ML Assignment-1

midigarahallishanm.r

September 2018

1 Question 1

$$X^T AX = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots x_n \end{pmatrix}$$

After multiplying three matrices we have:

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

consider some kth row in the above vector to partially differentiate

$$\frac{\partial}{\partial x_k} X^T AX = \frac{\partial}{\partial x_k} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j \right)$$

$$= x_1 a_{k1} + \ldots + \left(\sum_{i=1}^{n} a_{ik} x_i + x_k a_{kk} \right) + \ldots + x_n a_{kn}$$

this can be re-written as follows

$$= \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$$

= (Kth row of A + transpose of kth row of A) X

when we combine the partial differentiation value for all the rows of the above vector, we will have the following

$$\frac{\partial}{\partial x_k} X^T AX = [[\text{first row of A}] + [\text{Transpose of first row of A}]]X$$

$$\vdots \qquad \vdots$$

$$[\text{[last row of A]} + \text{[Transpose of last row of A]}]X$$

$$\frac{\partial}{\partial x_k} X^T A X = (A + A^T)X$$

Gaussian(first derivative) is as shown bellow:

$$\frac{\partial}{\partial x_k}(X^T A X + B^T X) = (A + A^T)X + B^T$$

Second derivative gives the Hessian, which is as follows:

$$\frac{\partial^2}{\partial^2 x_k}(X^T A X + B^T X) = (A^T + A)$$

2 Question 1.2

In order to fetch the rank of the matrix, the matrix needs to be reduced to row echelon matrix first. Once we have the row echelon matrix in hand, the count of non-zero rows in the matrix gives the rank of the given matrix.

A matrix is said to be in row echelon form, if it satisfies the following condition:

1. The leading non-zero entry in the row is 1
2. The leading entry of all the row is to the right of the leading entry of the previous row.
3. Rows with just zeros is below the rows with non-zero entries.

In this case, the matrix obtained is in the reduced echelon form, with leading entry(1) being the only non-zero element in each row

Assumptions:

1> Rn -> represents row n
2> Cn -> represents column n

In order to reduce the given matrix to row-echelon form, we perform a series of row and column operations as shown below:

Rank of the matrix:

$$A = \begin{pmatrix} 4 & 0 & 2 & 3 & 1 \\ 2 & -1 & 2 & 0 & 1 \\ 5 & 2 & 2 & 1 & -1 \\ 4 & 0 & 2 & 2 & 1 \\ 2 & -2 & 0 & 0 & 1 \end{pmatrix}$$

R1-R4 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 2 & -1 & 2 & 0 & 1 \\ 5 & 2 & 2 & 1 & -1 \\ 4 & 0 & 2 & 2 & 1 \\ 2 & -2 & 0 & 0 & 1 \end{pmatrix}$$

c3-2c5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 2 & -1 & 2 & 0 & 1 \\ 5 & 2 & 4 & 1 & -1 \\ 4 & 0 & 0 & 2 & 1 \\ 2 & -2 & -2 & 0 & 1 \end{pmatrix}$$

R2-R5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 5 & 2 & 4 & 1 & -1 \\ 4 & 0 & 0 & 2 & 1 \\ 2 & -2 & -2 & 0 & 1 \end{pmatrix}$$

R2-R5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 5 & 2 & 4 & 1 & -1 \\ 4 & 0 & 0 & 2 & 1 \\ 2 & -2 & -2 & 0 & 1 \end{pmatrix}$$

R5+R2 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 5 & 2 & 4 & 1 & -1 \\ 4 & 0 & 0 & 2 & 1 \\ 2 & -1 & 0 & 0 & 1 \end{pmatrix}$$

C4 - C5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 5 & 2 & 4 & 0 & -1 \\ 4 & 0 & 0 & 1 & 1 \\ 2 & -1 & 0 & -1 & 1 \end{pmatrix}$$

