Machine Learning Assignment 2

Anirudh Mukund Deshpande
USC ID: 2746365653
deshpana@usc.edu

September 2016

1 Logistic Regression

(1.a) **Answer:**

The formula for log likelihood as a loss function is -

$$E(w) = -ln(\Pi_{i=1}^{N} P(Y = y_i | X = x_i))$$

Using the binary logistic regression model, it can be written as,

$$\begin{aligned} E(w) &= -ln(\Pi_{i=1}^{N}[y_i ln[\sigma(\omega^T x_i)] + (1 - y_i) ln[1 - \sigma(\omega^T x_i)]) \\ &= -\sum_{i=1}^{N}[y_i ln[\sigma(\omega^T x_i)] + (1 - y_i) ln[1 - \sigma(\omega^T x_i)]] \end{aligned}$$

(1.b) **Answer:**

From (1.a), we have -

$$\varepsilon(w) = -\sum_{i=1}^{N}[y_i ln[\sigma(\omega^T x_i)] + (1 - y_i) ln[1 - \sigma(\omega^T x_i)]]$$

Taking the partial derivative with respect to w, we have -

$$\frac{\partial(\varepsilon(w))}{\partial(w)} = -\sum_{i=1}^{N}[y_i(1 - \sigma(\omega^T x_i)) - (1 - y^i)\frac{\sigma(\omega^T x_i)(1 - \sigma(\omega^T x_i))}{1 - \sigma(\omega^T x_i)}]x_i^T$$

Simplifying, we have -

$$\frac{\partial(\varepsilon(w))}{\partial(w)} = -\sum_{i=1}^{N}[y_i(1 - \sigma(\omega^T x_i)) - (1 - y_i)\sigma(\omega^T x_i)]x_i^T$$

Further simplifying, we have -

$$\begin{aligned} \frac{\partial(\varepsilon(w))}{\partial(w)} &= -\sum_{i=1}^{N}[y_i(1 - \sigma(\omega^T x_i)) - (1 - y_i)\sigma(\omega^T x_i)]x_i^T \\ &= -\sum_{i=1}^{N}[y_i - y_i\sigma(\omega^T x_i) - \sigma(\omega^T x_i) + y_i\sigma(\omega^T x_i)]x_i^T \\ &= -\sum_{i=1}^{N}[y_i - \sigma(\omega^T x_i)]x_i^T \\ &= \sum_{i=1}^{N}[\sigma(\omega^T x_i) - y_i]x_i^T \end{aligned}$$

The updated rule for w using Gradient Descent method is,

$$\omega^{T+1} = \omega^T - \eta[\sum_{i=1}^{N}[\sigma(\omega^T x_i) - y_i]x_i^T]$$

Since the expression $[\sigma(\omega^T x_i) - y_i]x_i^T$ represents a scalar, we know that transpose of a scalar is a scalar quantity.
From the properties of vectors (considering C a scalar), we have

$$Y^T = CX^T$$
$$(Y^T)^T = (CX^T)^T$$
$$Y = XC^T$$
$$C^T = C$$
$$Hence Y = XC$$

Hence, $[\sigma(\omega^T x_i) - y_i]x_i^T = x_i[\sigma(\omega^T x_i) - y_i]$
Using the above rules, the updated value of w can be written as -

$$\omega^{T+1} = \omega^T - \eta[\sum_{i=1}^{N} x_i[\sigma(\omega^T x_i) - y_i]]$$

$$\frac{\partial(\varepsilon(w))}{\partial(w)} = \sum_{i=1}^{N} x_i[\sigma(\omega^T x_i) - y_i]$$

Taking the second derivative,

$$\frac{\partial^2(\varepsilon(w))}{\partial(w)^2} = \sum_{i=1}^{N} x_i[\sigma(\omega^T x_i)(1 - \sigma(\omega^T x_i))]x_i^T$$

If we consider the matrix formed by multiplying $x_i * x_i^T$, the diagonal elements represent the second order derivative with respect to each 'w', the diagonal elements become positive. Also, the sigmoid function always gives a positive value.
Hence the function converges to a global minimum. With a proper value of η, we can find the minimum value of ω

