MACHINE LEARNING
ASSIGNMENT 2
10/03/2016

AJAY KUMAR LOGANATHAN RAVICHANDRAN
USC ID: 1669-4689-06

1) LOGISTIC REGRESSION

a)

Given: Binary logistic regression model with training examples : $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

To find: Negative log likelihood $L(w)$

We know that the probability of a training sample (x_n, y_n) :

$$p(y_n|x_n; b; w) = \sigma(b + w^T x_n)^{y_n}[1 - \sigma(b + w^T x_n)]^{1-y_n}$$

Then, the log likelihood can be shown as:

$$= log\left(\prod_{i=1}^{n} P(Y = y_i|X = x_i\right)$$

$$= \sum_n \{y_n log\sigma(b + w^T x_n) + (1 - y_n) log[1 - \sigma(b + w^T x_n)]\}$$

The negative log likelihood is:

$$L(w) = -\sum_n \{y_n log\sigma(b + w^T x_n) + (1 - y_n) log[1 - \sigma(b + w^T x_n)]\}$$

b)

To find: Update rule for w using Gradient Descent method

$$L(w) = -\sum_n \{y_n log\sigma(b + w^T x_n) + (1 - y_n) log[1 - \sigma(b + w^T x_n)]\}$$

General form for minimizing $L(w)$:

$$w^{t+1} \leftarrow w - \eta \frac{\partial L(w)}{\partial w}$$

$$w^{t+1} \leftarrow w - \eta \left(-\sum_n \{y_n[1 - \sigma(w^T x_n + b)]x_n - (1 - y_n)\sigma(w^T x_n + b)x_n\}\right)$$

$$w^{t+1} \leftarrow w - \eta \left(\sum_n \{\sigma(w^T x_n + b) - y_n\} x_n\right)$$

For the above function to have a global minimum , it should be a convex function, i.e. the second-order derivative of the error loss function should be greater than zero.

$$\frac{\partial^2 L(w)}{\partial w^2} = \frac{\partial}{\partial w}\left(\sum_n \{\sigma(w^T x_n + b) - y_n\} x_n\right)$$

$$= \frac{\partial}{\partial w}\left(\sum_n \left(\frac{1}{1+exp(-(w^T x_n+b))} - y_n\right) x_n\right)$$

$$= \sum_n \left(\frac{exp(-(w^T x_n+b))x_n}{\left(1+exp(-(w^T x_n+b))\right)^2}\right)$$

$$= \sum_n \frac{1}{(1+exp(-(w^T x_n+b)))}\left[1 - \frac{1}{(1+exp(-(w^T x_n+b)))}\right] x_n^2$$

$$= \sum_n \sigma\left(w^T x_n + b\right)\left(1 - \sigma(w^T x_n + b)\right) x_n^2$$

As we know that sigmoid values range between 0 and 1, the above error function will never be negative. Therefore the error function is convex. I.e. It has a global minimum

c)

To find: Negative log likelihood $L(w)$

Given:

Posterior probability for class k :

$$P(Y = k \mid X = x) = \frac{exp(w_k^T x)}{1 + \sum_1^{K-1} exp(w_t^T x)} \quad for\ k = 1, ..., K-1$$

$$P(Y = k \mid X = x) = \frac{1}{1 + \sum_1^{K-1} exp(w_t^T x)} \quad for\ k = K$$

$Since\ w_x^T = 0,$

$$P(Y = k \mid X = x) = \frac{exp(w_k^T x)}{1 + \sum_1^{K-1} exp(w_t^T x)} \quad for\ k = 1, ..., K$$

We consider a K-dimensional vector y_n using 1-of-K encoding, where $y_n = [y_{n1}, y_{n2}, ..., y_{nk}]^T$ and $y_{nk} = 1\ if\ y_n = K,\ 0\ otherwise.$

