Data Visualization – Lab 8
Naïve Bayes Classification

Name: Ayush Sharma
Reg. No: 15BCE1335
Faculty: Dr. Priyadarshini J

Ques. Prediction of Edible mushroom and visualization

Code for Naïve bayes classification without the library:

```
args<-commandArgs(TRUE)
nbc_mushroom <- function(training.dataset, test.dataset, output.filename){
  cat("Running...")
  ##read data from file to a data frame
  training.table <- read.table(training.dataset)
  test.table <- read.table(test.dataset)
  ## Already broken in to 30% and 70%
  ##retrieve class and features from training data
  training.class <- training.table[, 1]
  training.features <- training.table[,-1]
  remove(training.table)


  ##Learn the features by calculating likelihood
  likelihood.list <- list()
  #calculate CPD by feature
  for (i in 1:dim(training.features)[2]){
    feature.values <- training.features[, i]
    unique.feature.values <- unique(feature.values)
    likelihood.matrix <- matrix(rep(NA), nrow=dim(priors)[1], ncol=length(unique.feature.values))
    colnames(likelihood.matrix) <- unique.feature.values
    rownames(likelihood.matrix) <- priors[, "classification"]
    for (item in unique.feature.values){
      likelihood.item <- vector()
      for (class in priors[, "classification"]){
        feature.value.inclass <- feature.values[training.class==class]
        likelihood.value <- length(feature.value.inclass[feature.value.inclass==item])/length(feature.value.inclass)
        likelihood.item <- c(likelihood.item, likelihood.value)
      }
      likelihood.matrix[, item] <- likelihood.item
    }
    likelihood.list[[i]] <- likelihood.matrix
  }


  ##Predict class for the test dataset
  #retrieve the features and target class of the testing dataset
  test.features <- test.table[, -1]
  test.target.class <- test.table[, 1]
```

```
  test.predict.class <- rep(NA, length(test.target.class))
  remove(test.table)

for (item in 1:length(record)){
  likelihood.value <- likelihood.list[[item]][class, as.character(record[1, item])]
  likelihood.v <- c(likelihood.v, likelihood.value)
  }
  accuracy <- length(test.predict.class[test.predict.class==test.target.class])/length(test.target.class)
  test.output <- cbind(test.features, test.target.class, test.predict.class)

  #print result and export to file
  file.con <- file(output.filename)

  write("Naive Bayes Classification Results\n", file=output.filename)
  #write("\n", file = output.filename, append = TRUE)
  write("Next, print the likelihood value for all the 21 features\n", file=output.filename, append=TRUE)
  for (i in 1:length(likelihood.list)){
    write(paste("\nLikelihood for feature", i, "\n"), file=output.filename, append=TRUE)
    write.table(data.frame(likelihood.list[[i]]), file=output.filename, eol="\n", append=TRUE)
  }
  write(paste("\n\nFinally, with the above Naive Bayes classifier, the prediction accuracy is", accuracy,"\n"),
file=output.filename, append=TRUE)
  cat("\nCompleted! Output in file:", output.filename, "\n")
}
nbc_mushroom(args[1], args[2], args[3])
```

Visualisation:

```
print(df['class'].value_counts())
df['class'].value_counts().plot(kind='bar')
sns.factorplot("class", col="gill-color", data=df_forplot, kind="count", size=2.5, aspect=.8, col_wrap=6)
sns.factorplot("class", col="cap-shape", data=df_forplot, kind="count", size=2.5, aspect=.8, col_wrap=6)
plot_grid(ggplot(mushroom, aes(x=cap.shape,fill=class))+ geom_bar(),
     ggplot(mushroom, aes(x=cap.surface,fill=class))+ geom_bar()+bar_theme1,
     ggplot(mushroom, aes(x=cap.color,fill=class))+ geom_bar()+bar_theme1,
     ggplot(mushroom, aes(x=bruises,fill=class))+ geom_bar()+bar_theme1,
     ggplot(mushroom, aes(x=odor,fill=class))+ geom_bar()+bar_theme1,
     ggplot(mushroom, aes(x=habitat,fill=class))+ geom_bar()+bar_theme1,
     align = "h")
```

Result:

Likelihood for feature 1

"f" "x" "k" "b" "s" "c"
"p" 0.400055991041433 0.435890257558791 0.151735722284434 0.0111982082866741 0 0.00111982082866741
"e" 0.377564269021034 0.463256297065697 0.0542716177616204 0.0971176317839522 0.0077901843676967 0

Likelihood for feature 2

"y" "s" "f" "g"
"p" 0.443729003359462 0.361422172452408 0.193729003359462 0.00111982082866741
"e" 0.358867826538561 0.26850168787328 0.372630485588159 0

Likelihood for feature 3

"n" "b" "e" "w" "y" "g" "c" "p" "r" "u"
"p" 0.259518477043673 0.0293952967525196 0.226203807390817 0.0811870100783875 0.171332586786114 0.208846584546473 0.00279955207166853 0.0207166853303471 0 0
"e" 0.29862373409504 0.00986756686574916 0.149831212672033 0.170864710464814 0.093741885224617 0.246689171643729 0.00804985717995326 0.0142820046741106 0.00389509218384835 0.00415476499610491

