CSCI 567 Fall 2016
Homework 2

MohammedJunaid Hundekar
2548-4069-96
hundekar@usc.edu
Collaborators: Sagar Makhwana, Dhawal Shah, Ankit Kothari

October 3, 2016

1 Logistic Regression

a. Negative Log Likelihood as loss function: Consider the probability of a single training sample (x_n, y_n)

$$p(y_n|x_n;b;w) = \sigma(b + w^T x_n) \qquad if y_n = 1 \tag{1}$$
$$p(y_n|x_n;b;w) = 1 - \sigma(b + w^T x_n) \qquad if y_n = 0 \tag{2}$$
$$p(y_n|x_n;b;w) = \sigma(b + w^T x_n)^{y_n}[1 - \sigma(b + w^T x_n)]^{1-y_n} \tag{3}$$
$$L(P(D)) = \prod_n \{\sigma(b + w^T x_n)^{y_n}[1 - \sigma(b + w^T x_n)]^{1-y_n}\} \tag{4}$$

Taking log likelihood on the whole training set of size D$(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$

$$\log L(P(D)) = \sum_n \{y_n \log \sigma(b + w^T x_n) + (1 - y_n) \log[1 - \sigma(b + w^T x_n)]\} \tag{5}$$

Taking negative of the log likelihood

$$\varepsilon(b, w) = -\sum_n \{y_n \log \sigma(b + w^T x_n) + (1 - y_n) \log[1 - \sigma(b + w^T x_n)]\} \tag{6}$$

For convenience
Append 1 to x $[1 \quad x_1 \quad x_2 \quad x_3 \quad ... \quad x_n]$
Append b to w $[b \quad w_1 \quad w_2 \quad w_3 \quad ... \quad w_n]$
Negative Log likelihood simplifies to

$$\varepsilon(b, w) = -\sum_n \{y_n \log \sigma(w^T x_n) + (1 - y_n) \log[1 - \sigma(w^T x_n)]\} \tag{7}$$

b. Gradient Descent Model Consider $\sigma(a) = \frac{1}{1+e^{-a}}$

$\frac{d\sigma(a)}{da} = \sigma(a)[1 - \sigma(a)]$

$\frac{d\log\sigma(a)}{da} = 1 - \sigma(a)$

Taking derivative of equation 7 w.r.t w

$$\frac{\partial\varepsilon(w)}{\partial w} = -\sum_n \{y_n[1 - \sigma(w^T x_n)]x_n(1 - y_n)\sigma(w^T x_n)x_n\} \tag{8}$$

$$\frac{\partial\varepsilon(w)}{\partial w} = \sum_n \{\sigma(w^T x_n) - y_n\}x_n \tag{9}$$

$$w^{(t+1)} = w^{(t)} - \eta\sum_n \{\sigma(w^T x_n) - y_n\}x_n \qquad \eta > 0 \tag{10}$$

Gradient descent works by updating the weights by using equation 10.
For the gradient descent to converge we need to select the step size (η) carefully.
If η is too small then the algorithm will take a long time to converge, on the other hand if η is too long the algorithm will oscillate and may not converge.

c. Log Likelihood Multi-Class logistic regression Given

$$P(Y = k|X = x) = \frac{\exp(w_k^T x)}{1 + \sum_1^{k-1}\exp(w_t^T x)} \qquad \text{for k=1,2...k-1} \tag{11}$$

$$P(Y = k|X = x) = \frac{1}{1 + \sum_1^{k-1}\exp(w_t^T x)} \qquad \text{for k=K} \tag{12}$$

We can simplify the above expression by introducing another fixed parameter $w_k = 0$
Thus we get

$$P(Y = k|X = x) = \frac{\exp(w_k^T x)}{\sum_1^{K-1}\exp(w_t^T x)} \qquad \text{for k=1,2...k-1} \tag{13}$$

$$P(Y = k|X = x) = \frac{1}{\sum_1^{K-1}\exp(w_t^T x)} \qquad \text{for k = K} \tag{14}$$

$$P(Y = k|X = x) = \frac{\exp(w_k^T x)}{\sum_1^{K}\exp(w_t^T x)} \qquad \text{By adding 13 and 14} \tag{15}$$

