Solutions : Name : Achyut Kulkarni

For all values here I have taken W= $\theta \Rightarrow L(w) \Rightarrow L(\theta)$ **and so on ...**

1) Logistic Regression :

a) Lets take W= $\theta \Rightarrow L(w) \Rightarrow L(\theta)$

w.k.t for Binary logistic regression :

P(y=1|x; θ) = $h_\theta(x) = \frac{1}{1+e^{-\theta^T X}}$ => P(y=0|x; θ) = $1 - h_\theta(x) = 1 - \frac{1}{1+e^{-\theta^T X}}$

Since there are only two classes we can define this as :

$L(\theta) = -\log(\prod_{i=1}^{n} p(y = y_i|x_i)) = -\log(\prod_{i=1}^{n} (h_\theta(x)^{y_i} * (1 - h_\theta(x))^{1-y_i}))$

=> $-y_{mi} \sum_{i=1}^{n} log(h_\theta(x)) - (1 - y_i) \sum_{i=1}^{n} log((1 - h_\theta(x)))$

b)

From the final equation in above answer ., we have

$-y_i \sum_{i=1}^{n} log(h_\theta(x)) - (1 - y_i) \sum_{i=1}^{n} log((1 - h_\theta(x)))$ ---------(1)

And we also know that

P(y=1|x; θ) = $h_\theta(x) = \frac{1}{1+e^{-\theta^T X}}$ => P(y=0|x; θ) = $1 - h_\theta(x) = 1 - \frac{1}{1+e^{-\theta^T X}}$

Substituting in 1, we get

However w.k.t Gradient descent is

$\theta_j : \theta_j - \alpha \frac{\partial}{\partial \theta} J(\theta)$

For n>=1, we have $J(\theta) = (-y_i \sum_{i=1}^{n} log(\frac{1}{1+e^{-\theta^T X}}) - (1 - y_i) \sum_{i=1}^{n} log((1 - \frac{1}{1+e^{-\theta^T X}})))$

Hence;

$$\frac{\partial}{\partial\theta}J(\theta) = \left(-y_i \sum_{i=1}^{n} log(\ \sigma\left(\theta^T x_n\right)\ ) - (1-y_i) \sum_{i=1}^{n} log((1-\sigma\left(\theta^T x_n\right)\ ))\ \right)$$

$$\frac{\partial}{\partial\theta}J(\theta) = \left(-\sum_{i=1}^{n} y_n\ (1 - \ \sigma\left(\theta^T x_n\right)\ x_n) - (1-y_n)(\sigma\left(\theta^T x_n\right) x_n))\ \right)$$

$$\frac{\partial}{\partial\theta}J(\theta) = \sum_{i=1}^{n}((\ \sigma\left(\theta^T x_n\right)\ - y_n\ )\ x_n)$$

Hence,

$$\theta_j : \theta_j - \alpha\frac{\partial}{\partial\theta}J(\theta)$$

$$\theta_j : \theta_j - \alpha \sum_{i=1}^{n}((\ \sigma\left(\theta^T x_n\right)\ - y_n\ )\ x_n)$$

Intuitively, a (strictly) convex function has a “bowl shape”, and hence has a unique global minimum θ∗ corresponding to the bottom of the bowl. Hence its second derivative must be positive everywhere, d f(θ) > 0. A twice-continuously differentiable, multivariate function f is dθ convex if its Hessian is positive definite for all θ.

Now lets differentiate $\sum_{i=1}^{n}((\ \sigma\left(\theta^T x_n\right)\ - y_n\ )\ x_n)$ second time and see if its > 0

Hence ;’

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}((\ \sigma\left(\theta^T x_n\right)\ - y_n\ )\ x_n) = \sigma\left(\theta^T x_n\right)(1 - \ \sigma\left(\theta^T x_n\right))x_n^2 \qquad .....1$$

Since , the sigmoid function lies between zero and 1 => $\sigma\left(\theta^T x_n\right)$ >0
=> (1-$\sigma\left(\theta^T x_n\right)$) > 0 too hence
The whole equation is greater than zero hence, the

Hessian H is > 0.