Likelihood for feature 4

"f" "t"
"p" 0.840705487122061 0.15929451287794
"e" 0.344326149052194 0.655673850947806

Likelihood for feature 5

"y" "f" "a" "n" "s" "p" "c" "m" "l"
"p" 0.147256438969765 0.552631578947368 0 0.0296752519596865 0.148376259798432 0.0663493840985442 0.0461926091825308 0.00951847704367301 0
"e" 0 0 0.0965982861594391 0.809140482991431 0 0 0 0 0.0942612308491301

Likelihood for feature 6

"f" "a"
"p" 0.995240761478164 0.00475923852183651
"e" 0.957413658789925 0.0425863412100753

Likelihood for feature 7

"c" "w"
"p" 0.973124300111982 0.0268756998880179
"e" 0.714359906517788 0.285640093482212

Likelihood for feature 8

"n" "b"
"p" 0.567469204927212 0.432530795072788
"e" 0.0688132952479875 0.931186704752012

Likelihood for feature 9

"b" "h" "w" "y" "g" "k" "n" "u" "p" "e" "o" "r"
"p" 0.442889137737962 0.134098544232923 0.0621500559910414 0.00531914893617021 0.127379619260918 0.0167973124300112 0.0288353863381859 0.0123180291153415 0.16461366181411 0 0 0.00559910414333707
"e" 0 0.048039470267463 0.225915346663204 0.0140223318618541 0.0599844196312646 0.0820566086730719 0.221760581667099 0.107244871461958 0.20436250324591 0.0231108802908335 0.013502986237341 0

Likelihood for feature 10

"t" "e"
"p" 0.515957446808511 0.484042553191489
"e" 0.616463256297066 0.383536743702934

Likelihood for feature 11

"s" "k" "f" "y"
"p" 0.391097424412094 0.570548712206047 0.0363941769316909 0.00195968645016797
"e" 0.863152427940795 0.0353155024668917 0.0978966502207219 0.00363541937159179

Likelihood for feature 12

"s" "y" "k" "f"
"p" 0.391097424412094 0.0193169092945129 0.553471444568869 0.0361142217245241
"e" 0.806284082056609 0.050376525577772 0.0355751752791483 0.107764217086471

Likelihood for feature 13

"w" "g" "o" "b" "p" "n" "c" "e" "y"
"p" 0.433370660694289 0 0 0.111702127659574 0.333986562150056 0.10946248600224 0.00951847704367301 0 0.00195968645016797
"e" 0.657231887821345 0.136068553622436 0.0425863412100753 0 0.138405608932745 0.00363541937159179 0 0.0220721890418073 0

Likelihood for feature 14

"w" "g" "o" "p" "n" "e" "c" "b" "y"
"p" 0.427211646136618 0 0 0.332306830907055 0.112821948488242 0 0.00951847704367301 0.112262038073908 0.00587905935050392
"e" 0.645286938457543 0.138145936120488 0.0425863412100753 0.136847572059205 0.0153206959231368 0.0218125162295508 0 0 0

Likelihood for feature 15

"p"
"p" 1
"e" 1

Likelihood for feature 16

"w" "o" "n" "y"
"p" 0.998040313549832 0 0 0.00195968645016797
"e" 0.957413658789925 0.0212931706050377 0.0212931706050377 0

Likelihood for feature 17

"o" "t" "n"
"p" 0.972844344904815 0.0176371780515118 0.00951847704367301
"e" 0.87405868605557 0.12594131394443 0

Likelihood for feature 18

"e" "p" "l" "n" "f"
"p" 0.452687569988802 0.20548712206047 0.332306830907055 0.00951847704367301 0
"e" 0.238119968839263 0.749935081796936 0 0 0.0119449493638016

Likelihood for feature 19

"w" "h" "k" "n" "o" "y" "r" "b" "u"
"p" 0.46444568868981 0.405375139977604 0.0559910414333707 0.0565509518477044 0 0 0.0176371780515118 0 0
"e" 0.137626590495975 0.0119449493638016 0.395481693066736 0.411321734614386 0.0101272396780057 0.0106465853025188 0 0.0111659309270319 0.0116852765515451

Likelihood for feature 20

"v" "s" "y" "c" "a" "n"
"p" 0.727883538633819 0.0923852183650616 0.166013437849944 0.0137178051511758 0 0
"e" 0.281745001298364 0.209036613866528 0.255258374448195 0.0664762399376785 0.091404829914308 0.096078940534926

Likelihood for feature 21

"p" "u" "d" "l" "g" "w" "m"
"p" 0.256438969764838 0.0691489361702128 0.320828667413214 0.154535274356103 0.190089585666293 0 0.00895856662933931
"e" 0.0324591015320696 0.022851207478577 0.448454946767073 0.0542716177616204 0.336795637496754 0.0446637237081278 0.0605037652557777

Finally, with the above Naive Bayes classifier, the prediction accuracy is **0.998573466476462**

Screenshot: