**CSE3506 Essentials of Data Analytics**

Name : **Sparsh Raj**

Reg. No. : **19BPS1028**

Lab Exercise: 6 – Naïve Bayes Classifier

(Dataset: Titanic - Machine Learning from Disaster)

**Objective:**

Perform Naïve Bayes Classifier, build the model from the training dataset and use it to predict on the testing set.

**Methods:**

# Loading training and test data sets onto R

data<-read.csv("train.csv")

str(data)

## 'data.frame': 891 obs. of 12 variables:

## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...

## $ Sex : chr "male" "female" "female" "female" ...

## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...

## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

## $ Cabin : chr "" "C85" "" "C123" ...

## $ Embarked : chr "S" "C" "S" "S" ...

# Splitting data into training and testing set

data2<-sample(2,nrow(data),replace=T,prob = c(0.75,0.25))

train<-data[data2==1,]

test<-data[data2==2,]

head(train)

## PassengerId Survived Pclass

## 2 2 1 1

## 3 3 1 3

## 4 4 1 1

## 5 5 0 3

## 6 6 0 3

## 7 7 0 1

## Name Sex Age SibSp Parch

## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1 0

## 3 Heikkinen, Miss. Laina female 26 0 0

## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1 0

## 5 Allen, Mr. William Henry male 35 0 0

## 6 Moran, Mr. James male NA 0 0

## 7 McCarthy, Mr. Timothy J male 54 0 0

## Ticket Fare Cabin Embarked

## 2 PC 17599 71.2833 C85 C

## 3 STON/O2. 3101282 7.9250 S

## 4 113803 53.1000 C123 S

## 5 373450 8.0500 S

## 6 330877 8.4583 Q

## 7 17463 51.8625 E46 S

head(test)

## PassengerId Survived Pclass Name Sex Age SibSp

## 1 1 0 3 Braund, Mr. Owen Harris male 22 1

## 8 8 0 3 Palsson, Master. Gosta Leonard male 2 3

## 20 20 1 3 Masselmani, Mrs. Fatima female NA 0

## 23 23 1 3 McGowan, Miss. Anna "Annie" female 15 0

## 31 31 0 1 Uruchurtu, Don. Manuel E male 40 0

## 43 43 0 3 Kraeff, Mr. Theodor male NA 0

## Parch Ticket Fare Cabin Embarked

## 1 0 A/5 21171 7.2500 S

## 8 1 349909 21.0750 S

## 20 0 2649 7.2250 C

## 23 0 330923 8.0292 Q

## 31 0 PC 17601 27.7208 C

## 43 0 349253 7.8958 C

colnames(train) <- tolower(colnames(train))

colnames(test) <- tolower(colnames(test))

# Data Cleaning

clean\_variable <- **function**(df, variable, group, lookup\_table){

df[variable] <- apply(df[, c(variable, group)], 1, FUN=**function**(x) {**if** (is.na(x[1]) | x[1]==0) **return**(lookup\_table[lookup\_table[group]==x[2], variable]) **else** **return**(x[1])})

**return**(df)

}

fare\_pclass <- aggregate(fare ~ pclass, data=train, FUN=mean)

train <- clean\_variable(train, "fare", "pclass", fare\_pclass)

test <- clean\_variable(test, "fare", "pclass", fare\_pclass)

# Naive Bayes Classifier

**library**(naivebayes)

## Warning: package 'naivebayes' was built under R version 4.1.2

## naivebayes 0.9.7 loaded

**library**(caret)

## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: ggplot2

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.1.2

train$survived<-as.factor(train$survived)

test$survived<-as.factor(test$survived)

model<-naive\_bayes(survived ~ pclass + sex + age + sibsp + parch + fare, data=train)

summary(model)

##

## ================================== Naive Bayes ==================================

##

## - Call: naive\_bayes.formula(formula = survived ~ pclass + sex + age + sibsp + parch + fare, data = train)

## - Laplace: 0

## - Classes: 2

## - Samples: 674

## - Features: 6

## - Conditional distributions:

## - Bernoulli: 1

## - Gaussian: 5

## - Prior probabilities:

## - 0: 0.6142

## - 1: 0.3858

##

## ---------------------------------------------------------------------------------

pred<-predict(model,train)

## Warning: predict.naive\_bayes(): more features in the newdata are provided as

## there are probability tables in the object. Calculation is performed based on

## features to be found in the tables.

head(pred)

## [1] 1 1 1 0 0 0

## Levels: 0 1

pred<-as.factor(pred)

n<-length(test$survived)

pred<-pred[0:n]

length(test$survived)

## [1] 217

confusionMatrix(test$survived, pred)

## Confusion Matrix and Statistics

##

## Reference

## Prediction 0 1

## 0 93 42

## 1 54 28

##

## Accuracy : 0.5576

## 95% CI : (0.4888, 0.6248)

## No Information Rate : 0.6774

## P-Value [Acc > NIR] : 0.9999

##

## Kappa : 0.0313

##

## Mcnemar's Test P-Value : 0.2616

##

## Sensitivity : 0.6327

## Specificity : 0.4000

## Pos Pred Value : 0.6889

## Neg Pred Value : 0.3415

## Prevalence : 0.6774

## Detection Rate : 0.4286

## Detection Prevalence : 0.6221

## Balanced Accuracy : 0.5163

##

## 'Positive' Class : 0

##