Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Assignment-I

Akash Tadwai - ES18BTECH11019
Vinta Reethu - ES18BTECH11028

November 1, 2020

1. Linear Regression

Consider a linear model of the form

$$y(x, w) = w_0 + \sum_{d=1}^{D} w_d x_d$$

together with a sum-of-squares error function of the form,

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2$$

Now suppose that Gaussian noise ϵ_n with zero mean and variance σ^2 is added independently to each of the input variables x_i.That is, independent noise is added to each dimension of the input. By making use of

$$\mathbb{E}[\epsilon_n] = 0 \; and \quad \mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$$

Show that minimizing $E_D(w)$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay (L_2 norm) regularization term, in which the bias parameter w_0 is omitted from the regularizer.

A: As independent noise is added to each dimension of input variable x_i. Our new model becomes,

$$\begin{aligned} f(x_i, w) &= w_0 + \sum_{d=1}^{D} w_d (x_{id} + \epsilon_{id}) \\ &= w_0 + \sum_{d=1}^{D} w_d x_{id} + \sum_{d=1}^{D} w_d \epsilon_{id} \\ &= y(x_i, w) + \sum_{d=1}^{D} w_d \epsilon_{id} \end{aligned}$$

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

where the noise ϵ_{id} is independent across both the i and d indices. So our new error function is

$$\begin{aligned} E'_D(w) &= \frac{1}{2}\sum_{n=1}^{N}\{f(x_n,w)-t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\left\{y(x_n,w)+\sum_{d=1}^{D}w_d\epsilon_{nd}-t_n\right\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n,w)-t_n)^2+2(y(x_n,w)-t_n)\left(\sum_{d=1}^{D}w_d\epsilon_{nd}\right)+\left(\sum_{d=1}^{D}w_d\epsilon_{nd}\right)^2\right\} \end{aligned}$$

Taking the expectation of this and using the ***linearity of expectation***, we get,

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

As $\mathrm{E}[\epsilon_{nd}]$ is 0, so the second term vanishes .
Now the third term is,

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{d=1}^{D}w_d\epsilon_{nd}\right)^2\right] &= \mathbb{E}\left[\sum_{d=1}^{D}\sum_{d'=1}^{D}w_dw_{d'}\epsilon_{nd}\epsilon_{nd'}\right] \\ &= \sum_{d=1}^{D}\sum_{d'=1}^{D}w_dw_{d'}\mathbb{E}[\epsilon_{nd}\epsilon_{nd'}] \\ &= \sum_{d=1}^{D}\sum_{d'=1}^{D}w_dw_{d'}\delta_{dd'}\sigma^2 \\ &= \sigma^2\sum_{d=1}^{D}w_d^2 \end{aligned}$$

Using these results, we get

$$\begin{aligned} \mathbb{E}[E'_D(w)] &= \frac{1}{2}\sum_{n=1}^{N}\left\{(y(x_n,w)-t_n)^2+\sigma^2\sum_{d=1}^{D}w_d^2\right\} \\ &= E_D(w)+\frac{N}{2}\sigma^2\sum_{d=1}^{D}w_d^2 \end{aligned}$$

and we see that we get a L_2 regularization term without the bias parameter w_0, as desired.

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

2. Multi-Output Regression

Consider the problem where inputs are associated with multiple real valued outputs (K > **1**) known as multi output regression (For e.g. predicting student score across different courses).

$$\mathbf{y}\left(\mathbf{x}, \mathbf{w}\right) = \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

here y is a K-dimensional column vector, W is an $M^{\star}$ K matrix of parameters,and $\boldsymbol{\varphi(x)}$ is an M-dimensional column vector with elements $\varphi_{\mathrm{j}}(\mathrm{x})$, with $\varphi_0(\mathrm{x}) = 1$.

1. Provide the expression for the likelihood, and derive ML and MAP estimates of W in the multi output regression case.
A: Assuming multiple independent outputs we can model this regression as,

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{W}) = \prod_{j=1}^{K} \mathcal{N}\left(y_j \mid \mathbf{w}_j^T \mathbf{x}_i, \sigma_j^2\right)$$

- **MLE**

  The distribution of one of the component of y is given by,

  $$y_{ij} \sim \mathcal{N}\left(\mathbf{w}_j^T \mathbf{x}_i, {\beta_j}^{-1}\right)$$

  where $i \rightarrow 1$ to M and $j \rightarrow$ from 1 to K
  The likelihood of one of the component of y is given by,

  $$p(\mathbf{Y}_j \mid \mathbf{X}, \mathbf{w}_j, \beta_j) = \prod_{i=1}^{M} \mathcal{N}\left(\mathbf{w}_j^T \mathbf{x}_i, \beta_j^{-1}\right)$$

  Taking logarithm on both sides and expanding we get,

  $$\ln p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j) = \sum_{i=1}^{M} \ln \mathcal{N}\left({\mathbf{w_j}}^T \mathbf{x}_i, {\beta_j}^{-1}\right)$$

  Using the density function of a uni variate Gaussian we get,

  $$\begin{aligned} \ln p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j) &= \sum_{i=1}^{M} \ln \frac{1}{\sqrt{2\pi\beta_j^{-1}}} e^{-\left(y_{ij} - \mathbf{w}_j^T \mathbf{x}_i\right)^2 / 2\beta_j^{-1}} \\ &= \frac{M}{2} \ln \beta_j - \frac{M}{2} \ln(2\pi) - \frac{\beta_j}{2} \sum_{i=1}^{M} \left(y_{ij} - {\mathbf{w_j}}^T \mathbf{x}_i\right)^2 \end{aligned}$$

  Notice that this is a quadratic function in $\mathbf{w_j}$, which means that we can solve for it by taking the derivative with respect to $\mathbf{w_j}$, setting that expression to 0, and solving for $\mathbf{w_j}$:

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

$$\frac{\partial \ln p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j)}{\partial \mathbf{w_j}} = \beta_j \sum_{i=1}^{M} \left(y_{ij} - \mathbf{w_j}^T \mathbf{x}_i\right) \mathbf{x}_i^T$$

$$0 = \beta_j \sum_{i=1}^{M} \left(y_{ij} - \mathbf{w_j}^T \mathbf{x}_i\right) \mathbf{x}_i^T$$

$$0 = \sum_{i=1}^{M} y_{ij} \mathbf{x}_i^T - \mathbf{w_j}^T \sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^T$$

Solving for w_j we get

$$\hat{\mathbf{w}}_\mathbf{j} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y_j}$$

We know that the likelihood factorizes across dimensions, the same does MLE.

$$\hat{\mathbf{W}} = [\hat{w}_1, \hat{w}_2, ..., \hat{w}_M]$$

where $\hat{\boldsymbol{w}}_{\boldsymbol{j}}$ is derived above.

- **MAP**

Assuming the Normal Prior with mean 0 and variance $\boldsymbol{S}_0^{-1}$, we get,

$$\mathbf{w_j} \sim \mathcal{N}\left(0, \boldsymbol{S}_0^{-1}\mathbf{I}\right)$$

$$p(\mathbf{Y}_j \mid \mathbf{X}, \mathbf{w}_j, \beta_j) = \prod_{i=1}^{M} \mathcal{N}\left(\mathbf{w}_j^T \mathbf{x}_i, \beta_j^{-1}\right)$$

By Bayes' theorem, posterior can be written as

$$\underbrace{p(\mathbf{w_j} \mid \mathbf{X}, \mathbf{Y_j}, \beta_j)}_{\text{posterior}} \propto \underbrace{p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j)}_{\text{likelihood}} \underbrace{p(\mathbf{w_j})}_{\text{prior}}$$

We now wish to find the value of $\mathbf{w_j}$ that maximizes the posterior distribution. We can maximize the log of the posterior with respect to w

$$\ln p(\mathbf{w_j} \mid \mathbf{X}, \mathbf{Y_j}, \beta_j) \propto \ln p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j) + \ln p(\mathbf{w_j})$$

From above derivations we already know the value of

$$\ln p(\mathbf{Y_j} \mid \mathbf{X}, \mathbf{w_j}, \beta_j) = \sum_{i=1}^{M} \ln \mathcal{N}\left(\mathbf{w_j}^T \mathbf{x}_i, {\beta_j}^{-1}\right)$$

Calculating the value of $\ln \mathbf{p}(\mathbf{w_j})$

$$\begin{aligned}
\ln p(\mathbf{w_j}) &= \ln \mathcal{N}\left(0, \boldsymbol{S}_0^{-1}\mathbf{I}\right) \\
&= \ln \frac{1}{\left(\left|2\pi \boldsymbol{S}_0^{-1}\mathbf{I}\right|\right)^{\frac{1}{2}}} \exp\left\{-\frac{\boldsymbol{S}_0}{2} \mathbf{w_j}^T \mathbf{w_j}\right\} \\
&= \mathbf{C} - \frac{\boldsymbol{S}_0}{2} \mathbf{w_j}^T \mathbf{w_j}
\end{aligned}$$