Therefore we can say that the function is convex and thus converges at only global minimum.

c) Lets take w= θ
Hence for K classes :

P(y=k|x; θ) = $h_\theta(x) = \frac{e^{\theta^T X}}{1+\sum_{1}^{k-1} e^{\theta^T X}}$ for K = 1,2,3,..... K-1

=> P(y=K|x; θ) = $1 - h_\theta(x) = 1 - \frac{1}{1+\sum_{1}^{k-1} e^{\theta^T X}} => \frac{1}{1+\sum_{1}^{k-1} e^{\theta^T X}} for\ k = K$

Negative Log Likelihood of the above equation is L($\theta_1, \theta_2, \theta_3 ..... \theta_k$)

$$L(\theta_1, \theta_2, \theta_3 ..... \theta_k) = -log(\prod_{i=1}^{n} p(y = y_i|x_i) = -\sum_{1}^{n}\sum_{1}^{k-1}(log(\frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})) - \sum_{1}^{n} log(\frac{1}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})$$

Also; we can do ;

$$L(\theta_1, \theta_2, \theta_3 ..... \theta_k) = -log(\prod_{i=1}^{n} p(y = y_i|x_i) = -(log(\prod_{i=1}^{n}\prod_{i=1}^{k-1} \frac{e^{\theta_t^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}}(\frac{1}{1+\sum_{1}^{k-1} e^{\theta_t^T X}}) \text{ -----}$$

$\frac{1}{1+\sum_{1}^{k-1} e^{\theta_t^T X}} = \frac{e^0}{1+\sum_{1}^{k-1} e^{\theta_t^T X}} = \frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}}$ as we know $\theta_K^T X = 0$

$$-\sum_{1}^{n}(\sum_{1}^{k-1}(log(\frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})) + log(\frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})) = -\sum_{1}^{n}(\sum_{1}^{k}(log(\frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})^{y_{ik}}) \text{ ; } where\ y$$

$$= -\sum_{1}^{n}(\sum_{1}^{k}(log(\frac{e^{\theta_K^T X}}{1+\sum_{1}^{k-1} e^{\theta_t^T X}})^{y_{ik}}))$$

$$= -\sum_{1}^{n}\sum_{1}^{k}\theta_k^T x y_{ik} - y_{ik} log(1 + \sum_{1}^{k-1} e^{\theta_t^T X})$$

This reduces to :

$$-\sum_{1}^{n}\sum_{1}^{k}\theta_k^T x y_{ik} - y_{ik} log(1 + \sum_{1}^{k-1} e^{\theta_t^T X})$$

d)

Now w.k.t for gradient descent

$\theta_j : \theta_j - \alpha\frac{\partial}{\partial\theta} J(\theta)$

And

$$J(\theta) = -\sum_{i=1}^{n}\sum_{k=i}^{k}\theta_k^T x y_{ik} - y_{ik} log(1 + \sum_{1}^{k-1} e^{\theta_t^T X}) - \sum_{i=1}^{n}\sum_{k\neq i}^{k-1}\theta_k^T x y_{ik} - y_{ik} log(1 + \sum_{1}^{k-1} e^{\theta_t^T X})$$

Hence

$$\frac{\partial}{\partial\theta_i} J(\theta) = -\sum_{1}^{n}(\sum_{k=i}^{k} x_{ij} y_{ik} - y_{ik}(\frac{\sum_{1}^{k} e^{\theta_t^T X} x_{ij}}{1+\sum_{1}^{k} e^{\theta_t^T X}}) + \sum_{k\neq i}^{k} - y_{ik}(\frac{\sum_{1}^{k} e^{\theta_t^T X} x_{ij}}{1+\sum_{1}^{k} e^{\theta_t^T X}}))$$

$$\frac{\partial}{\partial\theta_i} J(\theta) = -\sum_{n}\left[y_{ik} x_i - \sum_{k} \frac{x_{ij} y_{ik} e^{\theta_k^T Xi}}{1+\sum_{t=1}^{K-1} exp(w_t^T x_n)} \right]$$

