IS 777- Data Analytics for Statistical Learning

Homework 1

Tejaswini Srinivas

[fg43775@umbc.edu](mailto:fg43775@umbc.edu)

Collaborated with -

Aishwarya Mohan Bhonde

Harish Ramamoorthy

1 Bias-Variance Trade-Off

1. **Bias**-

Bias is the amount at which the target function value varies from the true estimates of the training datasets or the expectations of the errors while predicting a model.Bias is the difference between the predicted values with the actual value

When the Bias is Low- less expectations about the form of target function

When the Bias is High – more expectations about the form of target function. Which is nothing but underfitting.

**Variance**-

Variance is the amount that estimates the target function that would change on different training data set. With different training datasets, the value of the target function also differs in the machine learning algorithm. Ideally, the value of the target function should not change way too much from one to other training sets. This fact is because the algorithm is good at pointing to the underlying mappings between the input and output variables.

When the Variance is Low- that there is small changes to estimate the target function with changing training datasets.

When the Variance is High – then there is large changes to estimate the target function with changing training datasets. (over fitting)

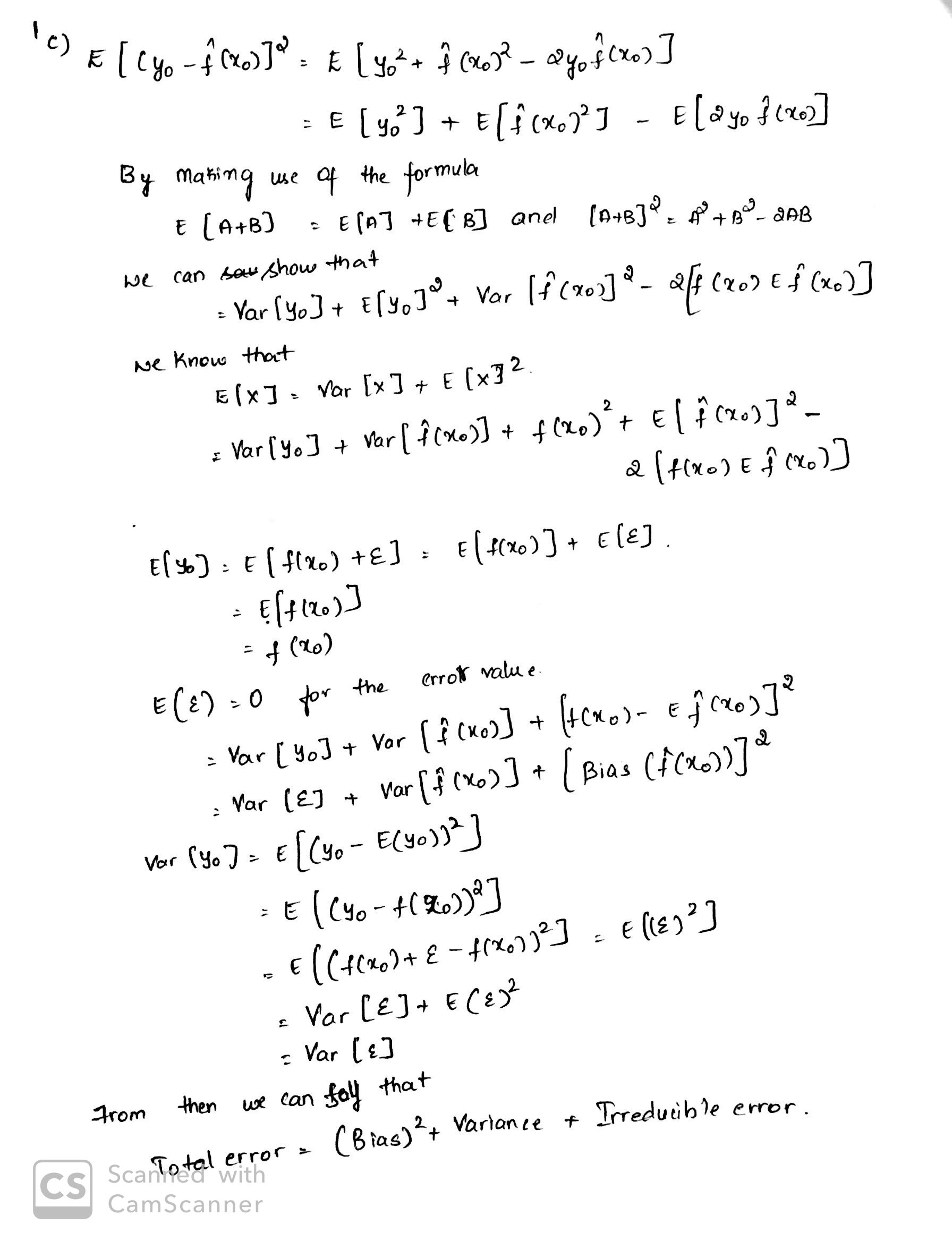
1. BIAS AND VARIANE TRADE-OFF-

Whenever using machine learning model, we would want the data to be fit perfectly.

The trade-off tries to get an optimal bias-variance model for the better consistency and more reliable model

Considering the possible bias-variance trade-off to optimize the value –

i.e. when we tried to increase bias, variance decrease and vice versa and to get better optimal bias-variance model there are multiple ways which depends on the training algorithm by reducing the dimensionality of data which removes features that adds to variance, in KNN by using optimal value of K can optimize bias- variance.



2 Linear Regression

1. Linear regression assumes uncertainty in the measurement of dependent variable as well as independent variable. This is because while calculating the linear relationship between the target variable (Y) and scalar response(X), there is always a reducible error ε that existed and we look for minimizing in the following equation by making use of Residual Sum of Square method.

y=β0+β1x1+⋯+βkxk+ε

Where ε contains the uncertainty measurements of X and Y .

1. Outlier impacts the model’s performance for test data, as it may occur due to data entry problems, technology limitation, or human error.

One can handle this issue by performing any one of the following methods-

* + Binning
  + Clustering
  + Regression

1. Non-parametric methods do not make explicitly assumptions about the function form of f, as they have the potential to precisely fit a wider range of likely shapes for “f”. This is true as a very large observation needed to find the accurate estimate of “f”.

This can be said by considering a model with 2 dependent variable which affects the dependent variable . However the variables are on a different scales i.e. X1 on[100-200] and X2 on [99-100], resulting the coefficient of X1 being too smaller than that of X2 in a linear regression.

1. Independent variables which are correlated to each other is said to be a Collinear. In regression model, Multicollinearity refers to the degree to which independent variables are correlated.

The degree of correlation between the independent variables should be small, if it is great then it is difficult to fit the model and predict the results.

Some of the problems occur when the Multicollinearity is great:

* + When an independent variable is ***perfectly*** correlated with another independent variable , an unique least-squares solution for regression does not exist
  + When one independent variable is ***highly*** correlated with another independent variable, the marginal contribution of that independent variable is influenced by other independent variables.

1. The outcome of posterior probability cut-off 0.5 for linear regression differs from that of posterior probability cut-off 0.5 for logistic regression. A 0.5 value on logistic regression tells us the mentioned points belong to 1 single class.

3) K-Nearest Neighbors

Suppose we know the locations (co-ordinates) and the disease types of 10 patients.

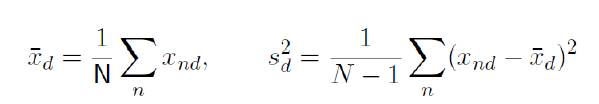
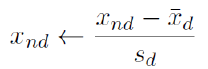
Disease type 1: {(10, 49) , (−12, 38) , (−9, 47)}

Disease type 2: {(29, 19) , (32, 31) , (37, 38)}

Disease type 3: {(8, 9) , (30, −28) , (−18, −19) , (−21, 12)}

1. Normalize the data

Compute the means and standard deviation for all the values using the below formula –

**Mean of** x̂ = (10+(-12)+(-9)+29+32+37+8+30+(-18)+(-21))/10

= 8.6

**Mean of** ŷ = (49+38+47+19+31+38+9+(-28)+(-19)+12)/10

= 19.6

Standard Deviation of Sdx2 = 1/9[(10-8.6) 2 + (-12-8.6) 2 +(-9-8.6) 2 +(29-8.6) 2 +(32-8.6) 2 +

(37-8.6) 2+(8- 8.6) 2 +(30-8.6) 2 +(-18-8.6) 2 +(-21-8.6) 2]

Sdx2 = 505.377

Sdx **=** 22.48

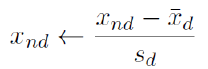
Standard Deviation of Sdy2 = 1/9[(49-19.6) 2 + (38-19.6) 2 +(47-19.6) 2 +(19-19.6) 2 +(31-19.6) 2 +

(38-19.6) 2 +(9-19.6) 2 +(-28-19.6) 2 + (-19-19.6) 2 + (12-19.6) 2]

Sdy2 = 705.377

Sdy= 26.55

Finding the normalized values of x and y using the formula



For X co-ordinates

x1 = (10 – 8.6) / 22.48= 0.062

x2 = (– 12 – 8.6) / 22.48= - 0.916

x3 = (– 9 – 8.6) / 22.48= - 0.782

x4 =( 29 – 8.6) / 22.48 = 0.907

x5 = (32 – 8.6) / 22.48= 1.040

x6 = (37 – 8.6 ) / 22.48= 1.263

x7 = (8 – 8.6 ) / 22.48 = -0.026

x8 = (30 – 8.6 ) / 22.48= 0.951

x9 = (– 18 – 8.6) / 22.48= - 1.183

x10 = (– 21 – 8.6) / 22.48= - 1.316

Normalized value of test data:

xtest ( x10)= (9 – 8.6) / 22.48

= 0.0177

For Y co-ordinates

y1 = (49 – 19.6) / 26.55= 1.107

y2 = (38 – 19.6)/ 26.55= 0.693

y3 = (47 – 19.6)/ 26.55= 1.032

y4 = (19 – 19.6)/ 26.55= -0.02

y5 = (31 – 19.6) / 26.55= 0.429

y6 = (38 – 19.6)/ 26.55= 0.693

y7 = (9 – 19.6) / 26.55= -0.399

y8 = (– 28 – 19.6) / 26.55= -1.792

y9 = (– 19 – 19.6) / 26.55= -1.453

y10 = (12 – 19.6)/ 26.55= -0.286

Normalized value of test data:

