BAUHAUS UNIVERSITÄT WEIMAR

INTRODUCTION TO MACHINE LEARNING

ASSIGNMENT 2

SUBMITTED BY

Anjana Muraleedharan Nair (125512)

Isabel Maria Binu (125514)

Vishal Sanjay Shivam (125353)

Sharat Anand (125404)

Exercise 1 : Machine Learning Basics

Solution:

a) *x* - a single feature
x - feature vector
X - feature space = domain of the feature vectors
X - multiset of feature vectors (world representation)
X - multivariate random variable whose instances are feature vectors(a random variable in a sample space)

b) The collection of all possible legal hypotheses is known as hypothesis space. This is a set of data that the machine would use to identify the best (and only one) description of the target function or outputs.
Hypothesis space H of linear regression with p features:

$$H = \{ h(x) = w_0 + w_1x_1 + w_2x_2 \ldots\ldots + w_px_p \}$$

where x_1, x_2,..x_p are the features and w_0, w_1, w_2,..,w_p are the parameters of the linear regression model.

c) Bayes error represents the error that is produced by the Bayes Classifier which is a model that maps each feature vector x to the highest-probability class c according to the true joint probability distribution p(c, x) that generates the data. This occurs when the class assignment depends on additional (unobserved) features not recorded in x. The Bayes error represents the theoretical minimal error that any classifier could achieve on average for a given dataset.

d) Bayes error is referred to as irreducible error, or unavoidable error. It can be reduced by refining the model or classifier. Adding more complexity to the model and obtaining important features.

e) Example of a dataset with noise:

D ={(x1,c1),(x1,c2), (x3,c3).....,}

here x represents the feature vector and c represents the identified class label.
For example,

D ={(x1,0),(x1,1), (x3,0),(x4,1).....,}

In the above given dataset, some of the datasets are labeled as 0 and some of them as 1. Some of the datasets may be misclassified due to noise in the data which may be different from the value we are expecting.

f) To create a stratified dataset sample D2,tr of D2 with |D2,tr | = 6: we have to make sure that the class distribution should remain the same. In the original dataset the distribution of c1, c2,c3 are given as: 2,6 and 4 respectively.

To get |D2,tr | = 6, we can select a instances of classes c1, c2, c3 are:

c1 =1, c2 =3, c3 = 2.

Hence the stratified dataset ={(x1,c1), (x2, c2),(x3, c3),(x4, c2),(x5, c2),(x3,c3)}.

Exercise 2 : Probabilistic Foundation of the True Misclassification Rate

Solution:

a)

X	c	P(X,c)
$\mathbf{(0,0)^T}$	0	0.1
$\mathbf{(0,0)^T}$	1	0
$\mathbf{(0,0)^T}$	0	0.3
$\mathbf{(0,1)^T}$	1	0.3
$\mathbf{(1,0)^T}$	0	0.1
$\mathbf{(1,0)^T}$	1	0.2

b)

x	y*(x)
$(0,0)^T$	0
$(0,1)^T$	0 or 1
$(1,0)^T$	1

c) Misclassification rate Err* = 0.5

Exercise 3 : Evaluating Effectiveness

Solution:

a)

	x1	x2	c	L+	L-
X1	1	1	1	0	1
X2	-1	1	-1	0	1
X3	1	1	-1	1	0
X4	1	1	1	0	1
X5	1	1	-1	1	0
X6	-1	1	1	1	0
X7	1	1	-1	1	0
X8	-1	-1	-1	0	1
X9	1	1	1	0	1
X10	1	-1	1	0	1

Since L+<= L- ω=1

Err(y(),D test) = 4/10=0.4

b) D Test={x8,x9,x10}; π=1 ;D tr = D\D test

Since L+ > L- ω=-1

Err(y'(),Dtest)= L-/3 =3/3=1

Exercise 4 : Receiver Operating Characteristic (ROC)

Solution:

a) True positive: A true positive is an outcome where the model correctly predicts the positive class. A classifier correctly identifies a spam email as spam.

False positive: A false positive is an outcome where the model incorrectly predicts the positive class. A classifier incorrectly identifies a non-spam email as spam.

False negative: A false negative is an outcome where the model incorrectly predicts the negative class. A classifier incorrectly identifies a spam email as non-spam.

True Negative: A true negative is an outcome where the model correctly predicts the negative class. A classifier correctly identifies a non-spam email as non-spam.

False positive rate: The false positive rate is the fraction of all negative test results that still produce positive test outcomes. The proportion of non-spam emails that are incorrectly classified as spam. It is calculated as False positive / (False positive + True Negative). It is also called Fallout.