We can also prove that loss function E(w) is a convex function as follows.
The loss function is given by,

$$\varepsilon(w) = -\sum_{i=1}^{N}[y_i ln[\sigma(\omega^T x_i)] + (1 - y^i) ln[1 - \sigma(\omega^T x_i)]]$$

Let us consider $f(w) = \sigma(w^T x)$, we have,

$$\frac{\partial(-lnf(w))}{\partial(w)} = (f(w) - 1)x$$
$$\frac{\partial^2(-lnf(w))}{\partial(ww^T)} = \frac{\partial(-lnf(w))}{\partial(w)}((f(w) - 1)x)$$
$$= f(w)(1 - f(w))xx^T$$

The Hessian matrix for any vector v could be written as,

$$v^T \frac{\partial^2(-lnf(w))}{\partial(ww^T)} v = v^T(f(w)(1 - f(w))xx^T)v$$
$$= f(w)(1 - f(w)((x^T)^2 v^2$$

Since $f(w)(1 - f(w))$ and $(x^T)^2 v^2$ is greater than or equal to zero, the above equation is greater than or equal to zero.
Now, taking the second term,

$$\frac{\partial(-ln(1 - f(w))}{\partial(w)} = \frac{\partial(-ln(1 - f(w))}{\partial(w)}(w^T x + ln(1 + exp(-w^T x))$$
$$= x + \frac{\partial(-ln(1 - f(w))}{\partial(w)}(ln(1 + exp(-w^T x))$$
$$\frac{\partial^2 -(ln(1 - f(w))}{\partial(ww^T)} = \frac{\partial(-log(1 - f(w))}{\partial(w)}(\frac{\partial(-log(1 - f(w))}{\partial(w)}(-log(1 - f(w))))$$
$$= \frac{\partial(-log(1 - f(w))}{\partial(w)}(x + \frac{\partial(-log(1 - f(w))}{\partial(w)}(log(1 + exp(-w^T x)))$$
$$= \frac{\partial^2(-log(f(w))}{\partial(ww^T)}$$

We have proved that $\frac{\partial^2(-log(f(w))}{\partial(ww^T)}$ >= 0 . Hence, $\frac{\partial(-ln(1-f(w))}{\partial(w)}$ >= 0.
Since both the terms are convex, the equation is convex. Hence we can prove that Gradient Descent will converge to a global minimum.

(1.c) **Answer:** If w_k is zero, then we can write the below formula

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

Considering $y_{ik} = 1$ if $y_i = k$, else 0, The likelihood is written as -

$$L(w_1..w_k) = \prod_{i=1}^{n}\prod_{k=1}^{K} P(Y=k|X=x_i)^{y_{ik}}$$

The negative log likelihood is -

$$l(w_1..w_k) = -ln(\prod_{i=1}^{n}\prod_{k=1}^{K} P(Y=k|X=x_i)^{y_{ik}})$$
$$= -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} * ln(P(Y=k|X=x_i))$$
$$= -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} * ln(\frac{\exp(w_k^T x_i)}{\sum_1^K \exp(w_t^T x_i)})$$
$$= -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(w_k^T x_i - ln\sum_{i=1}^{K}\exp(w_t^T x_i))$$

(1.d) **Answer:** If we differentiate negative log likelihood w. r. to. w_j, we get,

$$\frac{\partial l}{\partial w_j} = -\sum_{i=1}^{n} y_{ij}(x_i^T - \frac{exp(w_j^T x_i)}{\sum_{i=1}^{K} exp(w_t^T x_i)} x_i^T)$$
$$= -\sum_{i=1}^{n} x_i y_{ij}(1 - \frac{exp(w_j^T x_i)}{\sum_{i=1}^{K} exp(w_t^T x_i)})$$

We have the updated value of gradient descent as -

$$w_j^{t+1} = w_j^t - \eta(-\sum_{i=1}^{n} x_i y_{ij}(1 - \frac{exp(w_j^T x_i)}{\sum_{i=1}^{K} exp(w_t^T x_i)}))$$
$$= w_j^t + \eta\sum_{i=1}^{n} x_i y_{ij}(1 - \frac{exp(w_j^T x_i)}{\sum_{i=1}^{K} exp(w_t^T x_i)})$$