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

By taking derivative and solving for $\boldsymbol{w_j}$
Finally we get,

$$\hat{\boldsymbol{w_j}} = (\boldsymbol{X}^\tau \boldsymbol{X} + \lambda \boldsymbol{M}\boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{Y_j}$$

where $\lambda = \frac{S_0}{\beta}$
We know that the likelihood factorizes across dimensions, the same does MAP.

$$\hat{\mathbf{W}} = [\hat{w}_1, \hat{w}_2, ..., \hat{w}_M]$$

where $\hat{\boldsymbol{w}}_{\boldsymbol{j}}$ is derived above.

2. Consider a multi-output regression problem where we have multiple independent outputs in linear regression. Let's consider a 2 dimensional output vector $y_i \in \mathrm{R}^2$. Suppose we have some binary input data, $\mathrm{x_i} \in \{0, 1\}$. The training data is as given in the right side. Let us embed each x into 2d using the following basis function:
$\varphi\ (0) = (1, 0)^\top, \varphi(1) = (0, 1)^\top$. The model becomes $y = W^\top \varphi(x)$ where $W = [w_1, w_2]$ is a 2 x 2 matrix, with both w_1 and w_2 column vectors. Find the MLE for w_1 and w_2.

.

x	y
0	$(-1, -1)^T$
0	$(-1, -2)^T$
0	$(-2, -1)^T$
1	$(1, 1)^T$
1	$(1, 2)^T$
1	$(2, 1)^T$

A: We know that

$$\boldsymbol{Y} = \boldsymbol{W^T \varphi(x)}$$

By expanding this we get,

$$\begin{pmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{w}}_1^T \\ \hat{\mathbf{w}}_2^T \end{pmatrix} \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix}$$

Our design matrix $\mathbf{X}$ which is obtained by plugging the values if x in our data set in $\varphi(\mathbf{x})$

$$\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{y}_1 = \begin{pmatrix} -1 \\ -1 \\ -2 \\ 1 \\ 1 \\ 2 \end{pmatrix} \mathbf{y}_2 = \begin{pmatrix} -1 \\ -2 \\ -1 \\ 1 \\ 2 \\ 1 \end{pmatrix}$$

$$\hat{\mathbf{w}_1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y_1}$$

$$\hat{\mathbf{w}_2} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y_2}$$

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

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

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

By multiplying by y_1 we get,

$$\hat{\mathbf{w}}_1 = \begin{pmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{pmatrix}$$

By multiplying by y_2 we get,

$$\hat{\mathbf{w}}_2 = \begin{pmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{pmatrix}$$

Hence our MLE of $\mathbf{W}$ is,

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

3. ML and MAP estimation of Poisson Distribution

- **Derivation of maximum likelihood estimation :**

  Suppose that $X = (X_1, X_2, \ldots, X_n)$ are iid observations from a Poisson distribution with unknown parameter λ. The likelihood function is:

$$f(x_i, \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$L(\lambda, \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

$$I(\lambda) = \log L(\lambda, \boldsymbol{x}) = -n\lambda + \sum x_i \log\lambda - \log\left\{\prod_{i=1}^{n} x_i!\right\}$$

$$\frac{dl}{d\lambda} = -n + \frac{\sum x_i}{\lambda} = \frac{\sum x_i - n\lambda}{\lambda} \begin{cases} > 0 & \text{if } \lambda < \bar{x} \\ < 0 & \text{if } \lambda > \bar{x} \end{cases}$$

$\hat{\lambda}_{MLE} = \bar{X}$ is the MLE of λ.

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **Derivation of maximum aposteriori estimation:**

  According to Baye's theorem

$$\mathcal{P}(\lambda \mid x) = \frac{P(x \mid \lambda) \cdot P(\lambda)}{P\,(x)}, \quad \text{where } P(x) \text{ is independent of } \lambda$$

  We know that

$$\mathcal{P}(x \mid \lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

  Why we chose Gamma Prior?
  We chose gamma prior as it is conjugate distribution for poission.
  Taking Gamma Prior,

