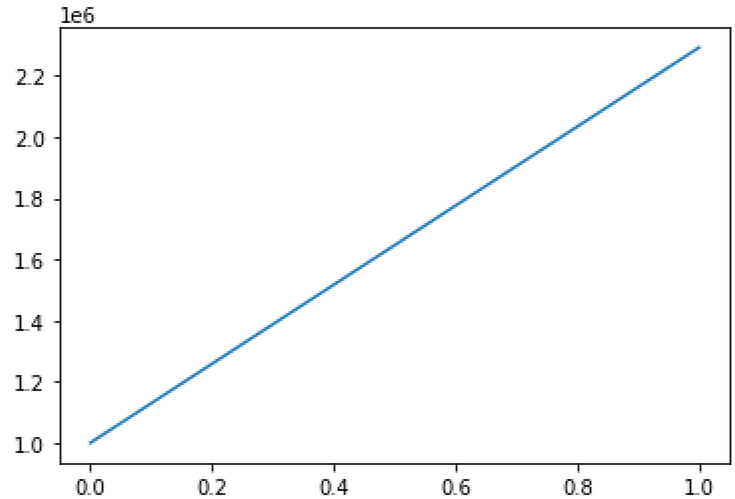
There are four seasons out of which 3rd season is most favourable for bike riding and 1st season is the least favourable

temp vs count



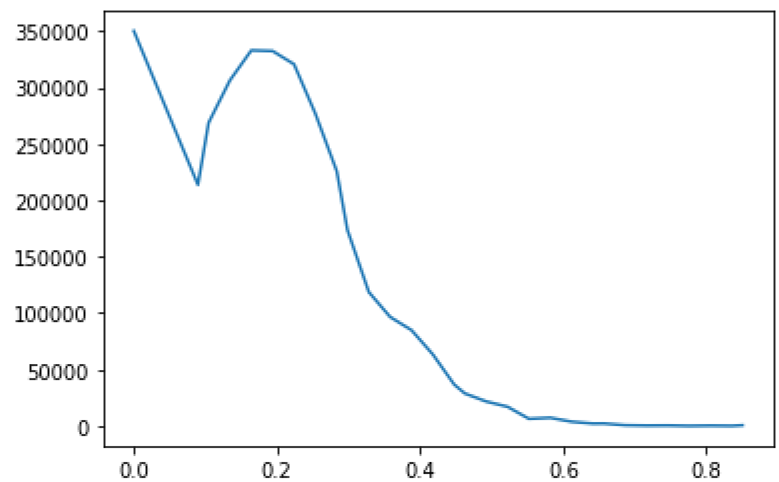
We can infer that temperature effect on the bike demand is not a fixed behaviour, but more sunny is preferred

working day vs count



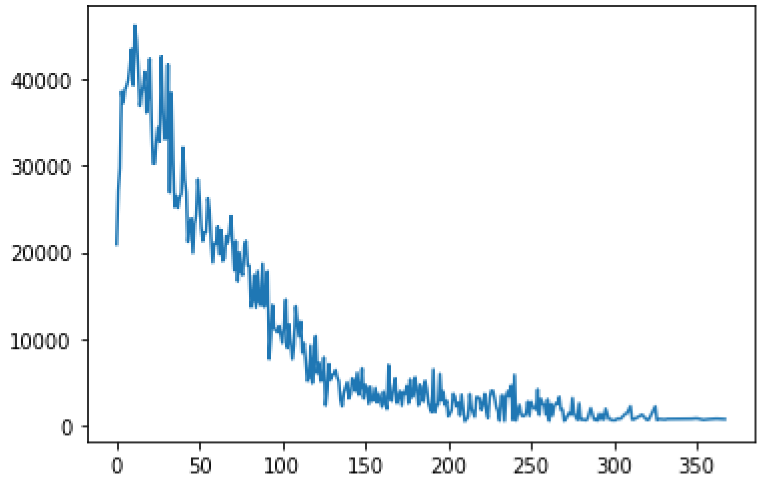
The above graph shows that bike demand is higher during the working day rather than holiday. This implies that most of the people might be transiting to work on bikes.

windspeed vs count



Lower the wind speed higher the demand. But we can observe a falling steep from 0 to 0.1 units. People preferring to have a little breeze during bike driving might be a reason

casual vs count



The above graph shows that most of the demand comes from the people who registered in a bike sharing system instead of people who apply casually now and then without registration.