Let us y_n by an vector $\boldsymbol{y}_n = [y_{n1} \quad y_{n2} \quad y_{n3} \quad y_{n4} \quad \cdots \quad y_{nK}]^T$
Where
$y_{nk} = 1$ if $y_n = k$
$y_{nk} = 0$ otherwise

Taking the negative of the log likelihood

$$-\log L(P(D)) = -\sum_n \log P(y_n|x_n) \tag{16}$$

$$= -\sum_n \log \prod_{k=1}^{K} P(C_k|x_n)^{y_{nk}} \tag{17}$$

$$= -\sum_n \sum_k y_{nk} \log P(C_k|x_n) \tag{18}$$

$$= \sum_n \sum_k y_{nk} \log(\frac{\exp(w_k^T x)}{\sum_1^K \exp(w_t^T x)}) \tag{19}$$

$$l(w_1, w_2 ... w_k) = -\sum_{i=1}^{n} \sum_k y_{ik} \log P(y = y_{ik}|x = x_i) \tag{20}$$

$$l(w_1, w_2 ... w_k) = -\sum_n \sum_k y_{nk}[w_k^T x - \log(\sum_1^K \exp(w_t^T x))] \tag{21}$$

d. Gradient Descent of (c) Taking derivative of 20 w.r.t ∂w_i

$$\frac{\partial - l(w_1, w_2 ... w_k)}{\partial w_i} = \sum_n [\frac{\exp(w_i^T x)}{\sum_1^K \exp(w_t^T x)} x_i - x_i y_{ki}] \tag{22}$$

$$= \sum_n (x_i[P(Y = y_{ki}|X = x_i) - y_{ki}]) \tag{23}$$

Update rule for w

$$w_k \leftarrow w_k - \sum_i (P(Y = y_{ki}|X = x_i) - y_{ki}) x_i \tag{24}$$

2 Linear/Gaussian Discriminant

a. Given: $D = \{(x_n, y_n)\}_{n=1}^N; \qquad y_n \in \{1, 2\}$

$$p(x_n, y_n) = p(y_n)p(x_n) \tag{25}$$

$$= p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(- \frac{(x_n - \mu_1)^2}{2\sigma_1^2} \right) \quad if y_n = 1 \tag{26}$$

$$= p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(- \frac{(x_n - \mu_2)^2}{2\sigma_2^2} \right) \quad if y_n = 2 \tag{27}$$

$$\log P(D) = \sum_n \log p(x_n, y_n) \tag{28}$$

$$= \sum_{n:y_n=1} \log \left(p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(- \frac{(x_n - \mu_1)^2}{2\sigma_1^2} \right) \right) + \sum_{n:y_n=2} \log \left(p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(- \frac{(x_n - \mu_2)^2}{2\sigma_2^2} \right) \right) \tag{29}$$

$$= \sum_{n:y_n=1} (\log p_1 - \log\sqrt{2\pi}\sigma_1 - \frac{(x_n - \mu_1)^2}{2\sigma_1^2}) + \sum_{n:y_n=2} (\log p_2 - \log\sqrt{2\pi}\sigma_1 - \frac{(x_n - \mu_2)^2}{2\sigma_2^2}) \tag{30}$$

Now we can maximize $\{p_1, \mu_1, \sigma_1, p_2, \mu_2, \sigma_2\}$ separately from the above equation by taking derivative and equating to zero for each term.

$$p_2 = 1 - p_1 \tag{31}$$

$$\frac{dl(D)}{dp_1} = \frac{\sum_{n:y=1} 1}{p_1} - \frac{\sum_{n:y=2} 1}{1 - p_1} = 0 \tag{32}$$

$$\frac{N_{y=1}}{p_1} = \frac{N_{y=2}}{1 - p_1} \tag{33}$$

$$p_1 = \frac{N_{y=1}}{N} \tag{34}$$

$$\text{Similarly,} \quad p_2 = \frac{N_{y=2}}{N} \tag{35}$$

$$\frac{dl(D)}{d\mu_1} = \sum_{n:y=1} [\frac{-2(x_n - \mu_1)(-1)}{2\sigma_1^2}] = 0 \tag{36}$$