ytest (y11)=(18 – 19.6) / 26.55

=-0.060

Values in table format

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Disease Types** | **x** | **y** | **Normalized value of x** | **Normalized value of y** |
| Disease Type 1 | 10 | 49 | 0.062 | 1.107 |
| Disease Type 1 | -12 | 38 | -0.916 | 0.693 |
| Disease Type 1 | -9 | 47 | -0.782 | 1.032 |
| Disease Type 2 | 29 | 19 | 0.907 | -0.022 |
| Disease Type 2 | 32 | 31 | 1.040 | 0.429 |
| Disease Type 2 | 37 | 38 | 1.263 | 0.693 |
| Disease Type 3 | 8 | 9 | -0.026 | -0.399 |
| Disease Type 3 | 30 | -28 | 0.951 | -1.792 |
| Disease Type 3 | -18 | -19 | -1.183 | -1.453 |
| Disease Type 3 | -21 | 12 | -1.316 | -0.286 |

## By using the below formula we can find Euclidean distance and Manhattan distance of training and test data values

**Euclidean distance** (L2):√(x2-x1)2 + (y2-y1)2

**Manhattan distance** (L1):| (x2-x1) | + | (y2-y1) |

Euclidean and Manhattan is calculated between the normalized values of training and test data and it is as shown below -

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Type** | **x** | **y** | **Normalized value of x** | **Normalized value of y** | **Euclidean distance**  **(L2)** | **Manhattan distance**  **(L1)** |
| Disease Type 1 | 10 | 49 | 0.062 | 1.107 | 1.168 | 1.2115 |
| Disease Type 1 | -12 | 38 | -0.916 | 0.693 | 1.1996 | 1.6869 |
| Disease Type 1 | -9 | 47 | -0.782 | 1.032 | 1.3537 | 1.8919 |
| Disease Type 2 | 29 | 19 | 0.907 | -0.022 | 0.8901 | 0.9275 |
| Disease Type 2 | 32 | 31 | 1.040 | 0.429 | 1.33 | 1.5515 |
| Disease Type 2 | 37 | 38 | 1.263 | 0.693 | 1.4553 | 1.9985 |
| Disease Type 3 | 8 | 9 | -0.026 | 0.399 | 0.3416 | 0.382 |
| Disease Type 3 | 30 | -28 | 0.951 | -1.792 | 1.9672 | 2.6651 |
| Disease Type 3 | -18 | -19 | -1.183 | -1.453 | 1.8389 | 2.5935 |
| Disease Type 3 | -21 | 12 | -1.136 | -0.286 | 1.3526 | 1.5628 |

1. To predict the new patients i.e. (9,18) disease type when k=1

Recollecting normalized value of new patient

xtest(xknearest)= 0.0177 ; ytest (yknearest) =-0.0602

Euclidean value When K=1

Here we consider the least Euclidean distance (L2) i.e. 0.341which is the value of the (8,9) co-ordinates. We can say that the test date (9,18) belongs to **Disease Type 3** as it is nearest to the training data(8,9)

When K=3

Considering the top 3 minimum Euclidean distances(L2) are as shown below –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease Type** | **x** | **y** | **Normalized value of x** | **Normalized of y** | **Euclidean distance**  **(L2)** |
| Disease Type 2 | 29 | 19 | 0.907 | -0.022 | 0.8901 |
| Disease Type 2 | 32 | 31 | 1.040 | 0.429 | 1.33 |
| Disease Type 3 | 8 | 9 | -0.026 | 0.399 | 0.3416 |

The given new patient i.e. test data(9,18) belongs to Disease Type 2 as 2 out of 3 Euclidean distance belongs to **Disease type 2**

Manhattan distance When K=1

Here we consider the least Manhattan distance (L1) i.e. 0.382 which is the value of the (8,9) co-ordinates. We can say that the test date (9,18) categorized to **Disease Type 3**

When K=3

Considering the top 3 minimum Manhattan distances(L1) are as shown below –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease Type** | **x** | **y** | **Normalized value of x** | **Normalized value of y** | **Manhattan distance**  **(L1)** |
| Disease Type 1 | 10 | 49 | 0.062 | 1.107 | 1.2115 |
| Disease Type 2 | 29 | 19 | 0.907 | -0.022 | 0.9275 |
| Disease Type 3 | 8 | 9 | -0.026 | 0.399 | 0.382 |

As we can see that the 2 minimum Manhattan distance belongs to different categories/types of disease. In such scenario, we consider the smallest distance among the values i.e. 0.382 and the test date (9, 18) can be categorized to **Disease type 3**.

## Comparing the results obtained between the above 4 predictors:

When K=1 for Euclidean (L2) and Manhattan (L1) metrics, the new patient i.e. test data (9,18) has the smallest distance with the training data (8,9) and this data belongs to **Disease type 3**.

