Prepare a classification model using Naive Bayes for salary data

> library(e1071)

> library(corrplot)

corrplot 0.84 loaded

> library(gmodels)

> salary\_train <-read.csv(file.choose())

> View(salary\_train)

> str(salary\_train)

'data.frame': 30161 obs. of 14 variables:

$ age : int 39 50 38 53 28 37 49 52 31 42 ...

$ workclass : Factor w/ 7 levels " Federal-gov",..: 6 5 3 3 3 3 3 5 3 3 ...

$ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...

$ educationno : int 13 13 9 7 13 14 5 9 14 13 ...

$ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...

$ occupation : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...

$ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...

$ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...

$ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...

$ capitalgain : int 2174 0 0 0 0 0 0 0 14084 5178 ...

$ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...

$ hoursperweek : int 40 13 40 40 40 40 16 45 50 40 ...

$ native : Factor w/ 40 levels " Cambodia"," Canada",..: 38 38 38 38 5 38 22 38 38 38 ...

$ Salary : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...

> salary\_test <- read.csv(file.choose())

> View(salary\_test)

> str(salary\_test)

'data.frame': 15060 obs. of 14 variables:

$ age : int 25 38 28 44 34 63 24 55 65 36 ...

$ workclass : Factor w/ 7 levels " Federal-gov",..: 3 3 2 3 3 5 3 3 3 1 ...

$ education : Factor w/ 16 levels " 10th"," 11th",..: 2 12 8 16 1 15 16 6 12 10 ...

$ educationno : int 7 9 12 10 6 15 10 4 9 13 ...

$ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 3 3 5 3 5 3 3 3 ...

$ occupation : Factor w/ 14 levels " Adm-clerical",..: 7 5 11 7 8 10 8 3 7 1 ...

$ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 4 1 1 1 2 1 5 1 1 1 ...

$ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 3 5 5 3 5 5 5 5 5 5 ...

$ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 2 2 1 2 2 2 ...

$ capitalgain : int 0 0 0 7688 0 3103 0 0 6418 0 ...

$ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...

$ hoursperweek : int 40 50 40 40 30 32 40 10 40 40 ...

$ native : Factor w/ 40 levels " Cambodia"," Canada",..: 38 38 38 38 38 38 38 38 38 38 ...

$ Salary : Factor w/ 2 levels " <=50K"," >50K": 1 1 2 2 1 2 1 1 2 1 ...

> salary <- rbind(salary\_train,salary\_test)

> str(salary)

'data.frame': 45221 obs. of 14 variables:

$ age : int 39 50 38 53 28 37 49 52 31 42 ...

$ workclass : Factor w/ 7 levels " Federal-gov",..: 6 5 3 3 3 3 3 5 3 3 ...

$ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...

$ educationno : int 13 13 9 7 13 14 5 9 14 13 ...

$ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...

$ occupation : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...

$ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...

$ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...

$ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...

$ capitalgain : int 2174 0 0 0 0 0 0 0 14084 5178 ...

$ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...

$ hoursperweek : int 40 13 40 40 40 40 16 45 50 40 ...

$ native : Factor w/ 40 levels " Cambodia"," Canada",..: 38 38 38 38 5 38 22 38 38 38 ...

$ Salary : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...

#Creating dummies

> level\_work <- levels(salary$workclass)

> View(salary)

> level\_edu <- levels(salary$education)

> level\_mari <- levels(salary$maritalstatus)

> level\_occ <- levels(salary$occupation)

> level\_rel <- levels(salary$relationship)

> level\_race <- levels(salary$race)

> level\_sex <- levels(salary$sex)

> level\_native <- levels(salary$native)

> level\_salary <- levels(salary$Salary)

> salary$workclass <- as.integer(factor(salary$workclass,levels=c(level\_work),labels=c(1,2,3,4,5,6,7)))

> salary$education <- as.integer(factor(salary$education,levels = c(level\_edu),labels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)))

> salary$maritalstatus <- as.integer(factor(salary$maritalstatus,levels = c(level\_mari),labels=c(1,2,3,4,5,6,7) ))

> salary$occupation <- as.integer(factor(salary$occupation,levels = c(level\_occ),labels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)))

> salary$relationship <- as.integer(factor(salary$relationship,levels = c(level\_rel),labels = c(1,2,3,4,5,6)))

> salary$race <- as.integer(factor(salary$race,levels = c(level\_race),labels = c(1,2,3,4,5)))

> salary$sex <- as.integer(factor(salary$sex,levels = c(level\_sex),labels = c(1,2)))

> salary$native <- as.integer(factor(salary$native,levels = c(level\_native),labels = c(seq(1,40,1))))

> salary$Salary <- as.integer(factor(salary$Salary,levels = c(level\_salary),labels = c(1,2)))

#Normalising the data

> norm <- function(x){

return((x-min(x))/(max(x)-min(x)))

}

> normsalary <- as.data.frame(lapply(salary,norm))

> View(normsalary)

#Splitting data to test and train

> salary\_train\_norm <- normsalary[1:30161,]