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

  Now applying Baye's theorem,

$$\begin{aligned} P(\lambda \mid x) &= \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{N} x_i!} \cdot \frac{\lambda^{\alpha-1} \cdot e^{-\beta\lambda}}{\Gamma(\alpha)} \\ &= k \cdot \lambda^{(\sum x_i + \alpha - 1)} \cdot e^{-\lambda(n+\beta)} \end{aligned}$$

$$\Rightarrow \text{Gamma}(\sum x_i + \alpha, \beta + N)$$

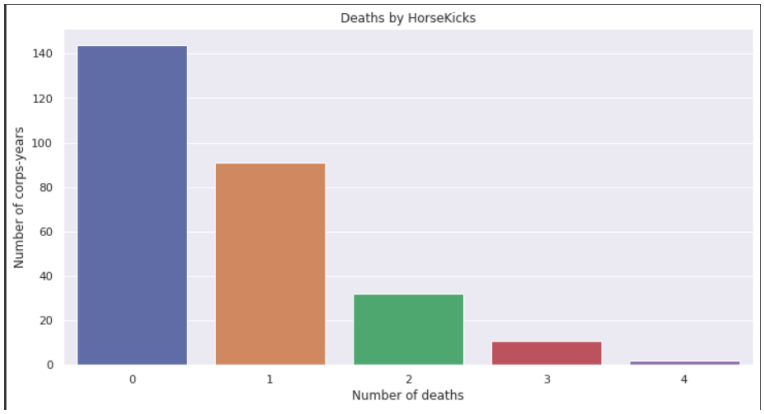
  Hence the MAP estimate would be the mode of the posterior distribution.
  We know that the Mode of Gamma distribution is, $\frac{\alpha-1}{\beta}$
  $\hat{\lambda}_{MAP} = \frac{\sum x_i + \alpha - 1}{N+\beta}$ is the MAP of λ.

- **Procedure :**
  - First we obtained the data from this link using **wget.**
  - Using the above mathematical equations we obtained MLE and RMSE.
  - We have plotted the Number of corp years vs Number of deaths to visualize the data.
  - Next we obtained MAP estimates by choosing suitable alpha, beta which is further explained in Observations below.
  - We plotted Likelihood,Prior,Posterior graphs by using the above derived mathematical equations.

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **Plots :**
  - **Final plot of number of deaths by Horsekicks:**



  - **Lambda estimate and RMSE (predictions) for each corp using MLE:**

```
Lambda estimate and RMSE (predictions) for each corps are :
Corp lambdaEstimate     RMSE
GC:     1.0             0.7559
C1:     0.6923          1.1124
C2:     0.6154          0.7298
C3:     0.6154          0.7298
C4:     0.4615          0.4848
C5:     0.3846          0.588
C6:     0.8462          0.9898
C7:     0.5385          0.898
C8:     0.3077          0.5094
C9:     0.6923          0.7384
C10:    0.5385          1.1597
C11:    1.0             1.1339
C14:    1.4615          1.0238
C15:    0.3077          0.9412
```

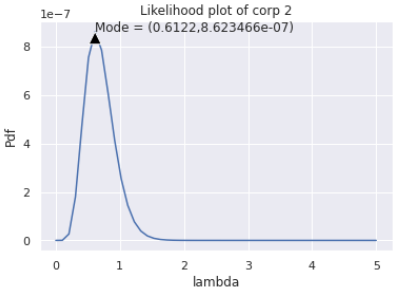
  - **Lambda estimate and RMSE (predictions) for each corp using MAP:**

```
Lambda estimate and RMSE (predictions) for each corps are :
Corp    lambdaEstimate  RMSE
GC:     0.9666          0.731
C1:     0.6809          1.1157
C2:     0.6095          0.7294
C3:     0.6095          0.7294
C4:     0.4666          0.4866
C5:     0.3952          0.5795
C6:     0.8238          0.9903
C7:     0.538           0.8981
C8:     0.3238          0.5058
C9:     0.6809          0.7366
C10:    0.538           1.1599
C11:    0.9666          1.1552
C14:    1.3952          0.9764
C15:    0.3238          0.9368
```

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

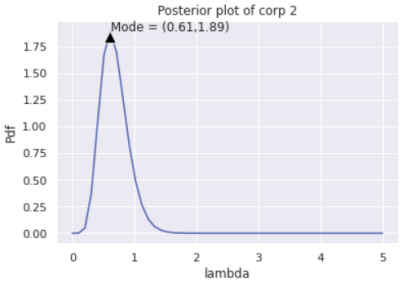
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Likelihood, Prior and Posterior Plots for 2,4,6 corps