2 Linear/Gaussian Discriminant

(2.a) **Answer**

The likelihood function could be written as -

$$L(x,y) = \prod_{i=1}^{N} P(x_i, y_i)$$
$$= \prod_{c}\prod_{i;y_i=c} P_c(x_i,y_i)^{(2-y_i)} P_c(x_i,y_i)^{(y_i-1)}$$

Taking log on both the sides, we have -

$$\begin{aligned}
ln(L(x,y)) &= \prod_c \prod_{i;y_i=c} (2-y_i)ln(P_c(x_i,y_i)) + (y_i-1)ln(P_c(x_i,y_i)) \\
&= \prod_c \prod_{i;y_i=c} 2ln(P_c(x_i,y_i)) - y_i ln(P_c(x_i,y_i))) + y_i ln(P_c(x_i,y_i) - ln(P_c(x_i,y_i)) \\
&= \prod_c \prod_{i;y_i=c} ln(P_c(x_i,y_i))
\end{aligned}$$

On further simplification, we have -

$$\begin{aligned}
ln(L(x,y)) = l &= \sum_c \sum_{i;y_i=c} ln(P_c(x_i,y_i)) \\
&= \sum_c \sum_{i;y_i=c} ln[p_c \frac{1}{\sqrt{2\Pi}\sigma_c} \exp^{-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}}] \\
&= \sum_c \sum_{i;y_i=c} ln(p_c) - \frac{1}{2} ln(2\Pi) - ln(\sigma_c) - \frac{(x_i-\mu_c)^2}{2\sigma_c^2}] \\
&= \sum_c \sum_{i;y_i=c} ln(p_c) - \frac{1}{2} ln(2\Pi) - ln(\sigma_c) - \frac{(x_i-\mu_c)^2}{2\sigma_c^2}]
\end{aligned}$$

Since this is a binary classifier, Expanding on each class -

$$\begin{aligned}
l &= \sum_c N_c ln(p_c) - \frac{N}{2} ln(2\Pi) - \sum_c N_c ln(\sigma_c) - \sum_c \sum_{i;y_i=c} \frac{(x_i-\mu_c)^2}{2\sigma^2}] \\
&= N_1 ln(p_1) + N_2 ln(p_2) - \frac{N}{2} ln(2\Pi) - N_1 ln(\sigma_1) - N_2 ln(\sigma_2) - \sum_{i;y_i=1} \frac{(x_i-\mu_1)^2}{2\sigma_1^2} - \sum_{i;y_i=2} \frac{(x_i-\mu_2)^2}{2\sigma_2^2}]
\end{aligned}$$

Where N_1 is number of features corresponding to class 1, N_2 corresponding to class 2, $p_1 + p_2 = 1$ Estimating p_1^*

$$\begin{aligned}
\frac{\partial l}{\partial(p_1)} &= \frac{N_1}{p_1} - \frac{N_2}{1-p_1} = 0 \\
&\Rightarrow \frac{N_1}{p_1} = \frac{N_2}{1-p_1} \\
&\Rightarrow N_1(1-p_1) = N_2 p_1 \\
&\Rightarrow (N_1+N_2)p_1 = N_1 \\
&\Rightarrow p_1 = \frac{N_1}{(N_1+N_2)} \\
&\Rightarrow p_1 = \frac{N_1}{N}
\end{aligned}$$

Hence we have $p_1^* = \frac{N_1}{N}$.
Estimating p_2^*

$$\begin{aligned}
\frac{\partial l}{\partial(p_2)} &= \frac{N_1}{1-p_2} - \frac{N_2}{p_2} = 0 \\
&\Rightarrow \frac{N_1}{1-p_2} = \frac{N_2}{p_2} \\
&\Rightarrow N_2(1-p_2) = N_1 p_2 \\
&\Rightarrow (N_1+N_2)p_2 = N_2 \\
&\Rightarrow p_2 = \frac{N_2}{(N_1+N_2)} \\
&\Rightarrow p_2 = \frac{N_2}{N}
\end{aligned}$$

Hence we have $p_2^* = \frac{N_2}{N}$.