True positive rate: The true positive rate is the percentage of real positive cases that the model correctly detected or categorized as positive. The proportion of spam emails that are correctly classified as spam. It can be calculated as True Positive / (True Positive + False Negative). It is also called Sensitivity.

b) The false positive rate is calculated by False positive / (False positive + True Negative). and true positive rate is calculated by True Positive / (True Positive + False Negative). Suppose we define the dataset as 0 and 1

- A classifier that classifies every email as spam
  In this scenario, True positive and False positive = 1 and True Negative and False Negative =0. False positive rate (FPR) = 1/(1+0) =1 and True positive rate (TPR) =1/(1+0)=1. Hence, FPR and TPR =1.

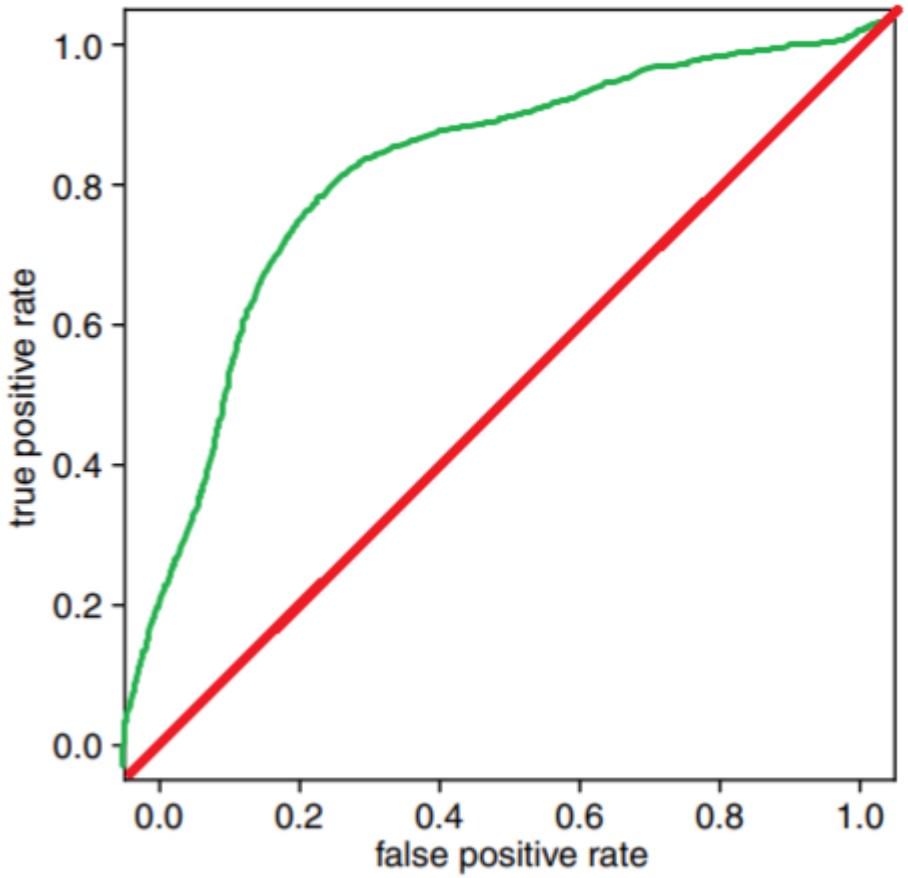
- A classifier that classifies every mail as not spam
  TP =0, FP =0, TN =1 and FN =1. FPR =0/(0+1) = 0 and TPR=0/(0+1) = 0
  Hence, FPR and TPR =0.

- A classifier that classifies every mail correctly
  TP =1, FP =0, FN =0, TN =1. FPR =0/(0+1) = 0 and TPR=1/(0+1) = 1
  Hence, FPR = 0 and TPR = 1.

- A classifier that classifies every mail incorrectly
  TP =0, FP =1, FN =1, TN =0. FPR =1/(1+0) = 0 and TPR=0/(0+1) = 0
  Hence, FPR = 1 and TPR = 0.

- A classifier that classifies every mail randomly with equal class probability
  TP =0.5, FP =0.5, FN =0.5, TN =0.5. FPR =0.5/(0.5+0.5) = 0.5 and TPR=0.5/(0.5+0.5) = 0.5
  Hence, FPR = 0.5 and TPR = 0.5.

c)



The ROC is the plot of true positive rate and false positive rate. The green line corresponds to y1 and the red line corresponds to y2. For y1, corresponding to spam examples (c=1), as the threshold value increases, there will be change in TPR and FPR. In the case of y2, corresponding to non-spam (c=0), it constantly changes with the values.

d) We choose the green color classifier (y1) because it is more close to the True positive rate.

Exercise 5 : Linear Regression

Solution: a) $y(x) = w_0 + w_1 . x$,

$$RSS(w0, w1) = \sum_{i=1}^{n} (yi - w0 - w1 \cdot xi)2$$

where w_1 & w_0 are the linear regression weights.