Thus,

Let $w \rightarrow w_1, w_2, ..., w_k$,

$$L(w) = -\prod_n P(Y_n \mid X_n)$$

$$L(w) = -log\left[\prod_n \prod_k [P(Y = k|x_n)]^{y_{nk}}\right]$$

$$= -\sum_n \sum_k y_{nk} log \frac{exp(w_k^T x_n)}{1 + \sum_{t=1}^{K-1} exp(w_k^T x_n)}$$

$$= -\sum_n \sum_k y_{nk} \left[\left(w_k^T x_n\right) - log\left(1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)\right)\right]$$

d)

To find: $\frac{\partial L(w_1, w_2, ..., w_k)}{\partial w_i}$

$$\frac{\partial L(w_1, w_2, ..., w_k)}{\partial w_i} = \frac{\partial}{\partial w_i}\left[-\sum_n \sum_k y_{nk}\left[\left(w_k^T x_n\right) - log\left(1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)\right)\right]\right]$$

$$= \frac{\partial}{\partial w_i}\left(-\sum_n\left(\left[\sum_{k=i} y_{nk}\left(\left(w_k^T x_n\right) - log\left(1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)\right)\right)\right] + \left[\sum_{k \neq i} y_{nk}\left(\left(w_k^T x_n\right) - log\left(1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)\right)\right)\right]\right)\right.$$

$$= -\sum_n \sum_{k=i} y_{nk} x_n - \frac{exp\left(w_i^T x_n\right) x_n y_{nk}}{1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)} + \sum_{k \neq i} - \frac{exp\left(w_i^T x_n\right) x_n y_{nk}}{1 + \sum_{t=1}^{K-1} exp\left(w_t^T x_n\right)}$$

$$= -\sum_n \left[y_{nk} x_n - \sum_k \frac{x_n y_{nk} exp(w_i^T x_n)}{1 + \sum_{t=1}^{K-1} exp(w_t^T x_n)}\right]$$

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

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

2) Linear/Gaussian Discriminant

a)

Given:

Training examples $D = \{(x_n,y_n)\}_{n=1}^N,\ y_n \in \{1,2\}$

$$p(x_n,y_n) = p(y_n)p(x_n|y_n) = \frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right) \quad if\ y_n = 1$$
$$= \frac{p_2}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right) \quad if\ y_n = 2$$

To find:

Log likelihood function $L(D)$ and MLE to find $(p_1^*, p_2^*, \mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*)$ that maximizes $L(D)$

Solution:

$$P(D) = \sum_n P(x_n,y_n)$$

$$logP(D) = \sum_n logP(x_n,y_n)$$

$$L(D) = logP(D)$$
$$= \sum_n logP(x_n,y_n)$$
$$= \sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} log\left(\frac{p_2}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

MLE:

To Estimate p_1^*

$$\frac{\partial L}{\partial p_1} = \frac{\partial}{\partial p_1}\left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)\right]$$
$$= \sum_{n:y_n=1} \frac{\partial}{\partial p_1}\left(log(p_1) - log(\sqrt{2\pi}\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} \frac{\partial}{\partial p_1}\left(log(1-p_1) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)$$
$$\Rightarrow \sum_{n:y_n=1}\left(\frac{1}{p_1}\right) + \sum_{n:y_n=2}\left(\frac{-1}{1-p_1}\right) = 0$$
$$\Rightarrow \frac{n_1}{p_1} = \frac{n_2}{1-p_1}$$
$$\Rightarrow \frac{1}{p_1} = 1 + \frac{n_2}{n_1}$$

$$\Rightarrow p_1 = \frac{n_1}{n_1+n_2}$$

$$\Rightarrow p_1 = \frac{n_1}{n} \quad where\ n = n_1 + n_2$$

To Estimate p_2^*

$$\frac{\partial L}{\partial p_2} = \frac{\partial}{\partial p_2}\left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)\right]$$

$$= \sum_{n:y_n=1} \frac{\partial}{\partial p_2}\left(log(1-p_2) - log(\sqrt{2\pi}\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} \frac{\partial}{\partial p_2}\left(log(p_2) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