We know that $P(Y=k \mid X=x) = \frac{exp(w_k^T x)}{1+\sum_1^{K-1} exp(w_l^T x)}$,

$$\frac{\partial L(w_1,w_2,...,w_k)}{\partial w_i} = -\sum_n x_n \left(\left[y_{nk} - \sum_k P(Y=k \mid X=x) y_{nk} \right] \right)$$

2) Linear/Gaussian Descriminant analysis :

a)
W.k.t :
Lets denote $x_n = x \; and \; y_n = y \; for \; the \; ease$

$P(x_n, y_n) = P(y_n) * P(x_n \mid y_n)$

=> $P(x_n, y_n) = \left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \; if \; y_n = 1 \right\}$ and

$P(x_n, y_n) = \left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \; if \; y_n = 2 \right\}$

Where $p1 + p2 = 1$
Hence p2 = 1 - p1

Log likelihood is :

$L(D) = \sum_1^2 log(P(x_n, y_n)$

=> $\sum_{n:\, y_n=1} log(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\}) + \sum_{n:\, y_n=2} log(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\})$

Now to calculate MLE's of parameters:

i)p_1^*

$p_1^* = \frac{\partial}{\partial p1}($

$\sum_{n:\, y_n=1} log(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\}) + \sum_{n:\, y_n=2} log(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\}) = 0$

p_1^* = $\sum_{n:\, y_n=1} (\frac{1}{p1}) + \sum_{n:\, y_n=1} (\frac{1}{1-p1}) \Rightarrow 0$

Lets denote

$$\sum_{n:\, y_n=1} 1 = n1 \qquad and \sum_{n:\, y_n=2} 1 = n2 \; .....hence$$

$$\frac{n1}{p1} - \frac{n2}{1-p1} = 0 \qquad => \quad p1 = \frac{n1}{n1+n2}$$

$$p_1^* = \frac{n1}{n}$$

ii) p_2^*

$$p_2^* = \frac{\partial}{\partial p1}\Big(\sum_{n:\, y_n=1} log\Big(\Big\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\Big(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\Big) \Big\} \Big) + \sum_{n:\, y_n=2} log\Big(\Big\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\Big(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\Big) \Big\} \Big) = 0$$

$$p_2^* = \sum_{n:\, y_n=1} \Big(\frac{1}{p1}\Big) + \sum_{n:\, y_n=1} \Big(\frac{1}{p2}\Big) \Rightarrow 0$$

Lets denote

$$\sum_{n:\, y_n=1} 1 = n1 \qquad and \sum_{n:\, y_n=2} 1 = n2 \; .....hence$$

$$\frac{n1}{1-p1} - \frac{n2}{p2} = 0 \qquad => \quad p2 = \frac{n2}{n2+n1}$$

$$p_2^* = \frac{n2}{n}$$

iii) μ_1^*

$$\mu_1^* = \frac{\partial}{\partial \mu_1}\Big(\sum_{n:\, y_n=1} log\Big(\Big\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\Big(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\Big) \Big\} \Big) + \sum_{n:\, y_n=2} log\Big(\Big\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\Big(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\Big) \Big\} \Big) = 0$$

This reduces to ;

$$\mu_1^* = \sum_{n1} log\Big(\{ exp\Big(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\Big) \}\Big) + 0 = 0$$

$$\frac{\partial}{\partial \mu_1}\Big(-\frac{1}{2\sigma_1^2}\sum_{n1} ((x_n - \mu_1)^2) \}\Big)\Big) + 0 = 0$$

$$\sum_{n1} \left((x_n - \mu_1) \right) = 0$$

$$\Rightarrow \mu_1^* = \frac{\sum_{n1} x_n}{n1}$$

iv) $\mu_2^* = \frac{\partial}{\partial \mu_2}($

$$\sum_{n:\, y_n=1} log(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\}) + \sum_{n:\, y_n=2} log(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\}) = 0$$