Estimating μ_1^*

$$\frac{\partial l}{\mu_1} = 0$$

$$\Rightarrow \sum_{i;y_i=1} \frac{-2(x_i - \mu_1)(-1)}{2\sigma_1^2} = 0$$

$$\Rightarrow \sum_{i;y_i=1} \frac{-2(x_i - \mu_1)(-1)}{2\sigma_1^2} = 0$$

$$\Rightarrow \sum_{i;y_i=1} x_i - \sum_{i:y_i=1} \mu_1 = 0$$

$$\Rightarrow \sum_{i;y_i=1} x_i - N_1\mu_1 = 0$$

$$\Rightarrow \mu_1 = \frac{\sum_{i;y_i=1} x_i}{N_1}$$

Hence we have $\mu_1^* = \frac{\sum_{i;y_i=1} x_i}{N_1}$.
Estimating μ_2^*

$$\frac{\partial l}{\mu_2} = 0$$

$$\Rightarrow \sum_{i;y_i=2} \frac{-2(x_i - \mu_2)(-1)}{2\sigma_1^2} = 0$$

$$\Rightarrow \sum_{i;y_i=2} \frac{-2(x_i - \mu_2)(-1)}{2\sigma_1^2} = 0$$

$$\Rightarrow \sum_{i;y_i=2} x_i - \sum_{i:y_i=2} \mu_2 = 0$$

$$\Rightarrow \sum_{i;y_i=2} x_i - N_2\mu_2 = 0$$

$$\Rightarrow \mu_2 = \frac{\sum_{i;y_i=2} x_i}{N_2}$$

Hence we have $\mu_2^* = \frac{\sum_{i;y_i=2} x_i}{N_2}$.
Estimating σ_2^*

$$\frac{\partial l}{\partial \sigma_1} = 0$$

$$\Rightarrow -\sum_{i;y_i=1} \frac{N_1}{\sigma_1} + \sum_{i:y_i=1} \frac{(x_i - \mu_1)^2}{\sigma_1^3} = 0$$

$$\Rightarrow \sum_{i;y_i=1} \frac{N_1}{\sigma_1} = \sum_{i:y_i=1} \frac{(x_i - \mu_1)^2}{\sigma_1^3}$$

$$\Rightarrow \sum_{i;y_i=1} N_1 = \sum_{i:y_i=1} \frac{(x_i - \mu_1)^2}{\sigma_1^2}$$

$$\Rightarrow \sigma_1^2 \sum_{i;y_i=1} N_1 = \sum_{i:y_i=1} (x_i - \mu_1)^2$$

$$\Rightarrow \sigma_1^2 = \sum_{i:y_i=1} \frac{(x_i - \mu_1)^2}{N_1}$$

$$\Rightarrow \sigma_1 = \sqrt{\sum_{i:y_i=1} \frac{(x_i - \mu_1)^2}{N_1}}$$

Hence we have $\sigma_1^* = \sqrt{\sum_{i:y_i=1} \frac{(x_i-\mu_1)^2}{N_1}}$.

Estimating σ_2^*

$$\begin{aligned}
&\frac{\partial l}{\partial \sigma_2} = 0 \\
&\quad \Rightarrow -\sum_{i;y_i=2} \frac{N_2}{\sigma_2} + \sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{\sigma_2^3} = 0 \\
&\quad \Rightarrow \sum_{i;y_i=2} \frac{N_2}{\sigma_2} = \sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{\sigma_2^3} \\
&\quad \Rightarrow \sum_{i;y_i=2} N_2 = \sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{\sigma_2^2} \\
&\quad \Rightarrow \sigma_2^2 \sum_{i;y_i=2} N_2 = \sum_{i:y_i=2} (x_i-\mu_2)^2 \\
&\quad \Rightarrow \sigma_2^2 = \sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{N_2} \\
&\quad \Rightarrow \sigma_2 = \sqrt{\sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{N_2}}
\end{aligned}$$

Hence we have $\sigma_2^* = \sqrt{\sum_{i:y_i=2} \frac{(x_i-\mu_2)^2}{N_2}}$.