$$\Rightarrow \sum_{n:y_n=1}\left(\frac{-1}{1-p_2}\right) + \sum_{n:y_n=2}\left(\frac{1}{p_2}\right) = 0$$

$$\Rightarrow \frac{n_1}{1-p_2} = \frac{n_2}{p_2}$$

$$\Rightarrow \frac{1}{p_2} = 1 + \frac{n_1}{n_2}$$

$$\Rightarrow p_2 = \frac{n_2}{n_2+n_1}$$

$$\Rightarrow p_2 = \frac{n_2}{n} \ where\ n = n_1 + n_2$$

To Estimate μ_1^*:

$$\frac{\partial L}{\partial \mu_1} = \frac{\partial}{\partial \mu_1}\left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)\right]$$

$$= \sum_{n:y_n=1} \frac{\partial}{\partial \mu_1}\left(log(1-p_2) - log(\sqrt{2\pi}_1\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} \frac{\partial}{\partial \mu_1}\left(log(p_2) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

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

$$\Rightarrow \left(\frac{1}{\sigma_1^2}\left(\sum_{n:y_n=1} x_n - n_1\mu_1\right)\right) = 0$$

$$\Rightarrow \sum_{n:y_n=1} x_n = n_1\mu_1$$

$$\Rightarrow \mu_1 = \frac{\sum_{n:y_n=1} x_n}{n_1}$$

To Estimate μ_2^*:

$$\frac{\partial L}{\partial \mu_2} = \frac{\partial}{\partial \mu_2}\left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)\right]$$

$$= \sum_{n:y_n=1} \frac{\partial}{\partial \mu_2}\left(log(1-p_2) - log(\sqrt{2\pi}_1\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2}\right)\right) + \sum_{n:y_n=2} \frac{\partial}{\partial \mu_2}\left(log(p_2) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

$$\Rightarrow \sum_{n:y_n=2} -\frac{-2(x_n-\mu_2)^2}{2\sigma_2^2} = 0$$

$$\Rightarrow \left(\frac{1}{\sigma_2^2} \left(\sum_{n:y_n=1} x_n - n_2\mu_2 \right) \right) = 0$$

$$\Rightarrow \sum_{n:y_n=2} x_n = n_2\mu_2$$

$$\Rightarrow \mu_2 = \frac{\sum_{n:y_n=1} x_n}{n_2}$$

To Estimate σ_1^*:

$$\frac{\partial L}{\partial \sigma_1} = \frac{\partial}{\partial \sigma_1} \left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2} \right) \right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2} \right) \right) \right]$$

$$= \sum_{n:y_n=1} \frac{\partial}{\partial \sigma_1} \left(log(1-p_2) - log(\sqrt{2\pi}_1\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2} \right) \right) + \sum_{n:y_n=2} \frac{\partial}{\partial \sigma_1} \left(log(p_2) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2} \right) \right)$$

$$\Rightarrow \sum_{n:y_n=1} \left(-\frac{1}{\sigma_1} - \frac{(x_n-\mu_1)^2}{\sigma_1^3} \right) = 0$$

$$\Rightarrow \left(-\frac{n_1}{\sigma_1} - \sum_{n:y_n=1} \frac{(x_n-\mu_1)^2}{\sigma_1^3} \right) = 0$$

$$\Rightarrow \sum_{n:y_n=1} (x_n-\mu_1)^2 = n_1\sigma_1^2$$

$$\Rightarrow \sigma_1 = \sqrt{\frac{\sum_{n:y_n=1} (x_n-\mu_1)^2}{n_1}}$$

To Estimate σ_2^*:

$$\frac{\partial L}{\partial \sigma_2} = \frac{\partial}{\partial \sigma_2} \left[\sum_{n:y_n=1} log\left(\frac{p_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x_n-\mu_1)^2}{2\sigma_1^2} \right) \right) + \sum_{n:y_n=2} log\left(\frac{1-p_1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x_n-\mu_2)^2}{2\sigma_2^2} \right) \right) \right]$$