`C2 + C5 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 5 & 1 & 4 & 0 & -1 \\ 4 & 1 & 0 & 1 & 1 \\ 2 & 0 & 0 & -1 & 1 \end{pmatrix}$$

`C1 - 4C5 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 1 & 4 & 0 & -1 \\ 0 & 1 & 0 & 1 & 1 \\ -2 & 0 & 0 & -1 & 1 \end{pmatrix}$$

`R4 - R1 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 1 & 4 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & -1 & 1 \end{pmatrix}$$

`R5 + R1 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 1 & 4 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 1 \end{pmatrix}$$

`R3 - R2 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 0 & 2 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 1 \end{pmatrix}$$

`R3 + R4 results in`

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 1 \end{pmatrix}$$

R3 - R2 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 1 \end{pmatrix}$$

C2 - C5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -2 & -1 & 0 & 0 & 1 \end{pmatrix}$$

C1 - 2C2 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ -2 & 1 & 2 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

R5 - R4 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ -2 & 1 & 2 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 \end{pmatrix}$$

R2+ R5 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ -2 & 0 & 2 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 \end{pmatrix}$$

c1 -c3 results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 \end{pmatrix}$$

c3 -> c3/2 and c1-> c1/9 and c2 => (-1 *c2) results in

$$= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Swap C4 and C1 followed by
swapping C3 with C2 followed by
swapping C3 and C4 followed by
swapping C4 and C5,
Resulting matrix is as shown below:

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The above matrix is in reduced row-echelon form.
The number of non-zero rows in the row-echelon matrix
represents the rank of the matrix.

Therefore Rank of above Matrix = 5

3 Question 2.1

Given : P(A, B, C) = P(A)P(B|A)P(C|B)

To say A and C are conditionally independent, given B, we need to prove:
P(A, C | B) = P(A|B) P(C|B)

Using Bayes rule on P(A | B) in given equation we have :

P(A, B, C) = P(A) P(A|B) P(B) P(C|B) * 1/ P(A)

```
 => P(A, B, C)
    ----------   =    P(A|B) P(C|B)   ---->  equation (1)
       P(B)
```

```
We know that P(A|B) = P(A, B)    ----> equation (2)
                      --------
                        P(A)
```

Using Equation (2) on LHS of Equation (1) we have:

```
P(A,C|B)  = P(A|B) P(C|B)
```

Hence proved! that A and C are conditionally independent, given B.

4 Question 2.2

The number of heads is larger than 140 but smaller than 160
using Central Limit Theorem:

Let X1, X2, ... Xn be the random sample of coin toss outcome of size n(=300).

$$X_i = \begin{cases} 1 & \text{if } i^{th}\text{toss results in heads} \\ 0 & \text{otherwise} \end{cases}$$

Expected value of the coin toss is as shown below:

$$E[X_i] = \mu = n * p$$
$$= 300 * 1/2 = 150$$

Finite variance is calculated as shown below:

$$var(X_i) = (\sigma^2) = np(1-p) = 300 * 0.5(1-0.5) = 75$$

$\implies$ Standard Deviation$\sigma = \sqrt{75}$

By using central limit theorem, we can calculate approximate probability that the coin tossed 300 times results in heads more than 140 times and less than 160 times.

$$S_300 = \sum_{i=1}^{300} X_i$$
$$z = \frac{\bar{x} - \mu}{\sigma}$$

$$P(S_{300} \in [140,160]) = P\left(\frac{S_{300} - 300\mu}{\sqrt{300\sigma^2}} \in \left[\frac{140-150}{\sqrt{75}}, \frac{160-150}{\sqrt{75}}\right]\right)$$

$$= P\left(\frac{S_{300} - 300\mu}{\sqrt{300\sigma^2}} <= 1.15\right) - P\left(\frac{S_300 - 300\mu}{\sqrt{300\sigma^2}} <= -1.15\right) \text{——equation(1)}$$