$$\sum_{n:y=1} (1)\mu_1 = \sum_{n:y=1} x_n \tag{37}$$

$$\mu_1 = \frac{\sum_{n:y=1} x_n}{N_{y=1}} \tag{38}$$

$$\text{Similarly,} \quad \mu_2 = \frac{\sum_{n:y=2} x_n}{N_{y=2}} \tag{39}$$

$$\frac{dl(D)}{d\sigma_1} = \sum_{n:y=1} \left([\frac{-1}{\sqrt{2\pi}\sigma_1}\sqrt{2\pi}] - [\frac{(x_n - \mu_2)^2}{2\sigma_1^3}(-2)] \right) = 0 \tag{40}$$

$$\frac{\sum_{n:y=1}(x_n - \mu_2)^2}{\sigma_1^3} = (\frac{\sum_{n:y=1} 1}{\sigma_1}) \tag{41}$$

$$\sigma_1^2 = \frac{\sum_{n:y=1}(x_n - \mu_1)^2}{N_{y=1}} \tag{42}$$

$$\text{Similarly,} \quad \sigma_2^2 = \frac{\sum_{n:y=2}(x_n - \mu_2)^2}{N_{y=2}} \tag{43}$$

b. Given $P(x|y=c_1) = \mathcal{N}(\mu_1, \Sigma)$ and $P(x|y=c_2) = \mathcal{N}(\mu_2, \Sigma)$
Assume $P(y=1) = \pi; \qquad P(y=2) = 1-\pi$

$$P(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Bigg(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\Bigg) \tag{44}$$

$$P(x|y=2) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Bigg(-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)\Bigg) \tag{45}$$

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1) + P(x|y=2)P(y=2)} \tag{46}$$

$$= \frac{1}{1 + \frac{P(x|y=2)P(y=2)}{P(x|y=1)P(y=1)}} \tag{47}$$

$$= \frac{1}{1 + \exp(-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\frac{1-\pi}{\pi}} \tag{48}$$

$$= \frac{1}{1 + \frac{1-\pi}{\pi}\exp(\sum_{i=1}^{N}[\frac{(x_i-\mu_{1i})^2}{2\sigma_i^2} - \frac{(x_i-\mu_{2i})^2}{2\sigma_i^2}]} \tag{49}$$

$$= \frac{1}{1 + \exp\Bigg[\ln\frac{1-\pi}{\pi} + \sum_i \Bigg(x_i\frac{\mu_{2i}-\mu_{1i}}{\sigma_i^2} + \frac{\mu_{1i}^2-\mu_{2i}^2}{2\sigma_i^2}\Bigg)\Bigg]} \tag{50}$$

$$\theta_1 = -(\ln(\frac{1-\pi}{\pi}) + \sum_i (\frac{\mu_{2i}^2 - \mu_{1i}^2}{2\sigma_i^2})) \tag{51}$$