This reduces to ;

$$\mu_2^* = \sum_{n2} log(\{ exp\, (-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \}) + 0 = 0$$

$$\frac{\partial}{\partial \mu_2}(-\frac{1}{2\sigma_2^2} \sum_{n2} ((x_n - \mu_2)^2) \})) + 0 = 0$$

$$\sum_{n2} \left((x_n - \mu_2) \right) = 0$$

$$\Rightarrow \mu_2^* = \frac{\sum_{n2} x_n}{n2}$$

v) σ_1^*

$\sigma_1^* = \frac{\partial}{\partial \sigma_1^*}($

$$\sum_{n:\, y_n=1} log(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\}) + \sum_{n:\, y_n=2} log(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\}) = 0$$

This reduces to ;

$$\sigma_1^* = \frac{\partial}{\partial \sigma_1^*} (\sum_{n:\, y_n=1} log(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\}) + 0 = 0$$

$$\sigma_1^* = \frac{\partial}{\partial \sigma_1^*} (\sum_{n:\, y_n=1} \left\{ log\{ \frac{1}{\sqrt{2\pi\sigma_1^2}} \} + log\{exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2})\} \right\} = 0)$$

$$\sigma_1^* = \frac{\partial}{\partial \sigma_1^*} \sum_{n:\, y_n=1} \{ \frac{-1}{2} log\sigma_1^2 = \frac{(x_n - \mu_1)^2}{2\sigma_1^2}$$

$$\sigma_1^* = \frac{\partial}{\partial \sigma_1^*} \sum_{n:\, y_n=1} \{ log\sigma_1^2 = \frac{(x_n - \mu_1)^2}{\sigma_1^2} = 0 \}$$

$$\sigma_1^* = 2\frac{n1}{\sigma_1} = \frac{-2(x_n - \mu_1)^2}{\sigma_1^3}$$

$$\sigma_1^* = \frac{\sum(x_n - \mu_1)}{\frac{n1}{\sqrt{n1}}}$$

vi)

σ_2^*

$$\sigma_2^* = \frac{\partial}{\partial \sigma_2^*}($$

$$\sum_{n:\, y_n=1} log\left(\left\{ p1 \frac{1}{\sqrt{2\pi\sigma_1^2}} exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) \right\} \right) + \sum_{n:\, y_n=2} log\left(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\} \right) = 0$$

This reduces to ;

$$\sigma_2^* = \frac{\partial}{\partial \sigma_2^*} \left(\sum_{n:\, y_n=2} log\left(\left\{ p2 \frac{1}{\sqrt{2\pi\sigma_2^2}} exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \right\} \right) + 0 = 0 \right.$$

$$\sigma_2^* = \frac{\partial}{\partial \sigma_2^*} \left(\sum_{n:\, y_n=2} \left\{ log\{ \frac{1}{\sqrt{2\pi\sigma_2^2}} \} + log\{exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2})\} \right\} = 0 \right)$$

$$\sigma_2^* = \frac{\partial}{\partial \sigma_2^*} \sum_{n:\, y_n=2} \{ \frac{-1}{2} log\sigma_2^2 = \frac{(x_n - \mu_2)^2}{2\sigma_2^2}$$

$$\sigma_2^* = \frac{\partial}{\partial \sigma_2^*} \sum_{n:\, y_n=2} \{ log\sigma_2^2 = \frac{(x_n - \mu_2)^2}{\sigma_2^2} = 0 \}$$

$$\sigma_2^* = 2\frac{n2}{\sigma_2} = \frac{-2(x_n - \mu_2)^2}{\sigma_2^3}$$

$$\sigma_2^* = \frac{\sum(x_n - \mu_2)}{\frac{n1}{\sqrt{n2}}}$$

2)
b)

W.k.t Naive Bayes

$p(x|y = c_1)$ *and* $p(x|y = c_2)$ follows a multivariate gaussian distributions

$N(\mu_1, \sum)$ *and* $N(\mu_2, \sum)$ respectively.