By substituting the values of age for x_i and stopping distance for y_i, we get

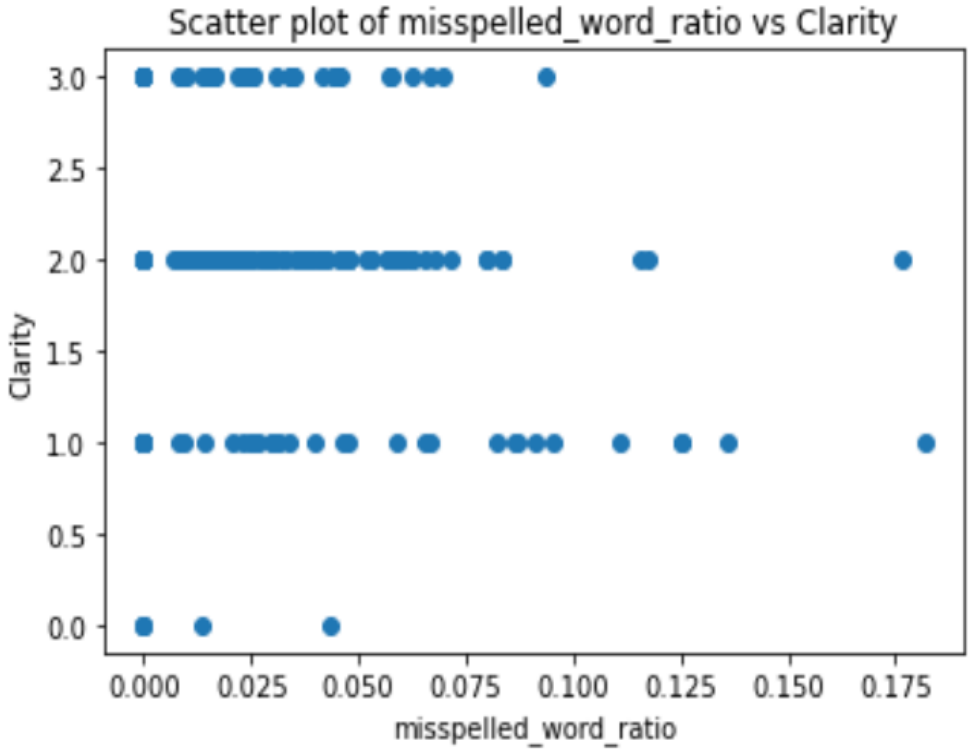
$w_1 = 10.58$ and $w_0 = -7.225$

b) $y(x) = w_0 + w_1 . x$

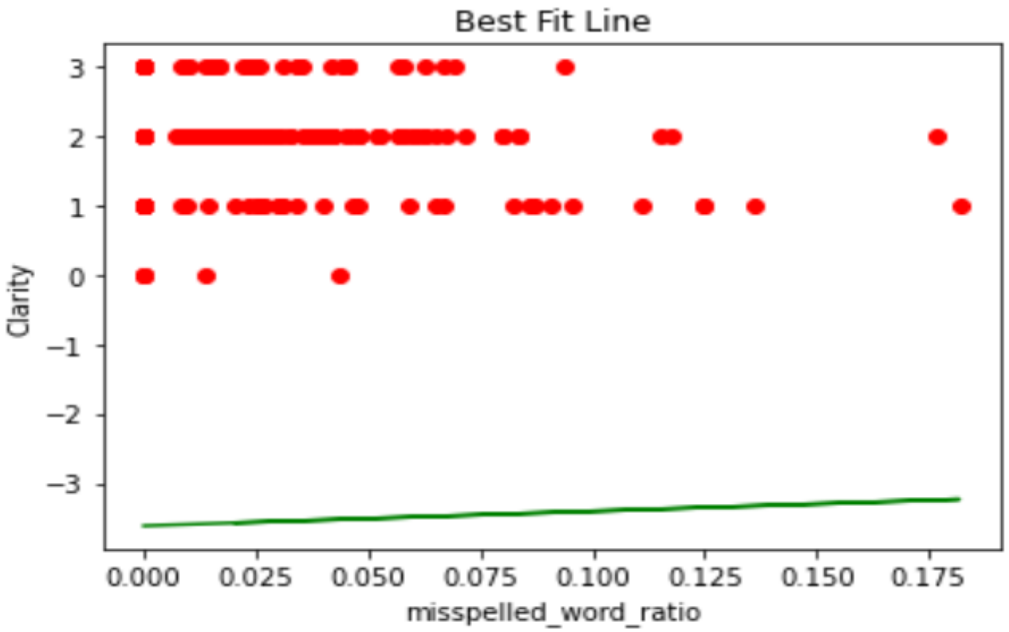
Hence for the age of 15, the stopping distance $y = 151.475$

Exercise 6: Basic Data Analysis and Linear Regression
Solution:

a) The scatter plot between the feature misspelled_word_ratio and clarity is plotted.



b)



c)

```
rss=0
for i in range(len(feature_data)):
    rss+= (clarity_dataset["clarity"][i]-(w0 +w1* feature_data["misspelled_word_ratio"][i]))**2
rss
```

```
8091.909212169797
```