**CSE 572 DATA MINING**

**Instructor- Prof. Arunabha Sen**

**ASSIGNMENT 4**

Group name: Sriharsha

Group ID: 15

Group members:

1. Sriharsha Uppu (1215318000)
2. Praveen Muruganandam (1215159608)
3. Chandrika Cuddapah Sudhinder (1215321133)
4. Nishant Trivedi (1214072053)

**Linear Regression:**

Regression is a method of modeling a target value based on independent predictors. This method is mostly used for forecasting and finding out the cause and effect relationship between variables.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

It consists of:

* Cost Function
* Gradient Descent

**Decision Tree:**

Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines*.*(repetitively because there may be two distant regions of same class divided by other).

So when does it terminate?

1. Divided into pure classes (only has members of single class )
2. Some criteria of classifier attributes are met.

Related terms:

* Impurity
* Entropy
* Information Gain

**Task 1:**

To perform linear regression on PB1\_Test.csv using the model developed by using PB1\_Train.csv.

The Model parameters are:[6.43234991 0.80954036]

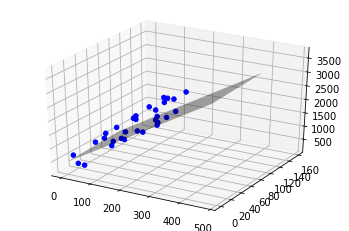
Intercept:467.3445829012643

The predicted values for the test\_csv are: [ 488.30468628 1445.76533622 689.45853301 632.86062651 1077.23847301 1344.61676464 579.16712538 1495.5171089 1410.99907384 1521.92413008 704.83979978 1400.65094097 1442.57114773 1512.84329441 976.24776574 788.17848352 862.04057513 829.17525856 1118.04132884 1125.7229485 1100.40733276 1402.8157244 931.00145167 897.57240492 1214.35068573 1489.30462369 1511.57599721 752.47076197 1122.94254392 1298.82474786]

The mean square error for the model are: 0.6091991650759796

The response variable for the given sample features [46,53] are: [806.13831763]

The 3D plot also including the regression plane is shown below.



**Task 2:**

To perform linear regression on PB2\_Test.csv using the model developed by using PB2\_Train.csv.

The Model parameters are:[7.98554746 0.03573724]

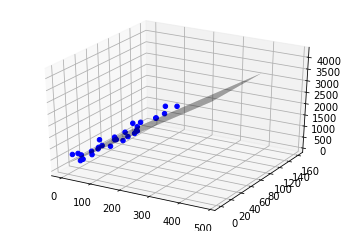
Intercept:290.6377035073201

The predicted values for the test\_csv is: [1472.85610043 1329.54509292 1491.54322528 606.00515051 291.45965993 1139.75029006 964.42561823 378.62167455 658.72336877 1102.71726884 506.74780632 645.53977823 459.51352902 1209.94056715 506.46190843 854.95087406 778.89969031 459.97811308 713.9074563 1561.16170659 1472.57020255 587.53244908 1226.41198338 834.40541296 646.07583677 796.55043532 1075.83017312 291.5311344 718.83919483 687.61174969]

The mean square error if the model for the given dataset are: 0.8560502775103181

The response value for the given sample feature values are : [445.07913523]

The 3D plot also including the regression plane is shown below.

****

**Task 3:**

Train a decision tree on PB3\_Train.csv and test it on PB3\_Test.csv.

Predicted values on PB3\_test.csv [1. 0. 1. 0. 1. 0. 0. 1. 0. 0. 1. 0. 1. 1. 1. 0. 1. 0. 0.]

Accuracy: 100.0

**Task 4:**

Train a decision tree on PB4\_Train.csv and test it on PB4\_Test.csv.

Predicted values on PB4\_test.csv [0. 0. 0. 0. 1. 0. 0. 1. 0. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1. 0.

1. 0. 0. 1. 1.]

Accuracy: 24.137931034482758