(2.b) **Answer**

We have the Bayes rule as -

$$\begin{aligned}
P(Y|X) &= \frac{P(X|Y=c_1)P(y=c_1)}{P(X|Y=c_1)P(y=c_1) + P(X|Y=c_2)P(y=c_2)} \\
&= \frac{1}{1 + \frac{P(X|Y=c_2)P(y=c_2)}{P(X|Y=c_1)P(y=c_1)}}
\end{aligned}$$

Assuming $P(y=c_1) = \pi$, since it is a Binary classifier we have $P(y=c_2) = 1-\pi$

$$P(Y|X) = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{P(X|Y=c_2)}{P(X|Y=c_1)}}$$

We have formula for Multivariate Gaussian distribution as -

$$\begin{aligned}
P(X|Y=c) &= \eta(\mu, \Sigma) \\
&= \frac{1}{\sqrt{(2\ \pi)^k |\Sigma|}} \exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)]
\end{aligned}$$

Substituting the above expression in the simplified Bayes rule expression, we have,

$$\begin{aligned}
P(Y|X) = &= \frac{1}{1 + \frac{1-\pi}{\pi} \frac{\exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}{\exp[-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)]}} \\
&= \frac{1}{1 + \frac{1-\pi}{\pi} \exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)] - (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)]}
\end{aligned}$$

Considering and the expression,

$$\begin{aligned}
&(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma^{-1}(x-\mu_2) \\
&\Rightarrow (x^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})x - (x^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})\mu_1 - (x^T\Sigma^{-1} + \mu_2^T\Sigma^{-1})x + (x^T\Sigma^{-1} + \mu_2^T\Sigma^{-1})\mu_2 \\
&\Rightarrow x^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}\mu_2 \\
&\Rightarrow (\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1})x - x^T(\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1}) - (\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2)
\end{aligned}$$

Considering the second term in the above expression, We know that transpose of a scalar is a scalar quantity, we have,

$$
\begin{aligned}
&(x^T(\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1}))^T \\
&= (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T(x^T)^T \\
&= (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T x
\end{aligned}
$$

Substituting the above found results in the formula for $P(Y|X) =$, Also we know that $\frac{1-\pi}{\pi} = exp^{log[\frac{1-\pi}{\pi}]}$

$$
\begin{aligned}
P(Y|X) &= \frac{1}{1 + \frac{1-\pi}{\pi}\exp[-\frac{1}{2}(\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1})x - (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T x - (\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2)]} \\
&= \frac{1}{1 + \frac{1-\pi}{\pi}\exp[-\frac{1}{2}((\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1}) - (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T)x - (\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2)]} \\
&= \frac{1}{1 + \exp[-\frac{1}{2}((\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1}) - (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T)x - (\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 - ln(\frac{1-\pi}{\pi}))]}
\end{aligned}
$$

The above equation is in the form of,

$$
\frac{1}{1 + exp[-\theta^T x + b]}
$$

Where

$$
\begin{aligned}
\theta^T &= \frac{1}{2}((\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1}) - (\mu_1\Sigma^{-1} - \mu_2\Sigma^{-1})^T) \\
b &= \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 - ln(\frac{1-\pi}{\pi}))
\end{aligned}
$$