By central limit theorem we know that:

$$Pr\Big[\frac{S_n - n\mu}{\sqrt{n * \sigma^2}} \leq z\Big] \approx \phi(z)$$

Therefore,

$$\approx \phi(1.15) - \phi(-1.15)$$

By referring chart of cumulative probabilities of the standard normal distribution we have,

$$\phi(1.15) = 0.8749$$

substituting in equation (1), we have:

$$= 0.8749 - (1 - 0.8749)$$

$$= 0.7498$$

5 Question 3

Question 3.a

Labelling graphs to represent their bias and variance

```
Model a : High bias and low variance. This results in under-fitting
Model b : Low variance and Low variance. This is what we are trying to achieve
          in bias-variance trade-off
Model c : Low bias and Low variance. This results in Over-fitting.
```

Question 3.b

In order to evaluate model for over-fitting and under-fitting, we can make use of bias and variance. Bias is the error that is introduced by the erroneous data whereas Variance is the measure of the model sensitivity to small changes in training data set.

If the model has high bias, then it fails to capture the pattern resulting in relationship between the features and target. This results in under-fitting. This is caused when we use small data-set for training the model.

High variance is caused in case of complex models when handling large data-set. In this case, model capture the noise along with underlying pattern in the training data-set. This results in Over-fitting.

In case of under-fitting, model fails to work well for both training as well as testing data set.

In case of over-fitting, model works well for training data set, but fails to generalize.

Mathematical representation of expected error of the model is as follows:

Let y be the target,

f(y) be the hypothesis function,

$f(\hat{X})$ be the approximation of f(X),

σ^2 represents the irreducible error that is associated with the model

$$E\Big[\big(y - f(\hat{X})\big)^2\Big] = \Big(Bias\big[f(\hat{x})\big]\Big)^2 + Var\big[f(\hat{X})\big] + \sigma^2$$

Where,

$$\Big(Bias\big[f(\hat{x})\big]\Big)^2 = E\Big[f(\hat{x}) - f(x)\Big]$$

and,

$$Var\Big[f(\hat{X}) = E\Big[f(\hat{x})^2\Big] - \Bigg(E\Big[F(\hat{x})\Big]\Bigg)^2$$

6 Question 4.1

Feature Vector is a n-dimensional vector which represents the measurable feature values.

In this question, Bag of Words i.e.bug, fix, correct, error, wrong forms the features representing the git commit message.

$$valueOfVector = \begin{cases} 1 & \text{if word is present in the sentence} \\ 0 & \text{otherwise} \end{cases}$$

Feature Vector for the given sentences is as follows:

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In the above matrix columns represent the words bug, fix, correct, error and wrong respectively and each rows is the given git commit message vector representation.

7 Question 4.2

```
Linear regression model:

Used Ordinary Least square to compute the linear regression model parameters.
```

$$\text{Let} X_i \text{ represent value for each feature in the feature vector}$$
$$X \text{ represent the feature vector}$$
$$Y \text{ represent the target}$$
$$Y' \text{ represent the predicted target}$$
$$\beta \text{ represent the model parameter for all the features}$$
$$\text{N represent number of predictions made}$$

$$\beta = \left(X^T X\right)^{-1} X^T Y$$

```
Root mean square error(RMSE):
```

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i' - y_i\right)^2}{N}}$$

```
Target value (y) was calculated using:
```

$$Y = X\beta^T$$

Results of the Linear Regression Model implementation:

```
Model Parameter(beta):

    2.98626776e-03,
    -1.12335030e+00,
    -1.85954809e-01,
    -5.47126305e-04,
    -1.82335652e+00,
    3.52988824e-03,
    -3.66394475e-03,
    4.35526351e+00,
    -4.35905056e-01,
    7.87023208e-01,
    3.00411212e-01

Plotted graph for Model paramters

Root mean square error(RMSE): 0.652332741227717
```