To prove:

$$p(y = 1|x) = \frac{1}{1+exp(-\theta^T x)} \textit{for some } \theta$$

$$P(Y = 1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)+P(X|Y=1)P(Y=1)}$$

$$= \left(\frac{1}{1+\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}}\right)$$

$$= \frac{1}{1+\frac{p_2 \frac{1}{\sqrt{(2\pi)^2|\sum|}} exp\left[\frac{-1}{2}(x-\mu_2)^T(\sum)^{-1}(x-\mu_2)\right]}{p_1 \frac{1}{\sqrt{(2\pi)^2|\sum|}} exp\left[\frac{-1}{2}(x-\mu_1)^T(\sum)^{-1}(x-\mu_1)\right]}}$$

By substituting $p_2 = 1 - p_1$ in the above equation,

$$P(Y = 1|X) = \frac{1}{1+\frac{(1-p_1) \frac{1}{\sqrt{(2\pi)^2|\sum|}} exp\left[\frac{-1}{2}(x-\mu_2)^T(\sum)^{-1}(x-\mu_2)\right]}{p_1 \frac{1}{\sqrt{(2\pi)^2|\sum|}} exp\left[\frac{-1}{2}(x-\mu_1)^T(\sum)^{-1}(x-\mu_1)\right]}}$$

$$P(Y = 1|X) = \frac{1}{1+\frac{(1-p_1)exp\left[\frac{-1}{2}(x-\mu_2)^T(\sum)^{-1}(x-\mu_2)\right]}{p_1 exp\left[\frac{-1}{2}(x-\mu_1)^T(\sum)^{-1}(x-\mu_1)\right]}}$$

Since $x = e^{log(x)}$ *we get*,

$$P(Y = 1|X) = \frac{1}{1+exp\left(log\left(\frac{(1-p_1)exp(+\frac{-1}{2}(x-\mu_2)^T(\sum)^{-1}(x-\mu_2)+\frac{-1}{2}(x-\mu_1)^T(\sum)^{-1}(x-\mu_1))}{(p_1)}\right)\right)}$$

$$P(Y=1|X) = \frac{1}{1+exp\left(log\left(\frac{(1-p_1)}{(p_1)}\right)+\left[\frac{-1}{2}(x-\mu_2)^T(\sum)^{-1}(x-\mu_2)+\frac{-1}{2}(x-\mu_1)^T(\sum)^{-1}(x-\mu_1)\right]\right)}$$

$$P(Y=1|X) = \frac{1}{1+exp\left(log\left(\frac{(1-p_1)}{(p_1)}\right)+\left[(\frac{-1}{2})(-x^T(\sum)^{-1}\mu_2-\mu_2^T(\sum)^{-1}x+\mu_2^T(\sum)^{-1}x+(\frac{1}{2})(-x^T(\sum)^{-1}\mu_1-\mu_1^T(\sum)^{-1}x+\mu_1^T(\sum)^{-1}\mu_1\right]\right)}$$

$$P(Y=1|X) = \frac{1}{1+exp\left(log\left(\frac{(1-p_1)}{(p_1)}\right)+\left[\frac{\mu_1^T(\sum)^{-1}\mu_1-\mu_2^T(\sum)^{-1}\mu_2}{2}\right]-\left[\mu_1^T-\mu_2^T\right](\sum)^{-1}x\right)}$$

$$P(Y=1|X) = \frac{1}{1+exp\left(-\left[-log\left(\frac{(1-p_1)}{(p_1)}\right)-\left[\frac{\mu_1^T(\sum)^{-1}\mu_1-\mu_2^T(\sum)^{-1}\mu_2}{2}\right]+\left[\mu_1^T-\mu_2^T\right](\sum)^{-1}x\right]\right)}$$

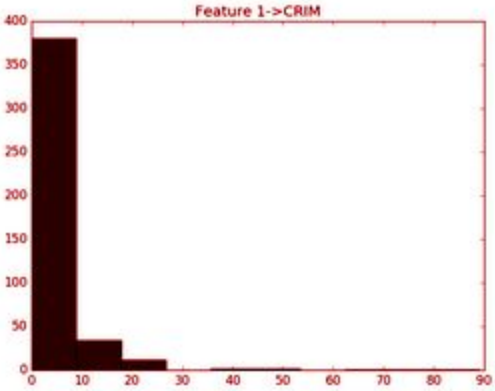
The above equation is of the required form