$$\theta_2 = -(\sum_i (\frac{\mu_{1i} - \mu_{2i}}{\sigma_i^2})) \tag{52}$$

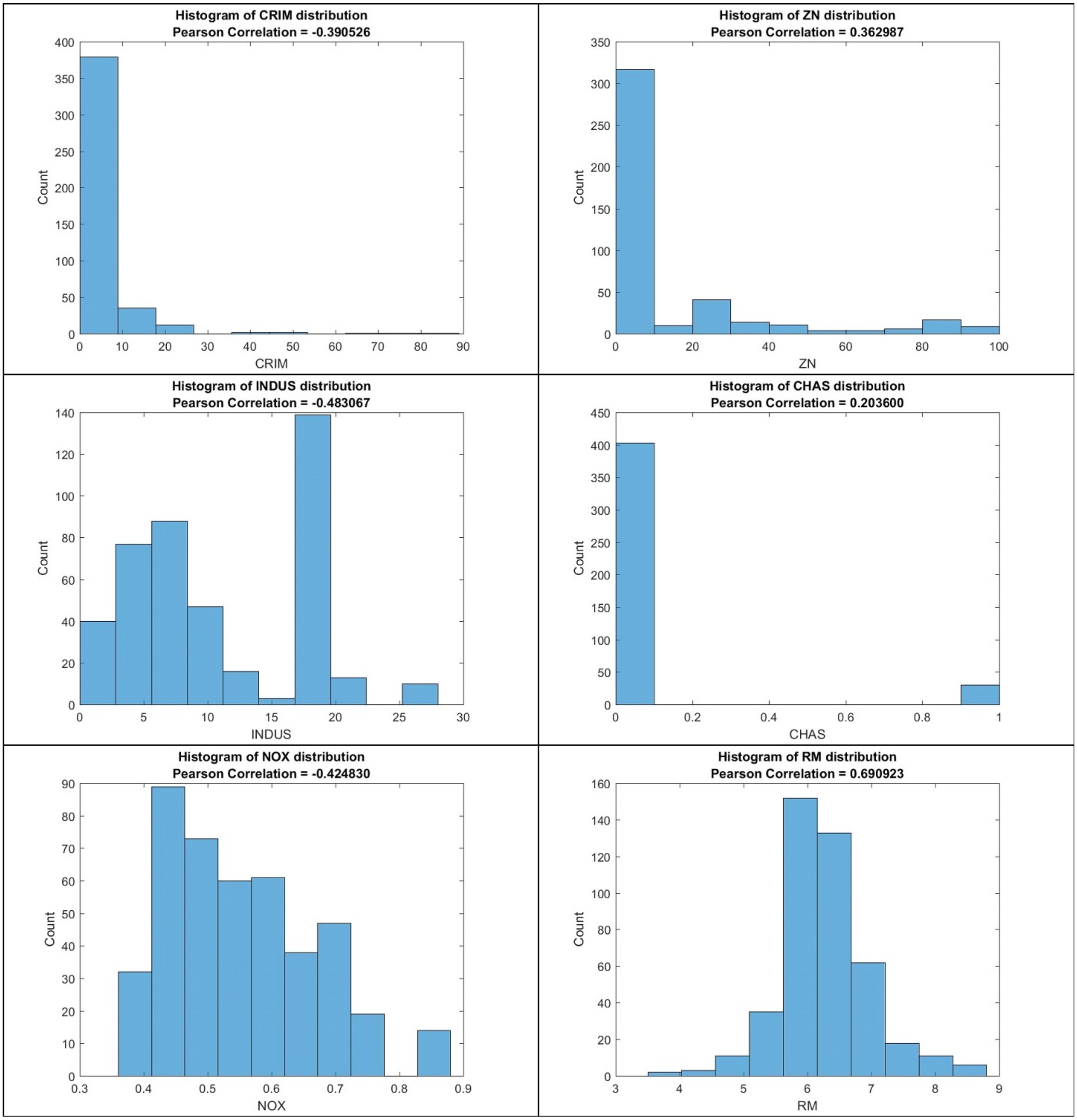
$$P(Y=1|x) = \frac{1}{1+\exp(-\theta_1 - \theta_2 X)} \tag{53}$$