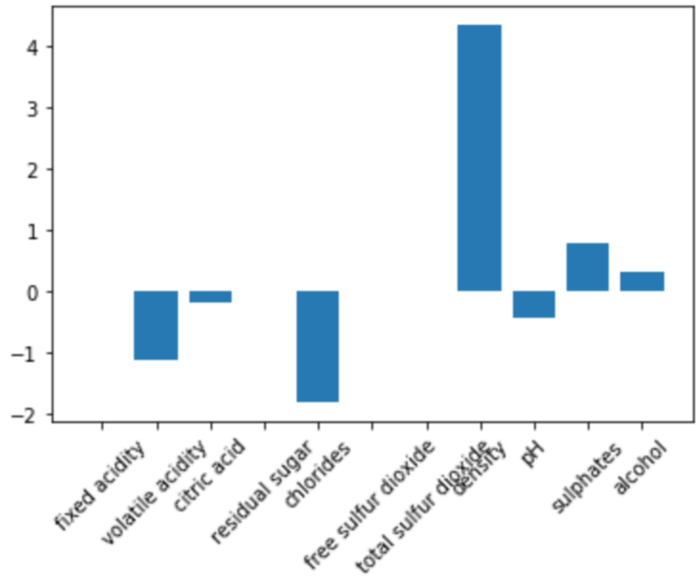

Figure 1: Bar plot of model parameters of Linear Regression Model

In order to determine the Quality of Wine, predictors can look closely in to the features with higher model parameter value.

Weighted Linear Regression :

In case of weighted linear regression, we assign same weights(=1) across all the feature parameters

In case of weighted linear regression, We make use of inverse of variance as weights.

In order to model this, following was implemented:

Let W represent a weight matrix of size N*N(diagonal matrix with weights as the diagonal elements).

In order to get weight matrix(W), fetch the co-variance matrix for the feature vector matrix(X). Diagonal elements of this matrix represents the variance of the variable when we are dealing with n-Dimensions in the model.

Just retaining diagonal elements and setting all the other row elements to zero, we get weight matrix W.

Model parameters are calculated using:

$$w_i = \frac{1}{\sigma^{(}2)}$$

$$\beta = \left(X^T W X\right)^{-1} X^T W Y$$

Root Mean square error:

$$\alpha_i \text{ represents the weight parameter}$$

$$RMSE = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \left(y_i^{'} - y_i\right)^2$$

Results of the Weighted Linear Regression Model implementation:

1. Model Parameter(beta):

```
-2.17466898e+00,
-2.40922917e-01,
-4.43590730e-01,
1.26250941e-01,
-2.04027501e+01,
3.88860983e-02,
-2.28020685e-02,
1.30243713e+02,
-3.19522664e+01,
2.88775330e+00,
2.26029529e-02
```

2. Root mean square error(RMSE): 0.13901718868644747

3. Plotted graph for Model paramters

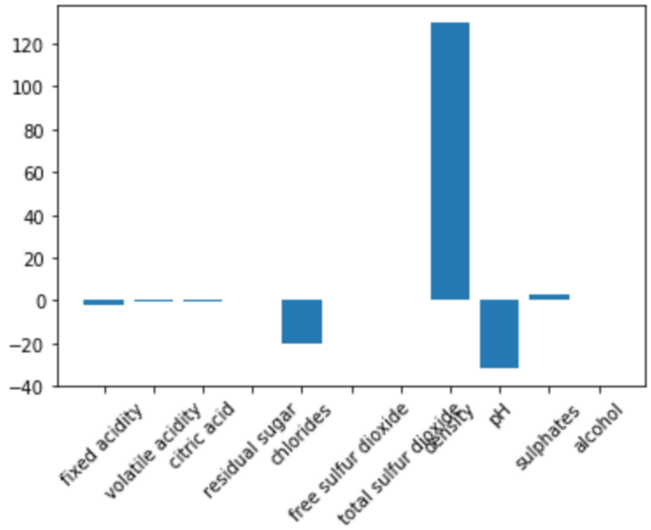

Figure 2: Bar plot of model parameters for Weighted Linear Model