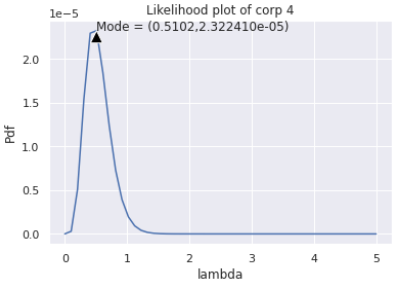
(a) *Likelihood*

(b) *Prior*

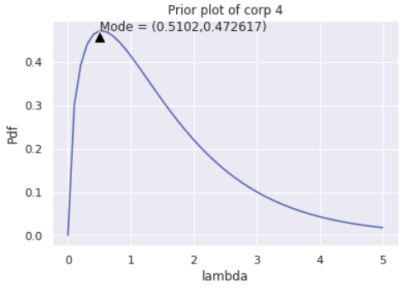


(c) *Posterior*

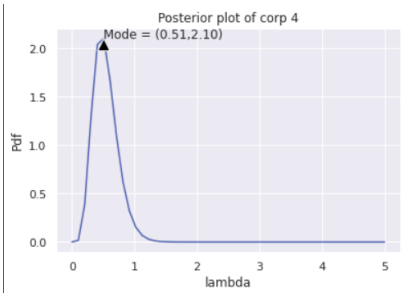
Figure 1: Corp 2



(a) *Likelihood*

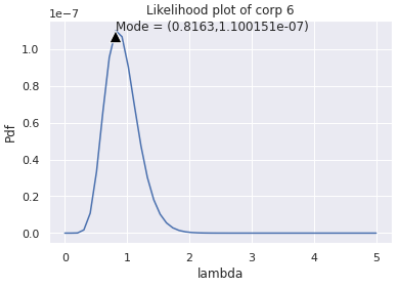


(b) *Prior*



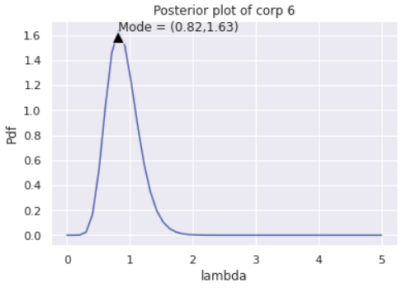
(c) *Posterior*

Figure 2: Corp 4



(a) *Likelihood*

(b) *Prior*



(c) *Posterior*

Figure 3: Corp 6

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **Observations:**
  - As Gamma distribution is the prior for Poisson distribution. We chose α and β such that, Mode of the prior is tending to zero, as parameters estimated using MLE are close to zero. As some of the lambdas are also tending to 1 we choose the Gamma distribution such that the Variance is maximum possible.

$$\begin{aligned} Mode &\to max(0, \frac{\alpha - 1}{\beta}) \\ Variance &\to \frac{\alpha}{\beta^2} \end{aligned} \tag{1}$$

    After clear examination of the effect of different alpha,betas on training data, we chose: $\alpha \to 1.5325$, $\beta \to 1$

  - **Mode of distributions :**
    We plotted Likelihood, Prior, Posterior plots for the following corps and used the graph to obtain the mode of distributions. We can see from the graphs that posterior density graph has density **somewhere between** prior and likelihood densities. We know that mode means the maximum number in the given set of observations, In our case mode of distribution gives the lambda for which we have the highest probability density. Among 2,4,6 corps *Corp 6 has the highest posterior Lambda* and hence are more likely to be dead by HorseKicks.
    - **Corp 2:**
      Likelihood : [0.6122449, 8.623466e-07]
      Prior : [0.5102, 0.472617]
      Posterior : [0.61, 1.89]
    - **Corp 4:**
      Likelihood : [0.51020408, 2.322410e-05]
      Prior : [0.5102, 0.472617]
      Posterior : [0.51, 2.10]
    - **Corp 6:**
      Likelihood : [0.81632653, 1.100151e-07]
      Prior : [0.5102, 0.472617]
      Posterior : [0.82, 1.63]

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

4. Bike Sharing Demand

- **Mean :**

  If $\mathbf{x} \in \mathbb{R}^n$ is a vector of independent variables, We know that Poisson regression model takes the form,

  $$\log(\mathrm{E}(Y \mid \mathbf{x})) = \alpha + \beta^T \mathbf{x}$$

  where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$.
  Given a Poisson regression model θ and an input vector x, the predicted mean of Poisson distribution is given by

  $$\mathrm{E}(Y \mid \mathbf{x}) = e^{\theta^T \mathbf{x}}$$

- **Maximum likelihood estimation in Poisson regression :**

  Given a set of parameters θ and an input vector x, the mean of the Poisson distribution, is given by

  $$\lambda = \mathrm{E}(Y \mid x) = e^{\theta^T x}$$

  and thus, the Poisson probability mass function is given by

  $$p(y \mid x; \theta) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{e^{y\theta^T x} e^{-e^{\theta^T x}}}{y!}$$