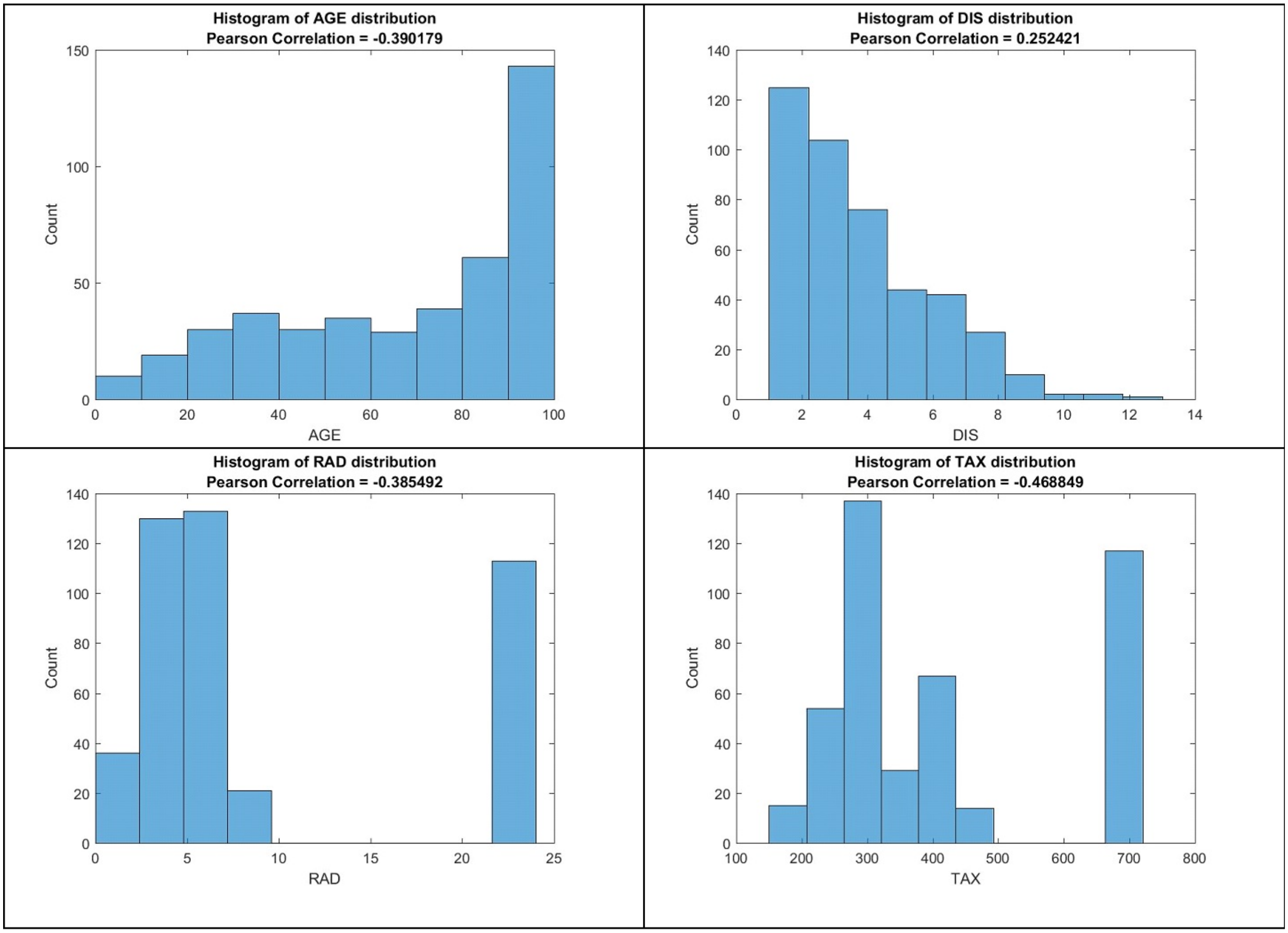
$$P(Y=1|x) = \frac{1}{1+\exp(-\theta^T X)} \tag{54}$$