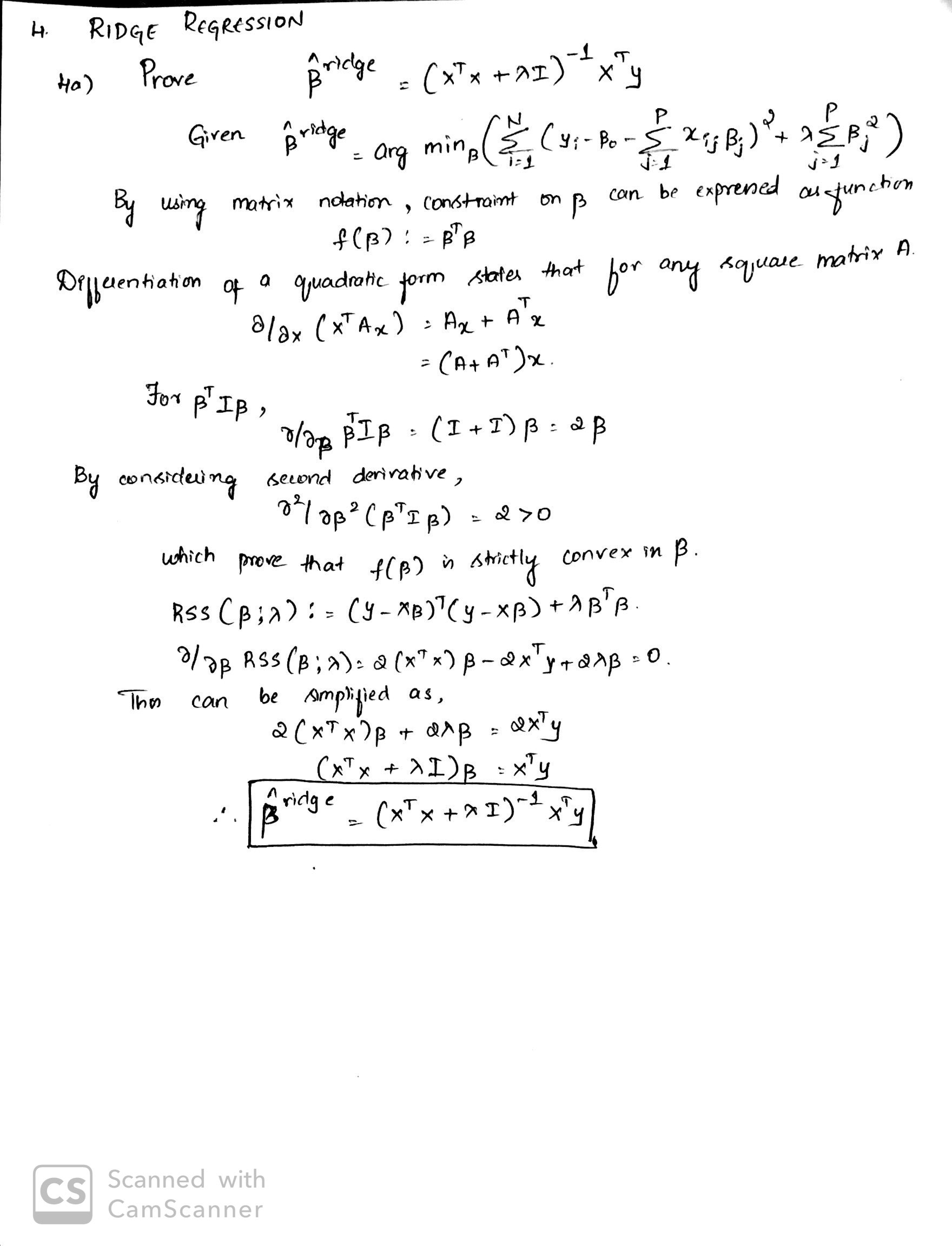
When K=3 for the L2 metric, we have computed the distance between test data(9,18) with the training data , and found that (9,18) has nearest distance with (8,9), (29,19) and (32,31) . Here the training data (29,19) and (32,31) belongs to DT2 and it is the maximum nearest neighbors for the test data(9,18), so it is classified as **Disease Type 2.**