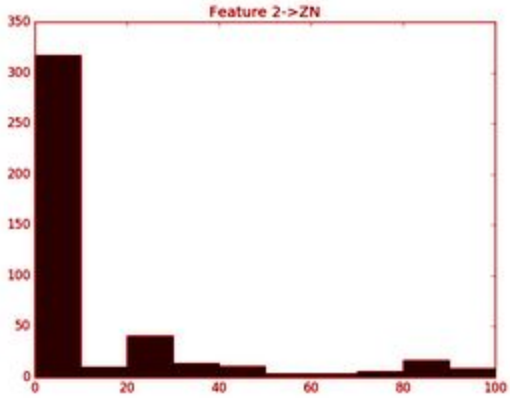
$p(y=1|x) = \frac{1}{1+exp(-\theta^T x)}$

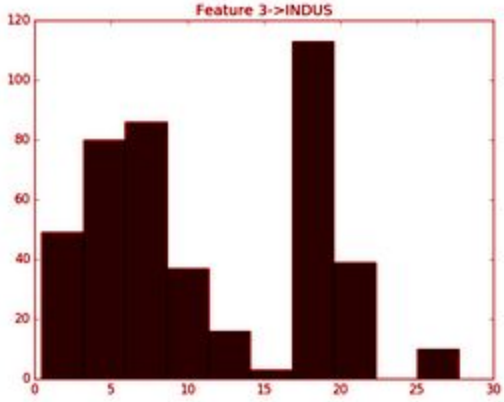
Where,

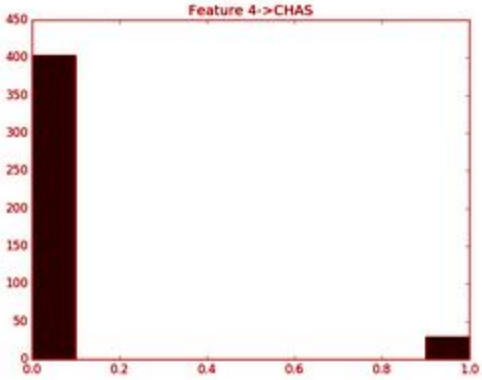
$$\theta^T x = \theta^0 + w^T x = -log\left(\frac{(1-p_1)}{(p_1)}\right) - \left[\frac{\mu_1^T(\sum)^{-1}\mu_1-\mu_2^T(\sum)^{-1}\mu_2}{2}\right] + \left[\mu_1^T-\mu_2^T\right](\sum)^{-1}x$$