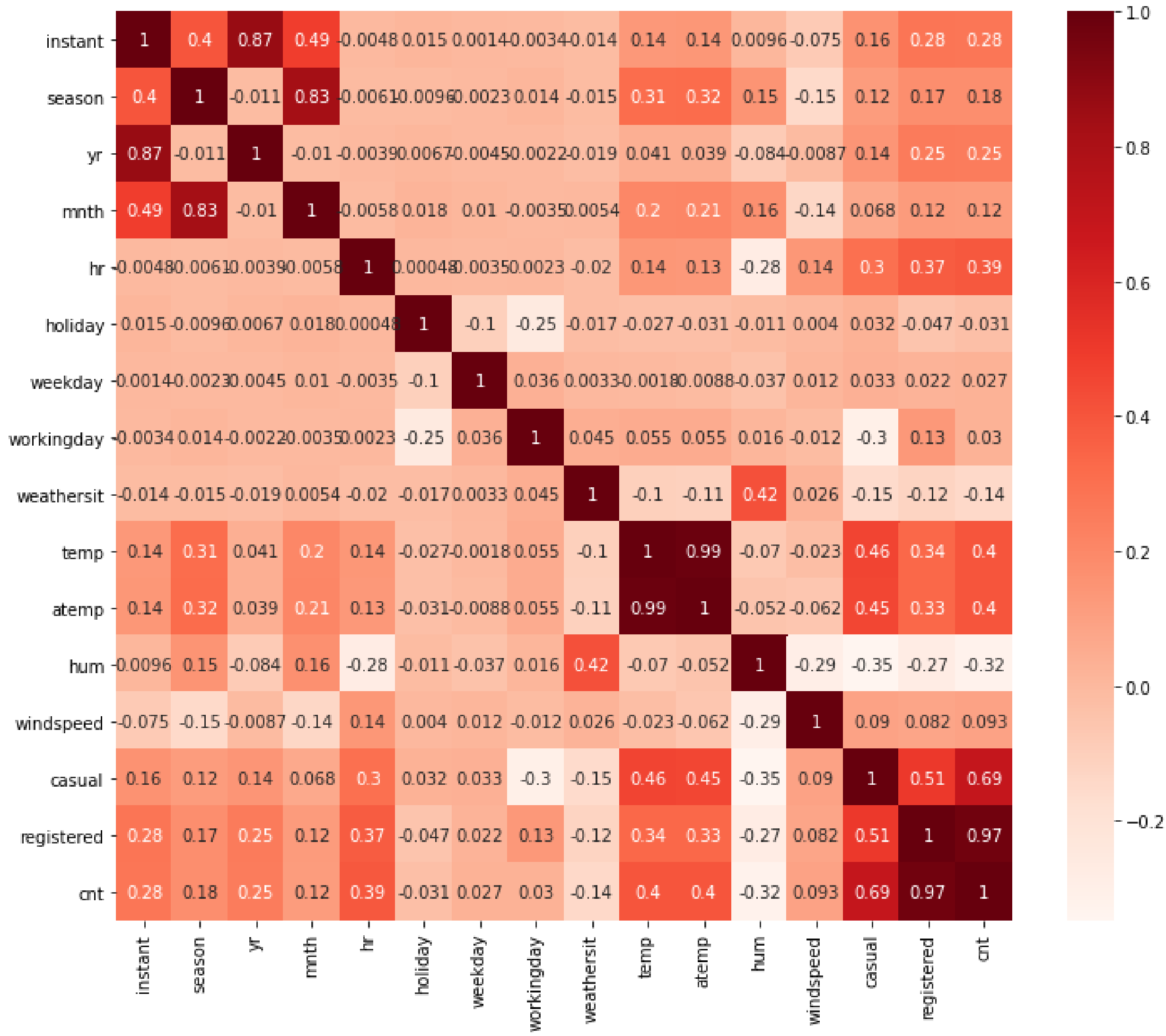
Thus, the data might be collected at a place where people prefer more to transit using bikes.

4.
Dataset is split into 3 parts
Firstly for total dataset 80-20 split is done into training set and testing set
The obtained training data is further is split 80-20 into training and validation sets

The accuracy is reported in terms of the RMSE values on the testing set. Higher the RMSE value, lower the accuracy
For l1-norm and l2-norm regularizations, values are reported based on the best hyperparameter found on the validation set.

	RMSE values
No regularization	9059.956
l1-norm regularization	8834.383
l2-norm regularization	8847.06
both l1 and l2 norm regularization	8913.57

5.
A heatmap is collected to check the correlation between features of the dataset.



This shows that the count of bikes is more correlated with registered, temp, hr, yr, season. Thus we can conclude these form some of the important features to which count is highly correlated with (based on the dataset)