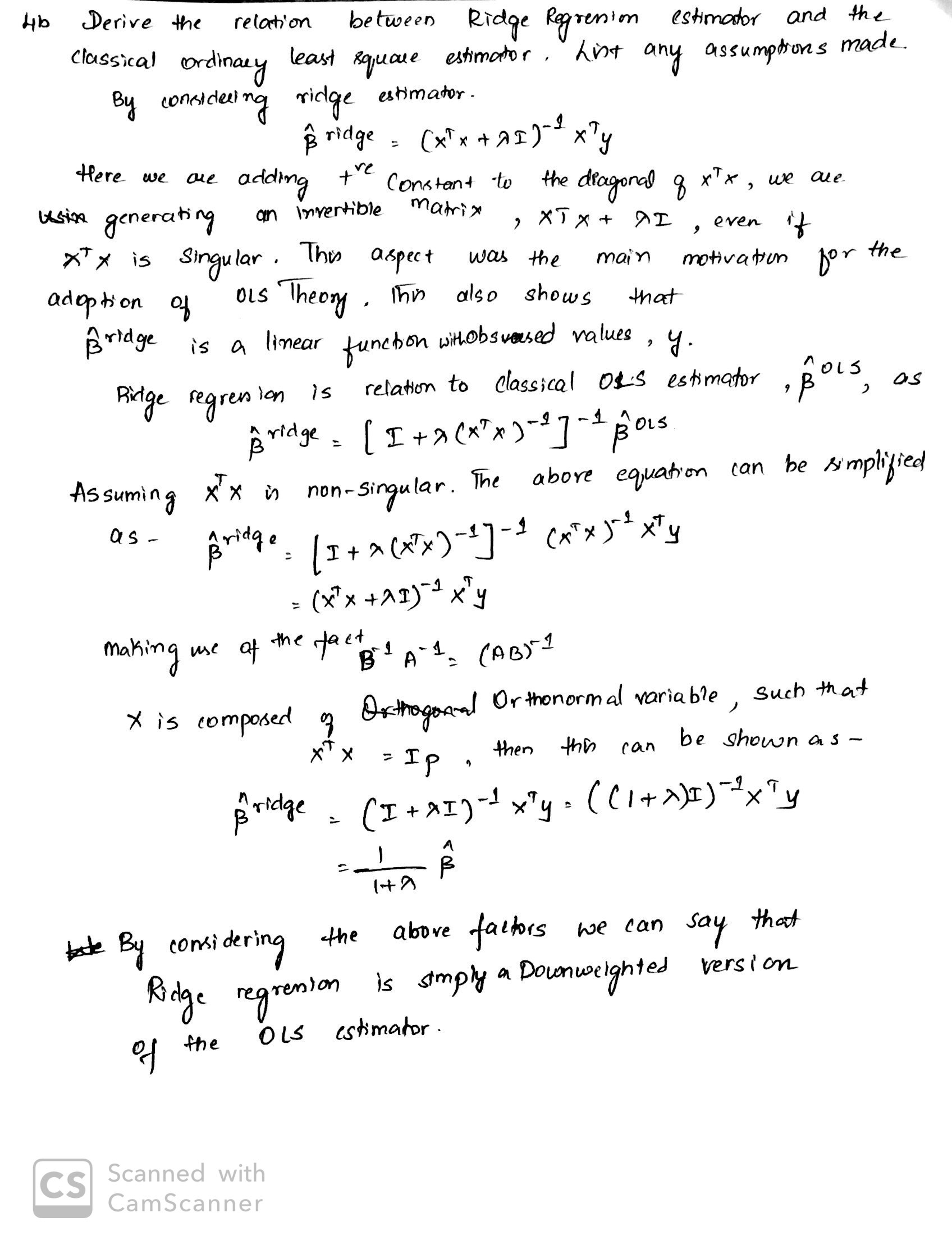
When K=3 in L1 metric, for the test data(9,18) its nearest distance are (8,9), (29,19) and (10,49) and all these training points belongs to different disease types and there is tie in choosing as to in which (9,18) should be categorized. In such cases, we have to consider the smallest L1 distance and it can be seen that (8,9) has the minimum distance among them. From this we can say that test data(9,18) can be classified as **Disease type 3**.

4)Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a l2 penalty. The ridge coefficients minimize a penalized residual sum of squares, [Equation 3.41, page 82, ESL book ]







# 5) Programming

# Data Analytics

Please write programs for the exercise problem 10 of the chapter 2, **ISLR** book (page 56).

# 

1. How many rows are in this data set? How many columns? What do the rows and columns represent?

A: In the Boston dataset, there are 506 rows and 14 columns.

Rows represent towns in Boston and Column represents different aspects that influence the house-price.

Attribute Information (in order):

- CRIM per capita crime rate by town

- ZN proportion of residential land zoned for lots over 25,000 sq.ft.

- INDUS proportion of non-retail business acres per town

- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

- NOX nitric oxides concentration (parts per 10 million)

- RM average number of rooms per dwelling

- AGE proportion of owner-occupied units built prior to 1940

- DIS weighted distances to five Boston employment centers

- RAD index of accessibility to radial highways

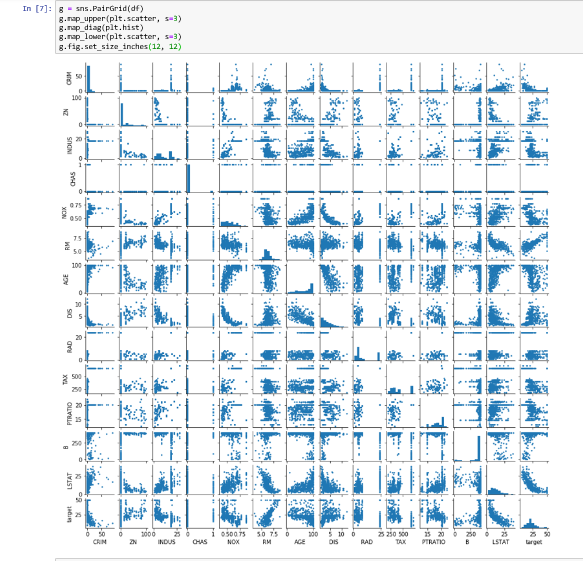
- TAX full-value property-tax rate per $10,000

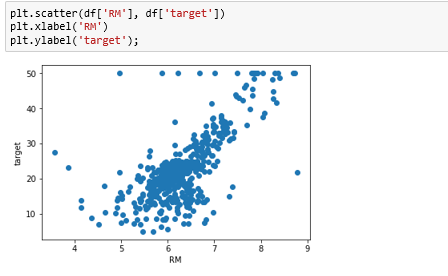
- PTRATIO pupil-teacher ratio by town

- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

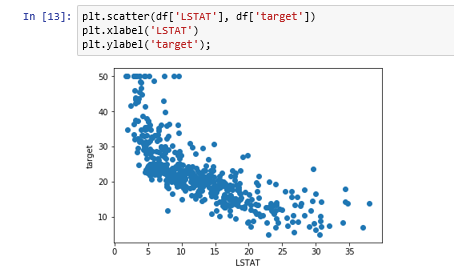
- LSTAT % lower status of the population

- MEDV Median value of owner-occupied homes in $1000's

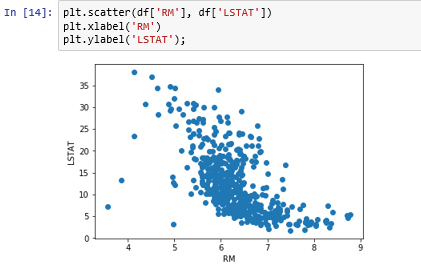
1. Make some pair wise scatter plots of the predictors (columns) in this data set. Describe your findings.