> salary\_test\_norm <- normsalary[30162:45221,]

> head(salary\_train)

age workclass education educationno maritalstatus occupation relationship

1 39 State-gov Bachelors 13 Never-married Adm-clerical Not-in-family

2 50 Self-emp-not-inc Bachelors 13 Married-civ-spouse Exec-managerial Husband

3 38 Private HS-grad 9 Divorced Handlers-cleaners Not-in-family

4 53 Private 11th 7 Married-civ-spouse Handlers-cleaners Husband

5 28 Private Bachelors 13 Married-civ-spouse Prof-specialty Wife

6 37 Private Masters 14 Married-civ-spouse Exec-managerial Wife

race sex capitalgain capitalloss hoursperweek native Salary

1 White Male 2174 0 40 United-States <=50K

2 White Male 0 0 13 United-States <=50K

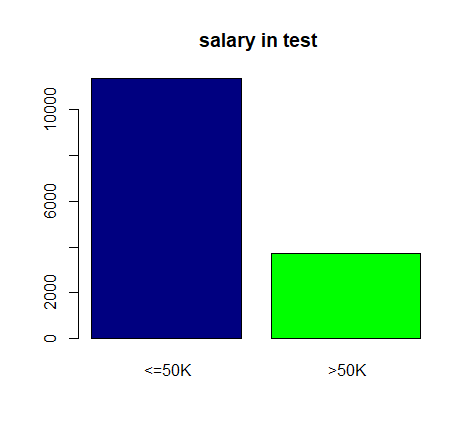
3 White Male 0 0 40 United-States <=50K

4 Black Male 0 0 40 United-States <=50K

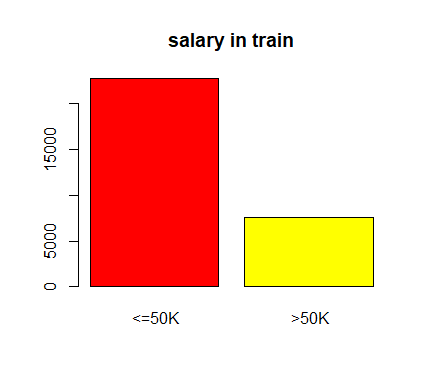
5 Black Female 0 0 40 Cuba <=50K

6 White Female 0 0 40 United-States <=50K

> barplot(table(salary\_test$Salary), main = " salary in test")



> barplot(table(salary\_train$Salary) ,main = "salary in train")



#Model building

> model1 <- naiveBayes(salary\_train\_norm,y=salary\_train$Salary)

> summary(model1)

Length Class Mode

apriori 2 table numeric

tables 14 -none- list

levels 2 -none- character

isnumeric 14 -none- logical

call 3 -none- call

> pred1 <- predict(model1,salary\_test$Salary)

There were 14 warnings (use warnings() to see them)

> mean(pred1==salary\_test$Salary)

[1] 0.7543161

**#Model Accuracy is 75.43%**

> table(salary\_test$Salary,pred1)

pred1

<=50K >50K

<=50K 11360 0

>50K 3700 0

> CrossTable(salary\_test$Salary,pred1)

Cell Contents

|-------------------------|

| N |

| N / Table Total |

|-------------------------|

Total Observations in Table: 15060

| pred1

salary\_test$Salary | <=50K | Row Total |

-------------------|-----------|-----------|

<=50K | 11360 | 11360 |

| 0.754 | |

-------------------|-----------|-----------|

>50K | 3700 | 3700 |

| 0.246 | |

-------------------|-----------|-----------|

Column Total | 15060 | 15060 |

-------------------|-----------|-----------|

#Model building with Laplace smoothing

> model2 <- naiveBayes(salary\_train\_norm,salary\_train$Salary,laplace = 1)

> summary(model2)

Length Class Mode

apriori 2 table numeric

tables 14 -none- list

levels 2 -none- character

isnumeric 14 -none- logical

call 4 -none- call

> pred2 <- predict(model2,salary\_test\_norm)

> mean(pred2==salary\_test$Salary)

[1] 0.9922311

#Model Accuracy is 99.22% After Laplace smoothing

> table(salary\_test$Salary,pred2)

pred2

<=50K >50K

<=50K 11290 70

>50K 47 3653

> CrossTable(salary\_test$Salary,pred2,prop.c = F,prop.t = F,prop.chisq = F)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

|-------------------------|

Total Observations in Table: 15060

| pred2

salary\_test$Salary | <=50K | >50K | Row Total |

-------------------|-----------|-----------|-----------|

<=50K | 11290 | 70 | 11360 |

| 0.994 | 0.006 | 0.754 |

-------------------|-----------|-----------|-----------|

>50K | 47 | 3653 | 3700 |

| 0.013 | 0.987 | 0.246 |

-------------------|-----------|-----------|-----------|

Column Total | 11337 | 3723 | 15060 |

-------------------|-----------|-----------|-----------|