$$= \sum_{n:y_n=1} \frac{\partial}{\partial \sigma_2} \left(log(1-p_2) - log(\sqrt{2\pi}_1\sigma_1) - \left(\frac{(x_n-\mu_1)^2}{2\sigma_1^2} \right) \right) + \sum_{n:y_n=2} \frac{\partial}{\partial \sigma_2} \left(log(p_2) - log(\sqrt{2\pi}\sigma_2) - \left(\frac{(x_n-\mu_2)^2}{2\sigma_2^2} \right) \right)$$

$$\Rightarrow \sum_{n:y_n=2} \left(-\frac{1}{\sigma_2} - \frac{(x_n-\mu_2)^2}{\sigma_2^3} \right) = 0$$

$$\Rightarrow \left(-\frac{n_2}{\sigma_2} - \sum_{n:y_n=1} \frac{(x_n-\mu_2)^2}{\sigma_2^3} \right) = 0$$

$$\Rightarrow \sum_{n:y_n=2} (x_n-\mu_2)^2 = n_1\sigma_2^2$$

$$\Rightarrow \sigma_2 = \sqrt{\frac{\sum_{n:y_n=1} (x_n-\mu_2)^2}{n_2}}$$

b)

Given:

$p(x|y = c_1)$ *and* $p(x|y = c_2)$ follows a multivariate gaussian distributions $N(\mu_1, \sum)$ *and* $N(\mu_2, \sum)$ respectively.

To prove:

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

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

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

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

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

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

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

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)

Program output:

!!!---------- Pearson Correlation ---------!!!

CRIM	=	-0.387696987621
ZN	=	0.362987295831
INDUS	=	-0.483067421758
CHAS	=	0.203600144696
NOX	=	-0.424829675619
RM	=	0.690923334973
AGE	=	-0.390179110401
DIS	=	0.252420566225
RAD	=	-0.385491814423
TAX	=	-0.468849385373
PTRATIO	=	-0.505270756892
B	=	0.343434137151
LSTAT	=	-0.73996982063

!!!---------- Linear Regression ---------!!!

MSE for train data	=	20.950144508
MSE for test data	=	28.4179164975

!!!---------- Ridge Regression ---------!!!

Lambda	=	0.0001
MSE for train data	=	20.950144508
MSE for test data	=	28.4179216831
Lambda	=	0.001
MSE for train data	=	20.9501445092
MSE for test data	=	28.4179683525
Lambda	=	0.01
MSE for train data	=	20.950144631
MSE for test data	=	28.4184349972
Lambda	=	0.1

MSE for train data = 20.950156753
MSE for test data = 28.4230964881
Lambda = 1
MSE for train data = 20.9513128127
MSE for test data = 28.4692071131
Lambda = 10
MSE for train data = 21.0296877783
MSE for test data = 28.8789028004

!!!---------- Ridge Regression : Cross Validation ---------!!!
Lambda = 0.0001
MSE for test data = 23.6431546149
Lambda = 0.001
MSE for test data = 23.643135083
Lambda = 0.01
MSE for test data = 23.6429399612
Lambda = 0.1
MSE for test data = 23.6410083486
Lambda = 1
MSE for test data = 23.6235468906
Lambda = 10
MSE for test data = 23.5669909329

Minimum value of lambda = 7.908244
MSE for training set = 23.3649453604
MSE on test data = 28.7916794612

!!!---------- FEATURE SELECTION - A ---------!!!
Highed correlated features - Top 4 = LSTAT,RM,PTRATIO,INDUS
MSE for train data = 26.4066042155
MSE for test data = 31.4962025449