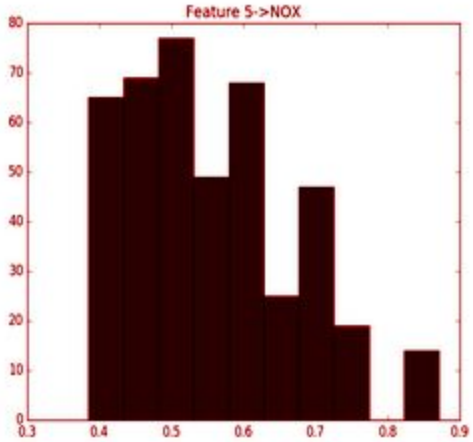
3) Histograms

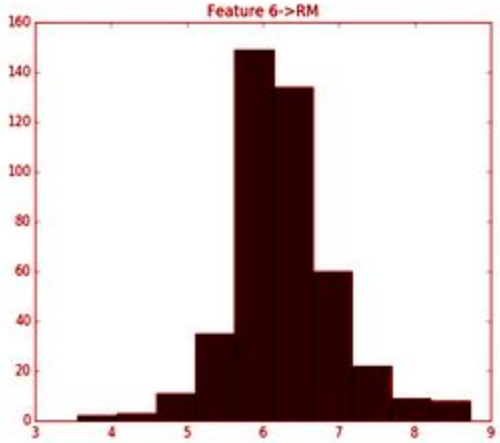

Feature 1->CRIM
400
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90

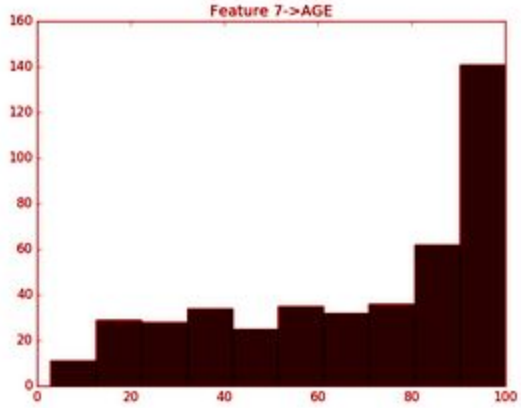

Feature 2->ZN
350
300
250
200
150
100
50
0
0
20
40
60
80
100

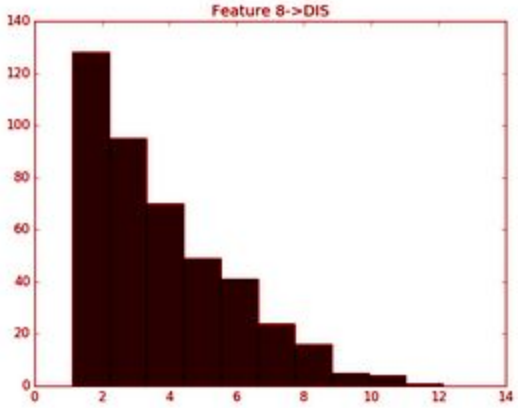

Feature 3->INDUS
120
100
80
60
40
20
0
0
5
10
15
20
25
30

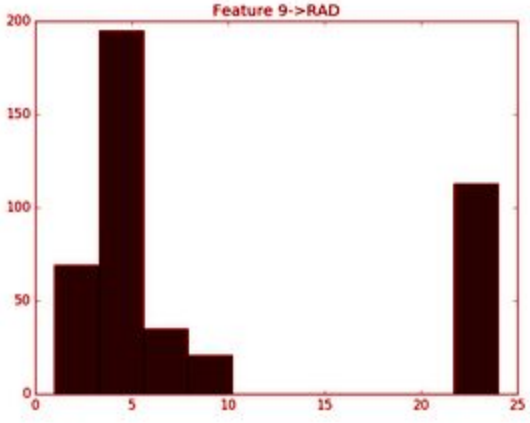

Feature 4->CHAS
450
400
350
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0

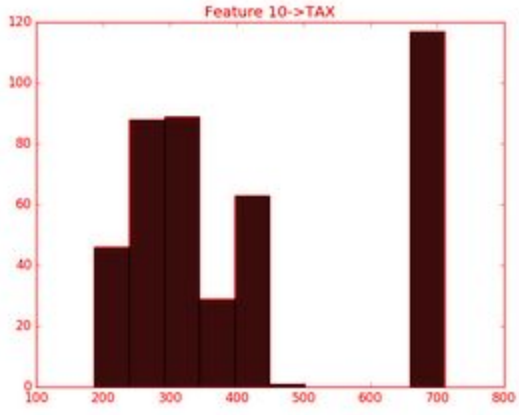

Feature 5->NOX
80
70
60
50
40
30
20
10
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9