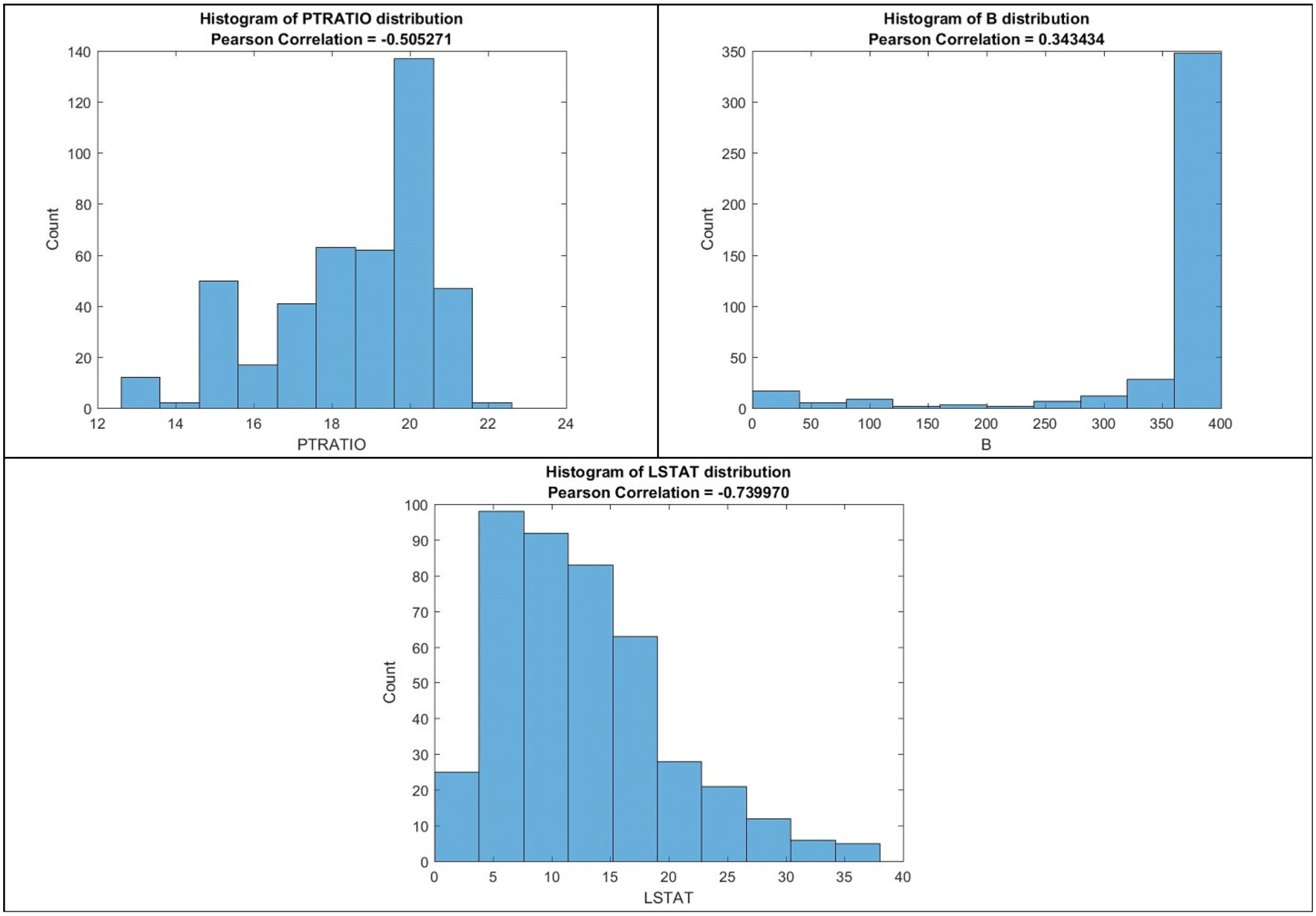
Appending a 1 in $X \leftarrow [1 \quad x_1 \quad x_2 \quad ... \quad x_n]$ and $\theta = \theta_1 + \theta_2$

3 Programming - Linear Regression

3.1 Data Analysis




Histogram of AGE distribution
Pearson Correlation = -0.390179
Count
AGE
Histogram of DIS distribution
Pearson Correlation = 0.252421
Count
DIS
Histogram of RAD distribution
Pearson Correlation = -0.385492
Count
RAD
Histogram of TAX distribution
Pearson Correlation = -0.468849
Count
TAX



3.2 Linear Regression

Table 1: Linear and Ridge Regression Performance on Training and Test Data

Algorithm	**Training Set MSE**	**Testing Set MSE**
Linear Regression	20.9441	28.4368
Rigde Regression L =0.01	20.9441	28.4371
Rigde Regression L =0.10	20.9442	28.4405
Rigde Regression L =1.00	20.948	28.476

Ridge Regression with Cross-Validation:

Incrementing λ by 0.01 after each iteration. Displaying only every 100^{th} row for conciseness. See variable **Res_cv** for all values.