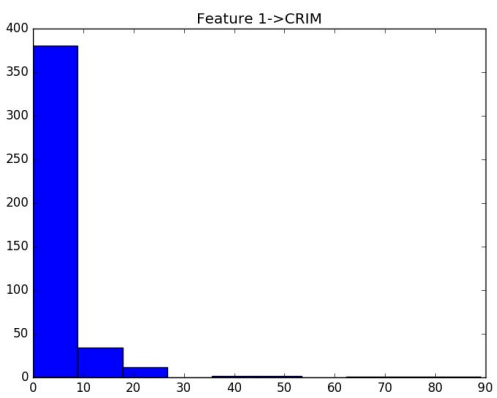
!!!---------- FEATURE SELECTION - B ---------!!!
Highly correlated features = LSTAT,RM,PTRATIO,CHAS
MSE for train data = 25.1060222464
MSE for test data = 34.6000723135

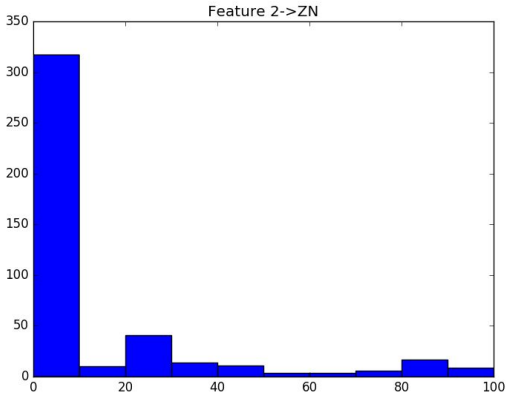
!!!---------- FEATURE SELECTION - C ---------!!!
Brute force search - Top 4 = CHAS,RM,PTRATIO,LSTAT
MSE for train data = 25.1060222464
MSE for test data = 34.6000723135

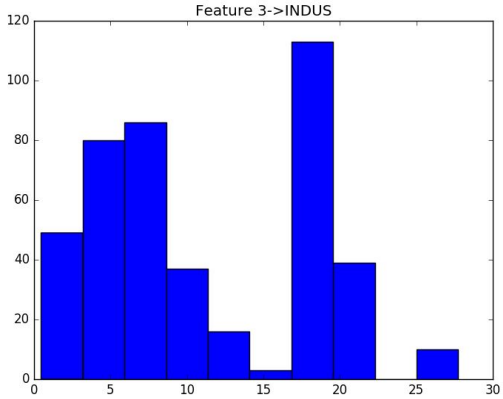
!!!---------- FEATURE EXPANSION ---------!!!
MSE for train data = 5.05978429711
MSE for test data = 14.5553049724

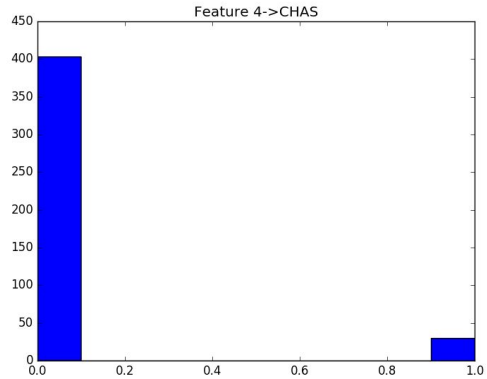
HISTOGRAM PLOTS:

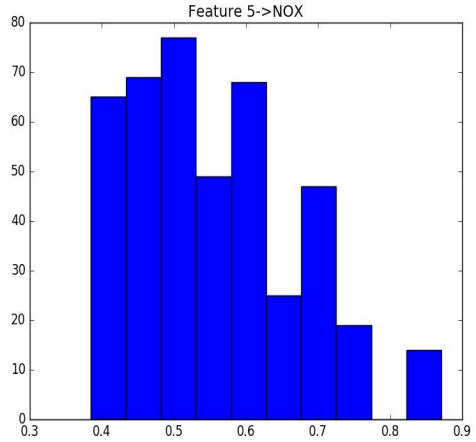
The histograms are plotted with the frequency of the training data the the total number of bins as 10.