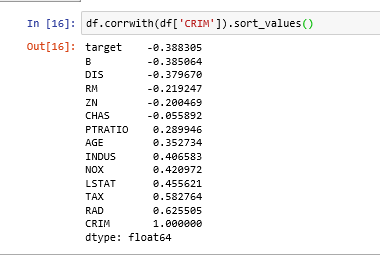
1. RM and Target Findings- there *seems to exist a positive linear relationship between RM and target*This is expected as RM is the number of rooms (more space, higher price)



1. LSAT and Target Findings- they seem to have a negative non-linear relationship This is expected as LSTAT is the percent of lower status people (lower status, lower incomes, cheaper houses)

**

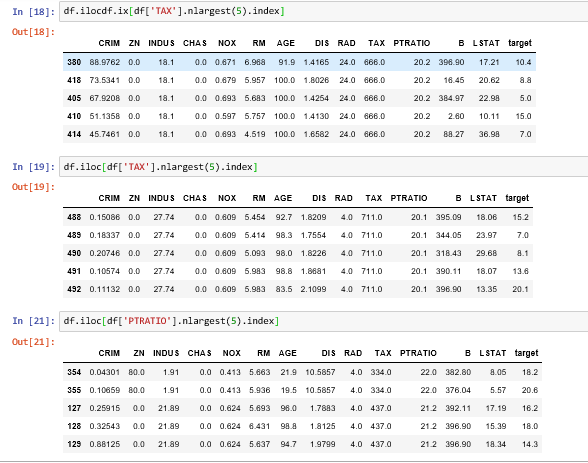
1. RM and LSTAT Findings- *It seems to exist a negative non-linear relationship between LSTAT and RM*It makes sense since people with less money (higher LSTAT) can't afford bigger houses (high RM)
2. Are any of the predictors associated with per capita crime rate? If so, explain the relationship



Yes, there are 14 predictors associated with per capita crime rates.

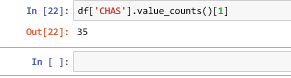
* + More crimes tend to occur around older homes (age increases)
  + More crimes tend to occur when there is easy accessibility to the highway roads.
  + More crimes occur in places closer to the occupational area.
  + Low crimes tend to happen with low ptratio and vice-versa
  + When there is high tax collected and the places are prone to crime.

d)Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



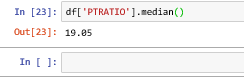
The 5 towns shown in CRIM table are particularly high, few towns has crime rates above 40 and some have even 80.All the towns shown in the TAX table have maximum TAX level above 700 PTRATIO table shows towns with high pupil-teacher ratios but not so uneven.

e) How many of the suburbs in this data set bound the Charles river?



In the mentioned Dataset there are 35 suburbs that are bound to Charles river.

f) What is the median pupil-teacher ratio among the towns in this data set?



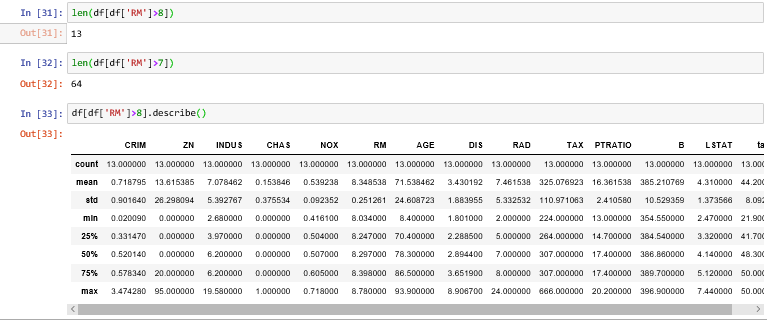
19.05 is the median pupil-teacher ratio among the towns in the data set.

g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.



The suburb with the lowest median value is 398. Relative to the other towns, this suburb has high CRIM, ZN below quantile 75%, above mean INDUS, does not bound the Charles river, above mean NOX, RM below quantile 25%, maximum AGE, DIS near to the minimum value, maximum RAD, TAX in quantile 75%, PTRATIO as well, B maximum and LSTAT above quantile 75%.

h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling



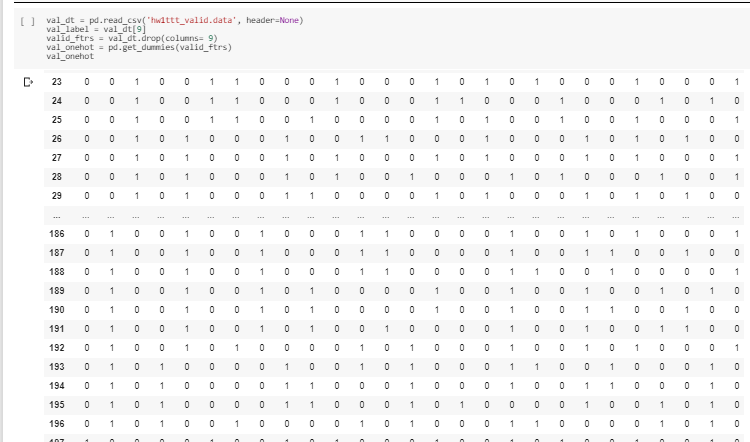
64 of the suburbs average more than seven rooms per dwelling. 13 of the suburbs average more than eight rooms per dwelling.