Table 2: Lamda and MSE on Training Data

Lamda Value	**MSE**
0.0001	32.887567
0.9801	32.790939
1.9801	32.716183
2.9801	32.66416
3.9801	32.633673
4.9801	32.623659
5.9801	32.633161
6.9801	32.661308
7.9801	32.707307
8.9801	32.770422
9.9801	32.849976

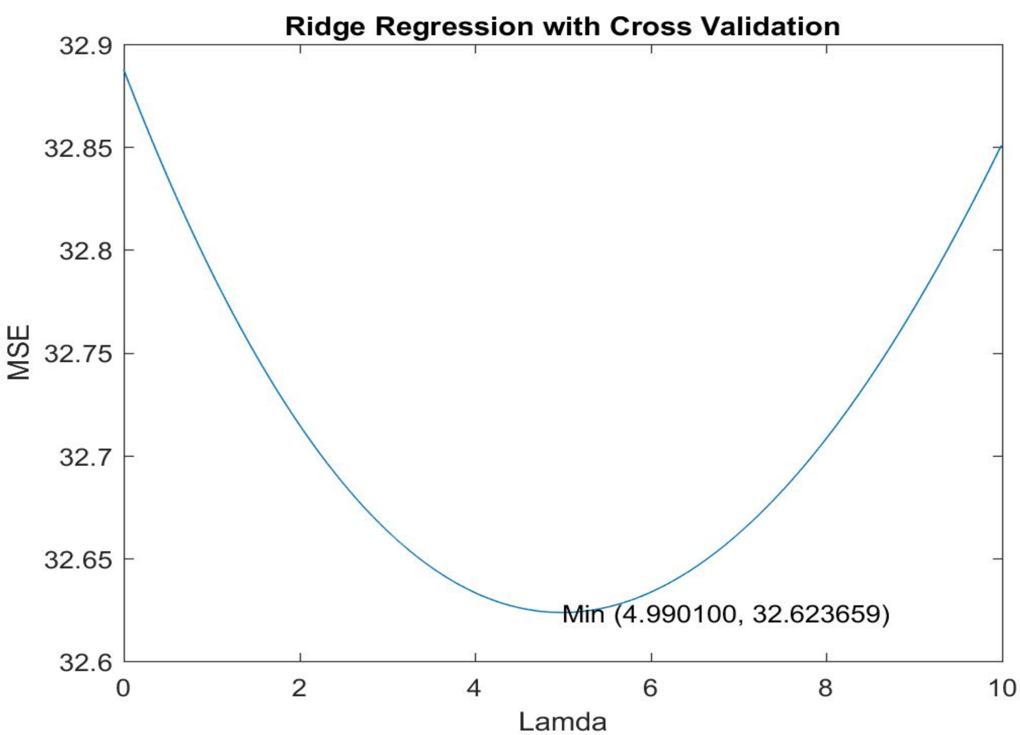


Table 3: Results of Cross validation on Testing Set

Lamda Value	**MSE**
4.990100	28.671087

From the graph we can see that when $\lambda = 4.990100$ we get the minimum MSE on Training Data: MSE = 32.623659.
Choosing this, we get MSE = 28.671087 on the testing set.

3.3 Feature Selection

a. Four features with highest absolute correlation

Table 4: Features with highest absolute correlation

Attribute	**Name**	**Correlation**
13	LSTAT	0.74
6	RM	0.6909
11	PTRATIO	0.5053
3	INDUS	0.4831

Using the above 4 features to train the linear regression
MSE on training data:: 26.406604
MSE on Testing data:: 31.496203

b. Four features with highest absolute correlation with Residue

Table 5: Features and their correlation with Residue

Attribute	**Name**	**Correlation**
13	LSTAT	0.74
6	RM	0.3709
11	PTRATIO	0.2975
4	CHAS	0.2196

Using the above 4 features to train the linear regression
MSE on training data:: 25.106022
MSE on Testing data:: 34.600072

Selection with Brute-force Search

The columns that give MIN MSE: 25.106022 on Training SET: [4 6 11 13]
Corresponding MSE: 34.600072 on Testing SET

The columns that give MIN MSE: 30.100406 on Testing SET: [6 11 12 13]
Corresponding value of MSE: 25.744417 on Training SET

3.4 Polynomial Feature Expansion

Expanding the existing features by polynomial expansion $x_i * x_j\{i, j = 1, 2, 3...13\}$ to get 104 features. The result of training the linear regression model on these feature are:
MSE on Training data:: 5.077346
MSE on Testing data:: 14.559306