  Now suppose we are given a data set consisting of m vectors $x_i \in \mathbb{R}^{n+1}, i = 1, \ldots, m$, along with a set of m values $y_1, \ldots, y_m \in \mathbb{N}$. Then, for a given set of parameters θ, the probability of attaining this particular set of data is given by

  $$p\left(y_1, \ldots, y_m \mid x_1, \ldots, x_m; \theta\right) = \prod_{i=1}^{m} \frac{e^{y_i \theta^T x_i} e^{-e^{\theta^T x_i}}}{y_i!}$$

  By the method of maximum likelihood, we want to find the set of parameters θ that makes this probability as large as possible.

  $$L(\theta \mid X, Y) = \prod_{i=1}^{m} \frac{e^{y_i \theta^T x_i} e^{-e^{\theta^T x_i}}}{y_i!}$$

  Now we take the log on both sides of Likelihood function:

  $$\ell(\theta \mid X, Y) = \log L(\theta \mid X, Y) = \sum_{i=1}^{m} \left(y_i \theta^T x_i - e^{\theta^T x_i} - \log\left(y_i!\right) \right)$$

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

Notice that the parameters θ only appear in the first two terms of each term in the summation. Therefore, given that we are only interested in finding the best value for θ we may drop the y_i^T and simply write

$$\ell(\theta \mid X, Y) = \sum_{i=1}^{m} \left(y_i \theta^T x_i - e^{\theta^T x_i} \right)$$

As we don't get closed form solution we minimise the negative log-likelihood and update the weights.

- **Cost function :**

We take the cost function as the negative log likelihood,

$$J\left(\theta_0, \theta_1, \theta_2, \ldots\right) = -\frac{1}{2m} \sum_{i=1}^{m} \left(y_i \theta^T x_i - e^{\theta^T x_i} \right)$$

By taking the derivative of loss function, $-\ell(\theta \mid X, Y)$ w.r.t θ,
we get,

$$\nabla \mathrm{E}(\mathbf{w}) = \frac{\partial J(\theta \mid X, Y)}{\partial \theta} = -\frac{1}{2m} \sum_{i=1}^{m} \left(y_i x_i - e^{\theta^T x_i} x_i \right)$$

Hence the Gradient Descent Update is as follows,

$$w^{new} = w - \frac{\eta}{2m} \sum_{i=1}^{m} \left(e^{\theta^T x_i} - y_i \right) x_i$$