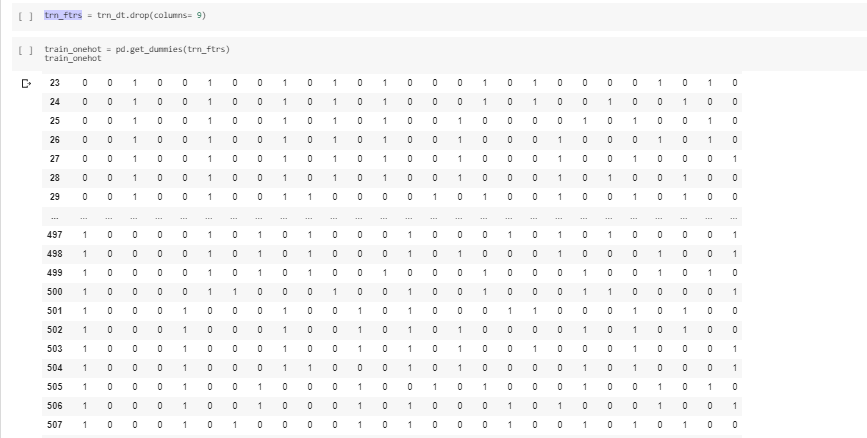
From the dataset we can infer the crime rates are high when there are eight rooms per dwelling with high median value and low lstat.

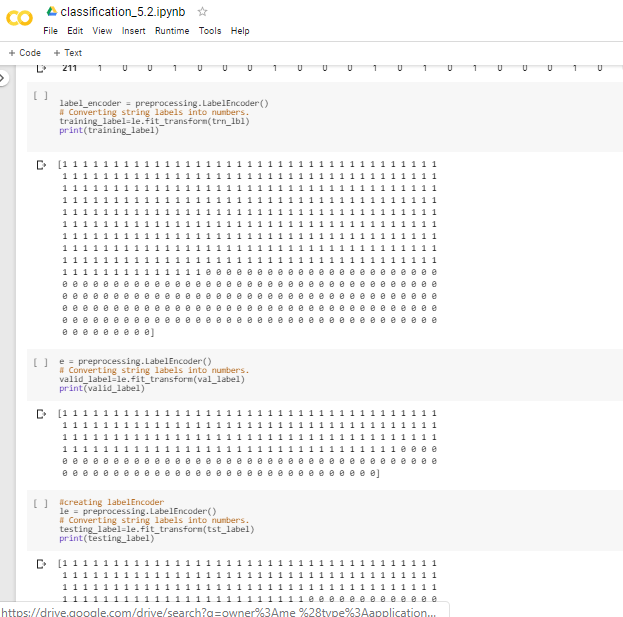
**5.2 Classification**

1. **Data- Preprocessing**

Snapshot of data preprocessing :

****



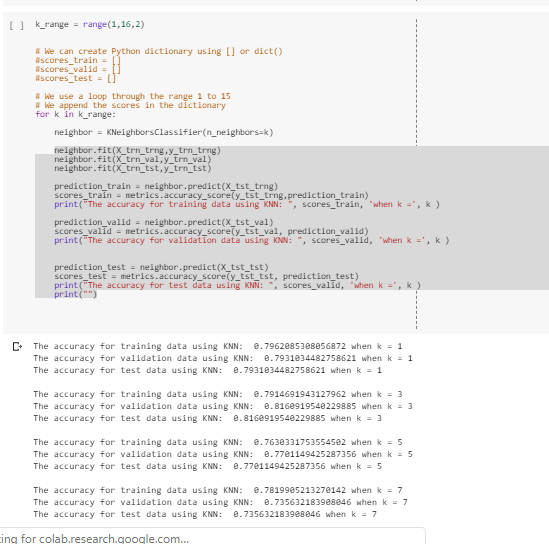


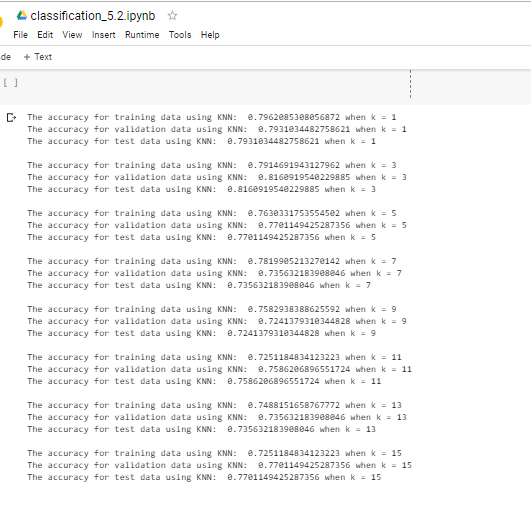


B) Performance Comparison

The below snapshot are the accuracy predicted for KNN, logistic regression , and Naïve Bayes algorithms .