3 Programming - Linear Regression

3.1 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

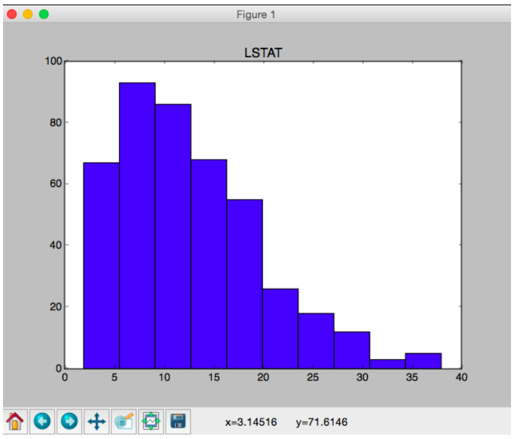
TAX : -0.468849385373

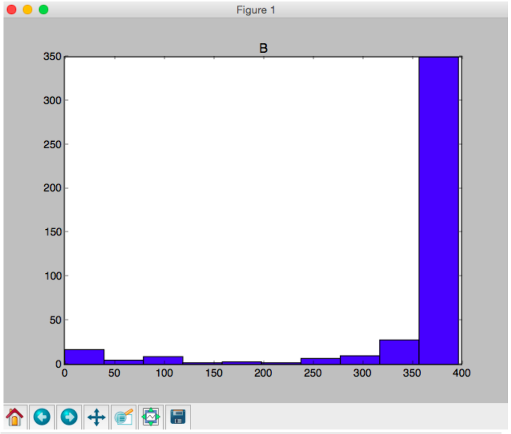
PTRATIO : -0.505270756892

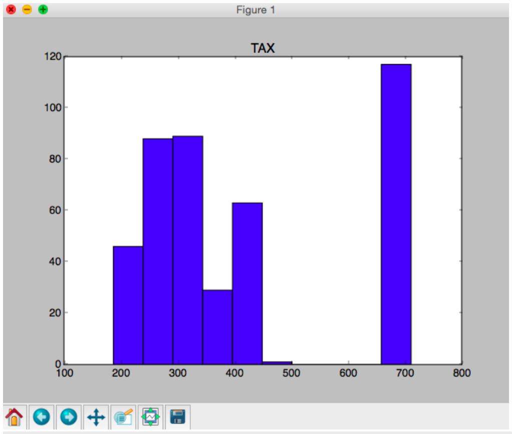
B : 0.343434137151

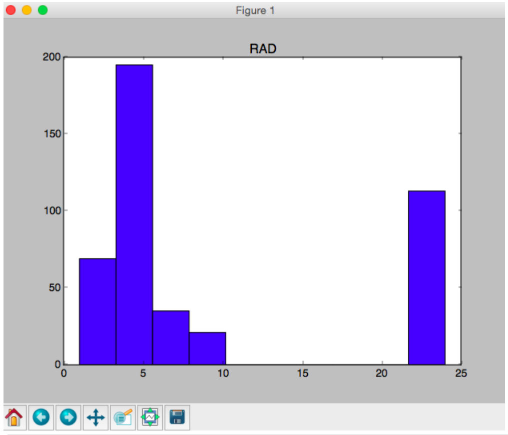
LSTAT : -0.73996982063

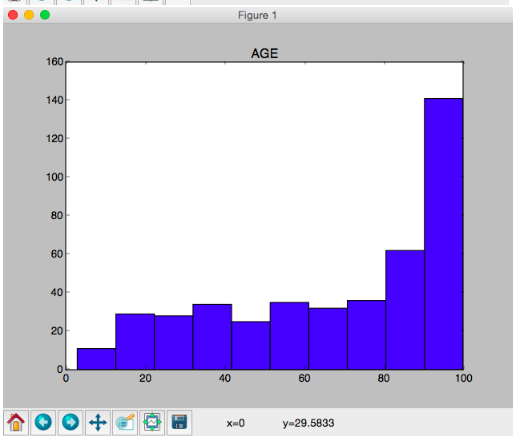
3.2 Hostograms

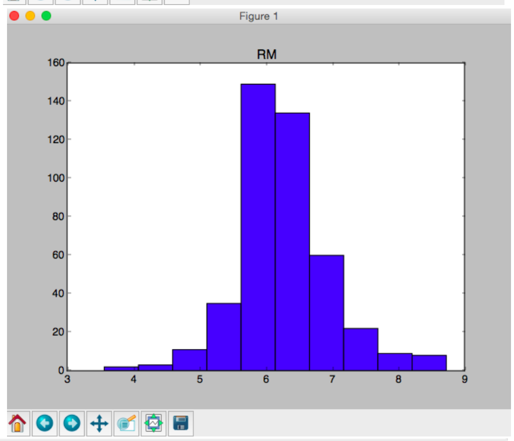


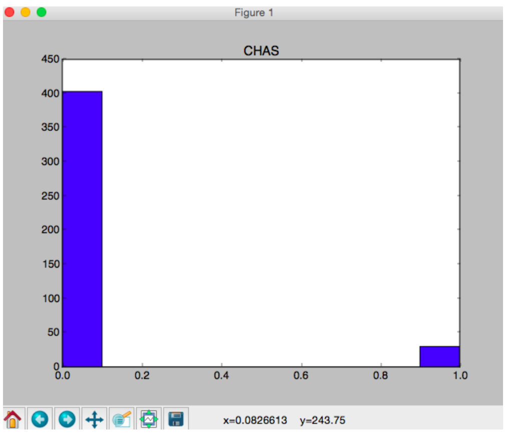


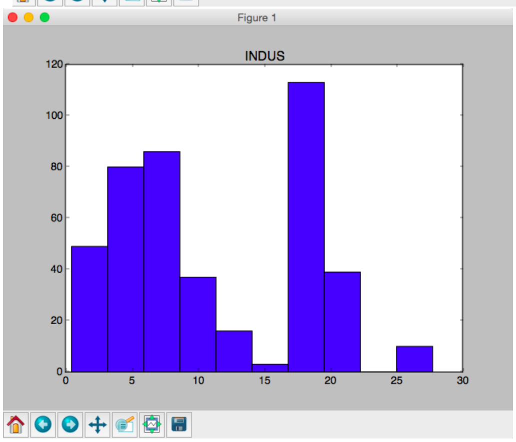


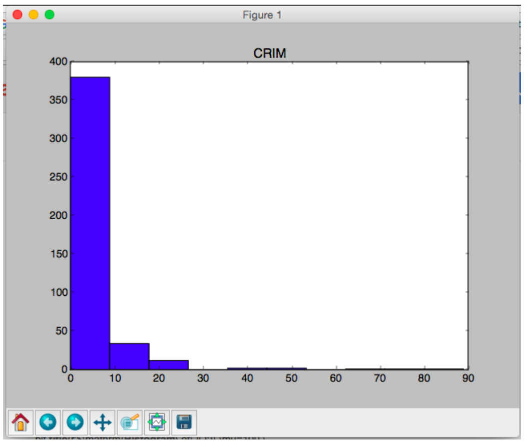












3.3 Linear Regression

Training Data

(a) Mean Squared Error loss on the training data = 20.950144508

Testing Data

(a) Mean Squared Error loss on the test data = 28.4179164975

3.4 Ridge Regression

Training Data

(a) Mean Squared Error loss on the training data with lambda as 0.01 = 20.9501449007
(b) Mean Squared Error loss on the training data with lambda as 0.1 = 20.9501837112
(c) Mean Squared Error loss on the training data with lambda as 1.0 = 20.9539971078

Testing Data

(a) Mean Squared Error loss on the testing data with lambda as 0.01 = 28.4182927619
(b) Mean Squared Error loss on the testing data with lambda as 0.1 = 28.4216969435
(c) Mean Squared Error loss on the testing data with lambda as 1.0 = 28.4574903672

3.5 Ridge Regression - Cross Validation

1. Lowest Train MSE is 33.2000882656 for lambda 5.4101
2. Test MSE for lambda 5.4101 is 28.677948865