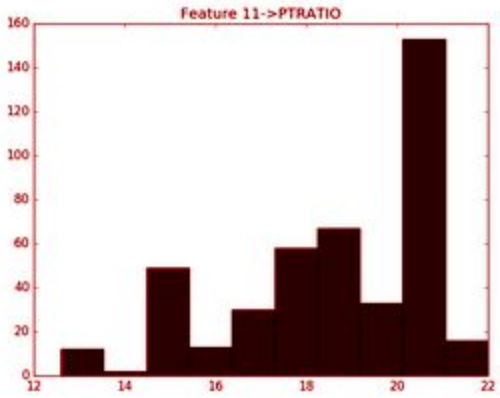

Feature 6->RM
160
140
120
100
80
60
40
20
0
3
4
5
6
7
8
9

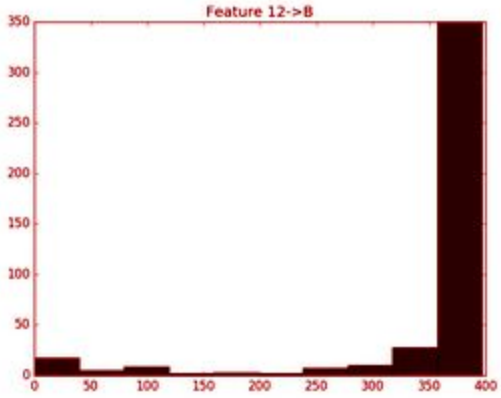

Feature 7->AGE
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100

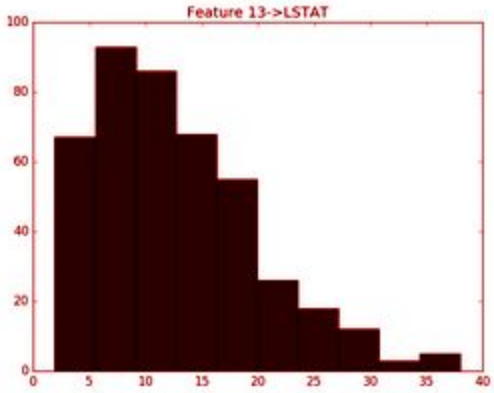

Feature 8->DIS
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14


Feature 9->RAD
200
150
100
50
0
0
5
10
15
20
25


Feature 10->TAX
120
100
80
60
40
20
0
100
200
300
400
500
600
700
800







Pearson Correlation : We can see from the data that the Pearson Correlation is highest for the feature 'LSTAT' which is -0.073

Programming :

1) **Linear Regression and Ridge Regression :**

MSE for Training Data : 20.950144508
MSE for testing Data : 28.4179164975

2) Ridge Regression :
MSE for training Data : 20.9502783985
MSE for testing data : 28.4253883068

2) Ridge Regression with Cross Validation - 10 fold :
lambda = 0.0001 Mean CV MSE : 23.5822099346

lambda = 0.001 Mean CV MSE : 23.5821939596
lambda = 0.01 Mean CV MSE : 23.5820343881
lambda = 0.1 Mean CV MSE : 23.5804563838
lambda = 1 Mean CV MSE : 23.5663434685
lambda = 10 Mean CV MSE : 23.5276152939

Here we can see that as Lambda Value increases the MSE decreases.However after a certain point in the curve even though when the lambda increases, the MSE will not decrease.
For example in my output for lambda value at 100, the MSE is

lambda = 100 Mean CV MSE : 24.7837340983

Hence, we can conclude that that the Lambda lies in the interval [1,10]

Performance :
We can see from our output that the performance of the Linear regression is better than the Ridge Regression in calculating MSE's.

Ridge Regression reduces overfitting and helps in Regularisation when the design matrix has features which are linear combination of each other making the matrix Singular or Non-invertible, however the data set we are provided with does neither has such feature nor does it reduces overfitting hence Ridge Regression does not improve the MSE of Linear Regression.

Feature Selection :

a)
The top 4 features selected are :
[INDUS RM PTRATIO LSTAT]

Train-MSE : 26.4066042155
Test-MSE : 31.4962025449
b)
The top 4 features selected are :
CHAS RM PTRATIO LSTAT

train-MSE : 25.1060222464
test-MSE : 34.6009447175

Selecting one feature at a time improved the MSE

Brute Force :

The top 4 features selected are :
CHAS RM PTRATIO LSTAT

train-MSE : 25.1060222464
test-MSE : 34.6009447175

Brute Force always selects the top 4 features as it exhausts all the combinations and gets the best results.
However Brute force performs solwest as the algorithm is itself brute force.

Feature Expansion :

Train MSE : 5.05978429711
Test MSE : 14.555304

We can see that as we expand the features the number of features, we end up increasing the points in the data space and thus we find better ways to fit the data with the new features which might have not been possible without the expansion of the features.
Thus we can say feature expansion is one to improve the accuracy of our Linear Regression