* Accuracy predicted for KNN algorithm for Training, validation , and testing data

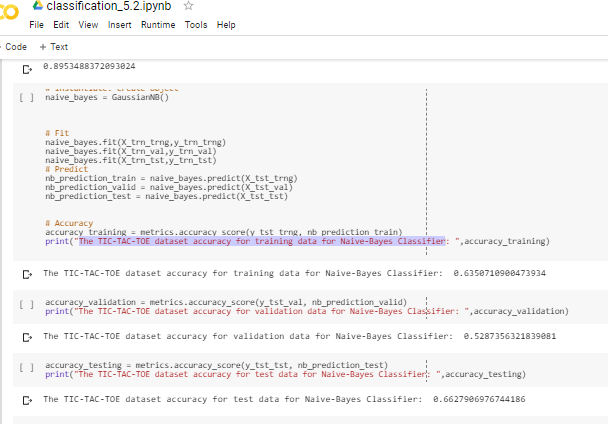




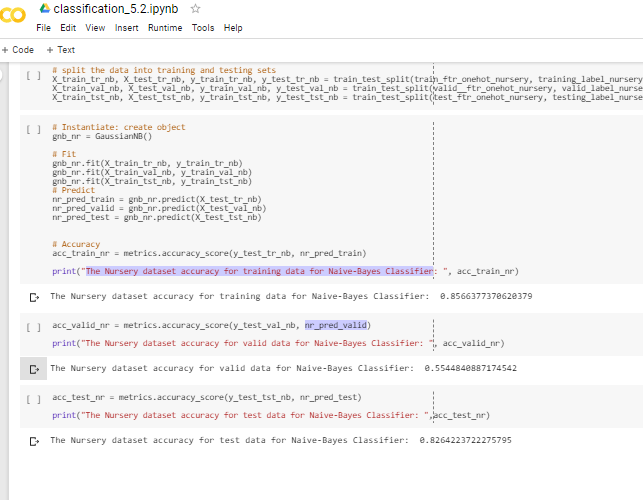
* Accuracy predicted for Logistic Regression algorithm for Training, validation , and testing data



* Accuracy predicted for Naïve Bayes algorithm for Training, validation , and testing data for TIC-TAC-TOC dataset



* Accuracy predicted for Naïve Bayes algorithm for Training, validation , and testing data for Nursery dataset



Comparing the accuracy of two datasets using Naïve Bayes Algorithm

Here we are comparing TIC-TAC-TOE and Nursery datasets, as the nature of both datasets is different it is difficult to compare the accuracy between them.

Accuracy for both datasets is as follows –

TIC-TAC-TOE

The TIC-TAC-TOE dataset accuracy for training data for Naive-Bayes Classifier: 0.6350710900473934

The TIC-TAC-TOE dataset accuracy for validation data for Naive-Bayes Classifier: 0.5287356321839081

The TIC-TAC-TOE dataset accuracy for test data for Naive-Bayes Classifier: 0.6627906976744186

Nursery Dataset

The Nursery dataset accuracy for training data for Naive-Bayes Classifier: 0.8566377370620379

The Nursery dataset accuracy for valid data for Naive-Bayes Classifier: 0.5544840887174542

The Nursery dataset accuracy for test data for Naive-Bayes Classifier: 0.826422372227579

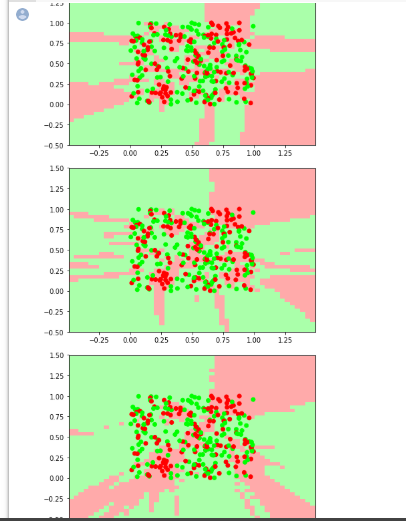
C) DECISION BOUNDARY

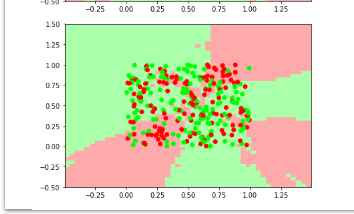
Data-preprocessing for the hw1boundary features.csv and hw1boundary labels.csv file .



Applied KNN for [1,5,15,25] values to plot the Decision Boundary







REFERENCES-

<https://stattrek.com/multipleregression/multicollinearity.aspx#targetText=When%20one%20independent%20variable%20is,regression%20coefficients%20does%20not%20exist.>

<https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/impact-of-removing-outliers-on-regression-lines>

<https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html>

<https://www.teachme.codes/docs/Classifiers/Logistic/Logistic/#1-plot-classification-probability-source>

<https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>

<https://www.ritchieng.com/machine-learning-k-nearest-neighbors-knn/>

<https://www.ritchieng.com/machine-learning-gaussian-naive-bayes/>

<https://botlnec.github.io/islp/sols/chapter2/exercise10/>

<https://towardsdatascience.com/mse-and-bias-variance-decomposition-77449dd2ff55>

<https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>