η : Learning Rate

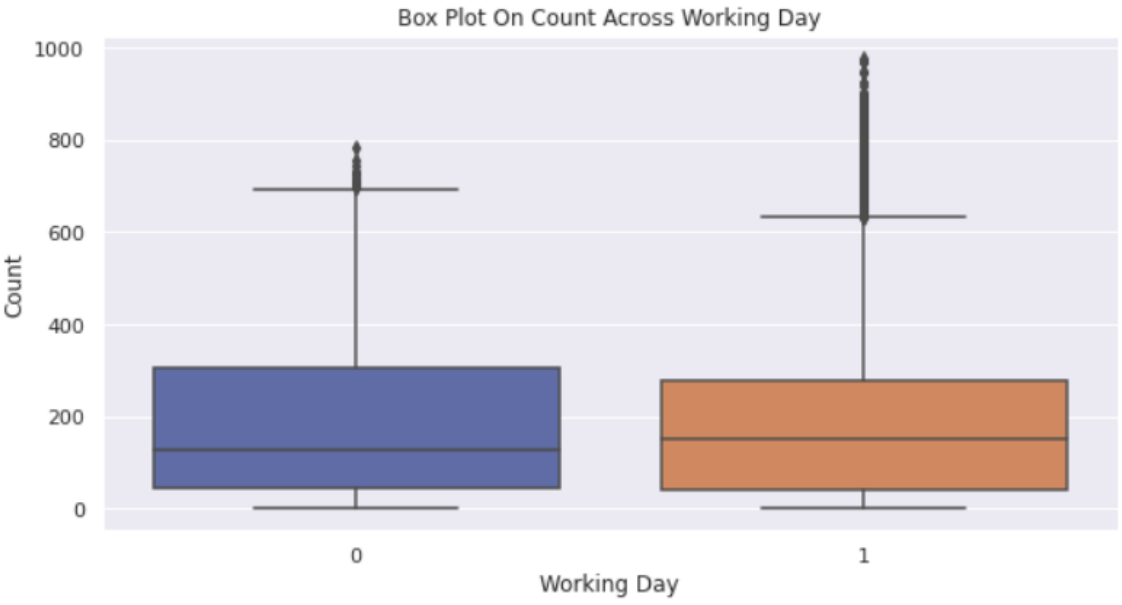
- **Statistics of the dataset :**

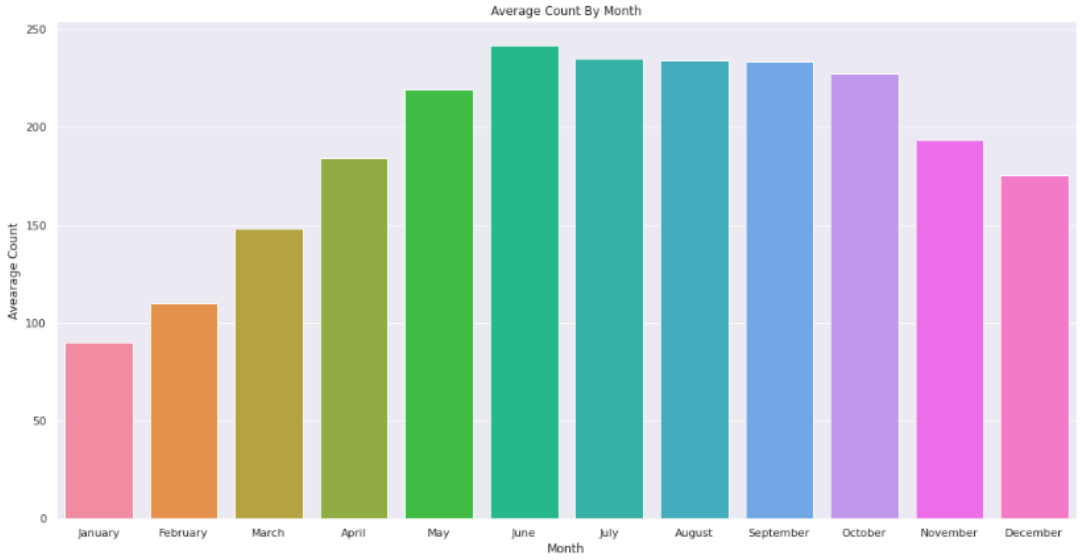
The mean count per month is 191.14315
The median count per month is 205.29504
The mean count per each day of week is 190.69694
Standard Deviation of temp is 0.192556
Minimum counts in a day is 1.00