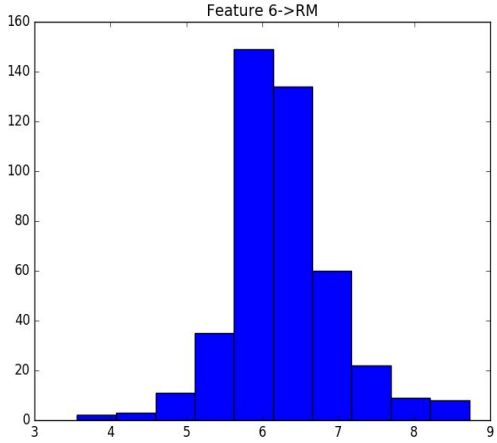


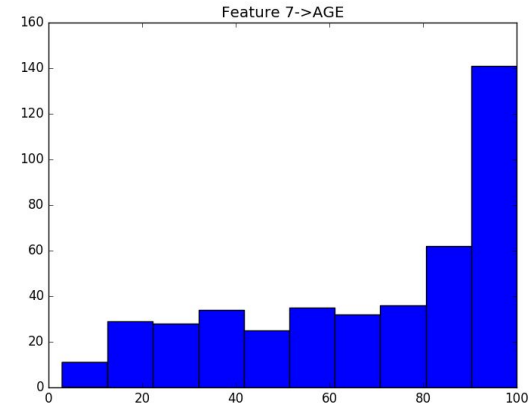


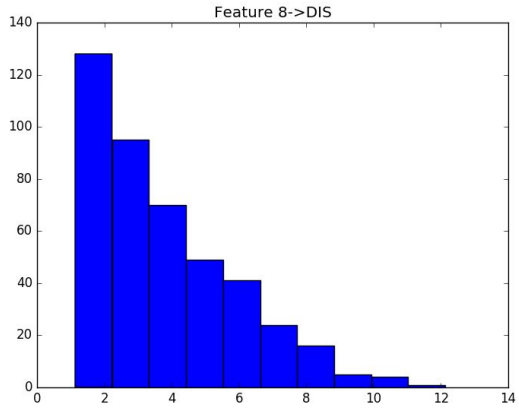


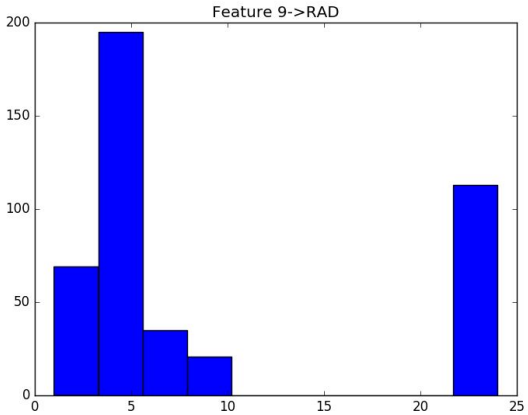


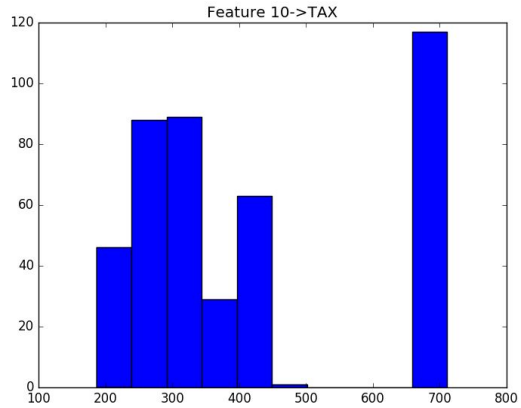


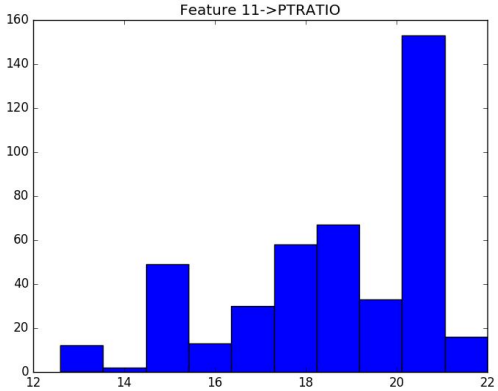


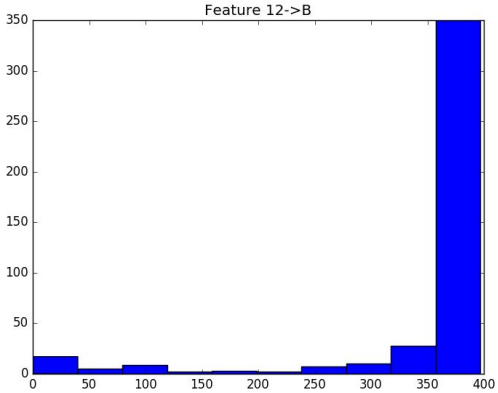

Feature 7->AGE
0
20
40
60
80
100
0
20
40
60
80
100
120
140
160

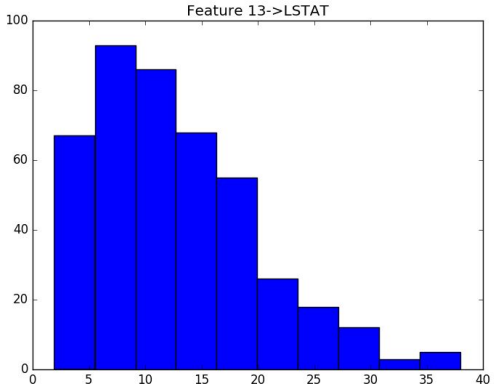

Feature 8->DIS
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120
140


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


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


Feature 11->PTRATIO
12
14
16
18
20
22
0
20
40
60
80
100
120
140
160


Feature 12->B
0
50
100
150
200
250
300
350
400
0
50
100
150
200
250
300
350



OBSERVATION:

1. **Pearson correlation:**
    a. Pearson correlation is calculated for all the features with the target Y
    b. Highly correlated feature is **LSTAT** with negative correlation
    c. Least correlated feature is **CHAS** with positive correlation
    d. Five features are positively correlated and eight features are negatively correlated.

2. **Linear regression and Ridge regression**
    a. In our dataset, We can observe that the linear regression performs better than ridge regression by looking at the values of MSE.
    b. This shows that either features in our dataset are not highly correlated and thus there isn't exist any overfitting or that the data samples are very less to show any relation between the features.
    c. The Ridge regression is consistent with the linear regression when the shrinkage parameter (λ) is almost zero,i.e the MSE is almost equal to the MSE of Linear regression
    d. When the shrinkage parameter increases, the MSE also increases.

3. **Ridge regression with Cross validation**
    a. We observe that the MSE decreases as λ increases.
    b. MSE is minimum when $\lambda = 10$
    c. Thus, we computed the MSE for the λ between 1 and 10, and found that the global minimum is between 7 and 8.

4. **Feature Selection**
    a. We notice that highly correlated features fetched using pearson correlation doesn't hold good with the ones selected with the brute force approach for the training data.

b. However, it performs better than the residual or brute force approach with the test data. The reason could be that it is dependent on the covariance between the data and the target and thus computes the linear dependence between the two data whereas the other two methods are dependent on the features which minimizes the MSE for train data which might not hold good for the unknown data samples.
c. Residual method provides the same results as the brute force method as they try to select the features which reduces the the MSE while training the samples.

5. **Feature Expansion**
    a. The less MSE after expansion of the features shows that the polynomial expansion helps to better fit the data than the linear or ridge regression.

COLLABORATORS:

SINDHUJHA SETHURAMAN,
RAJBHARATH RAJENDRAN