3. for 10 different Lambda values, the Average cross validation MSE and train MSE are as shown below.

Lambda : 1.0001	Average CV MSE: 33.4324448072	Test MSE: 28.4574945202
Lambda : 2.0001	Average CV MSE: 33.3364179448	Test MSE: 28.5009623961
Lambda : 3.0001	Average CV MSE: 33.2669610919	Test MSE: 28.5482776673
Lambda : 4.0001	Average CV MSE: 33.2225853807	Test MSE: 28.5994090937
Lambda : 5.0001	Average CV MSE: 33.2019611649	Test MSE: 28.6543355715
Lambda : 6.0001	Average CV MSE: 33.2038894663	Test MSE: 28.7130426381
Lambda : 7.0001	Average CV MSE: 33.2272802303	Test MSE: 28.7755199525
Lambda : 8.0001	Average CV MSE: 33.2711354883	Test MSE: 28.8417594847
Lambda : 9.0001	Average CV MSE: 33.3345361138	Test MSE: 28.9117542229

3.6 Feature Selection

1. For 4 features with highest correlation with the target,
   (a) Top 4 features: 'LSTAT', 'RM', 'PTRATIO', 'INDUS'
   (b) Training MSE: 26.4066042155
   (c) Testing MSE: 31.4962025449
2. For Brute Force approach,

   Training:
   i. Top 4 features: 'CHAS', 'RM', 'PTRATIO', 'LSTAT'
   ii. Minimum Training MSE: 25.1060222464
   iii. Testing MSE: 34.6000723135
3. For Residue approach,
   (a) The 4 features selected are:'LSTAT', 'RM', 'PTRATIO', 'CHAS'
      i. Minimum Training MSE: 25.1060222464
      ii. Testing MSE: 34.6000723135

3.7 Polynomial Feature Expansion

1. Training results: 5.05978429711
2. Testing results: 14.5553049723

Collaborators: Shitesh Saurav, Piyush Gupta