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

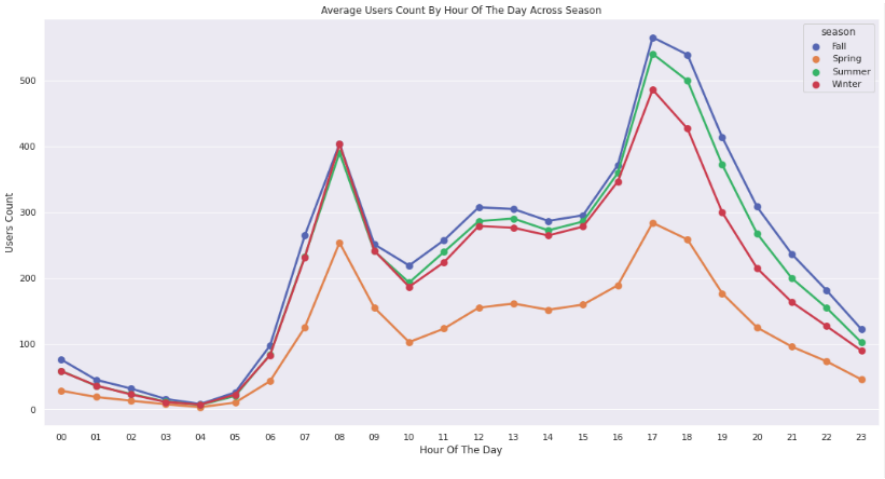
- **Plots:**

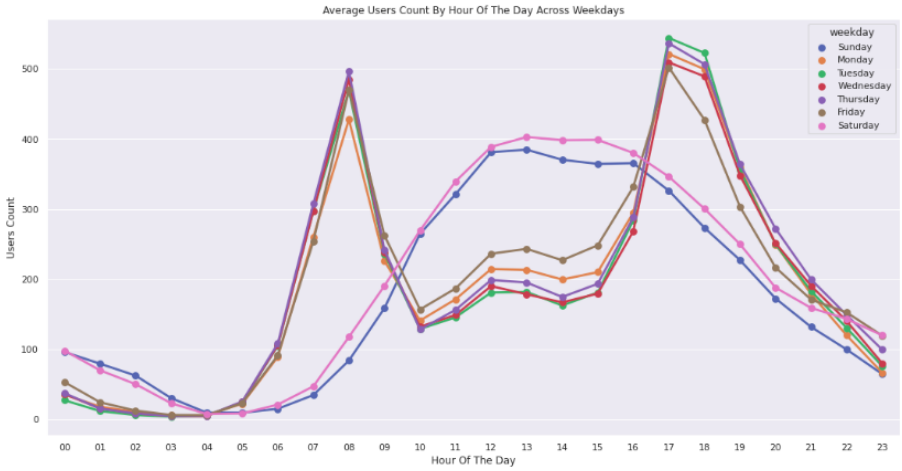


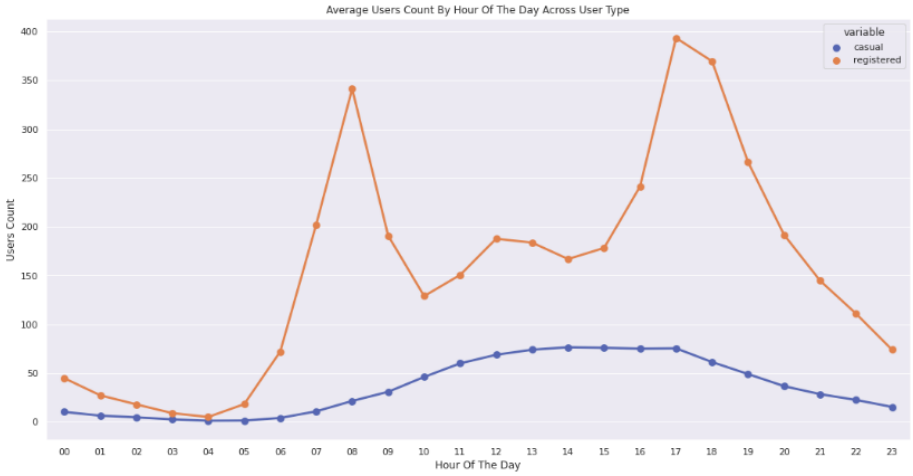


Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad







Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **Procedure :**
  - First we obtained the data from this link using **wget.**
  - We used the variables present in the dataset to plot the graphs on training set.
  - Using the numerical variables we plotted the heat map to see the correlated features. Further explanation of this is given in observation section below.
  - Next we took the categorical features and did the one hot encoding on those vectors.
  - We know that hour,weekday,season,year,weather,holiday,working day are categorical variables.
  - So we did one-hot encoding on these variables, We have dropped one of the highly correlated features (eg. temp and atemp)
  - We used Poisson regression to model the count and obtained the hyperparameter by tuning for different values.
  - We have done Poisson regression with and without regularisation and obtained the values, which are given in Observation section below.

- **Observations:**

From the correlation plot above we can see that,

- Count variable has got little dependency on ”temp” and ”humidity”.
- ”Casual” and ”Registered” features should be dropped since they are just leakage variables.
- ”atemp” and ”temp” are highly correlated and hence one of them is to be dropped.

Akash Tadwai
Vinta Reethu
Indian Institute of Technology Hyderabad
Assignment-I

• After applying L1 and L2 norm regularization over weight vectors, and find the best hyper-parameter settings for the mentioned problem using validation data we obtained :

- **No regularisation :**
  RMSE : 108.6739
  Learning Rate : 4*1e-3
  Iterations : 650
- **L1 Norm :**
  RMSE : 108.8211
  Learning Rate : 4.5*1e-3
  Iterations : 700
  Lambda : 2
- **L2 Norm :**
  RMSE : 108.8151
  Learning Rate : 4.5*1e-3
  Iterations : 650
  Lambda : 0.5

• **Important Features:**

- By inferring from the weights of the model with and without regularisation and the correlation plot we have drawn above, we see that **hour of the day, season and humidity** are playing an important role in estimation of Counts in the dataset.
- Variables such as **casual**,**registered** are leaky variables and hence removed while training.
- Highly correlated variables are **temp and atemp** and we removed **atemp** while training.

Colab Notebook links:

- Horse-Kicks
- Bike-Sharing

******THE END******