Workshop 4
Naive Bayes Algorithm in R

Exercise 1 Naive Bayes Algorithm: Revision

Read through the first part of the website "Naïve Bayes classification in R" by Zhongheng Zhang. This introduces the Naive Bayes Classifier including a medical example followed by fitting a Naive Bayes Model to the Titanic data. Do not run the code in this website; this is covered in the next exercise.

Exercise 2 A simple Example

We will repeat the Titanic example in Zhang using slightly different code given below. First load the library `e1071` which contains the function `naiveBayes()` installing it, if it is not already installed.

```
#install.packages("e1071")
library(e1071)
data(Titanic)
Titanic
```

Note that the data is in array form (frequency data). It is possible to pass frequency data into `naiveBayes()` but it is much easier to analyse the results, when it is in the form of a data frame.

```
#Save into a data frame and inspect it
Titanic_df<-as.data.frame(Titanic)
#expand the number of rows to correspond to the frequencies
repeating_sequence<-rep.int(seq_len(nrow(Titanic_df)), Titanic_df$Freq)
#This will repeat each combination equal to the frequency of each combination
#Create the dataset by row repetition created
Titanic_df<-Titanic_df[repeating_sequence,]
#We no longer need the frequency, drop the feature
Titanic_df$Freq<-NULL
```

Fit the Naive Bayes model using standard R model fitting notation, and print the model output.

```
NB_Titanic<-naiveBayes(Survived ~., data=Titanic_df)
NB_Titanic
```

Note that the conditional probabilities are just the observed marginal frequencies from the data.

```
prop.table(table(Titanic_df$Survived))
prop.table(table(Titanic_df$Survived,Titanic_df$Class),1)
prop.table(table(Titanic_df$Survived,Titanic_df$Sex),1)
prop.table(table(Titanic_df$Survived,Titanic_df$Age),1)
```

Obtain the model predictions and print the classification statistics.

```
NB_Preds<-predict(NB_Titanic,Titanic_df)
#Confusion matrix to check accuracy
tab<-table(NB_Preds,Titanic_df$Survived)
tab
#sensitivity
tab[2,2]/sum(tab[,2])
#specificity
tab[1,1]/sum(tab[,1])
#misclassification rate
1-sum(diag(tab))/sum(tab)
```

For these data the sensitivity is poor but the specificity is good.

Exercise 3 The IBM attrition data

The `attrition` data set in the `rsample` library, contains data on "employee attrition", which means employees resigning. Load the rsample library and read the short help page for `attrition`.

Data preprocessing There are two numeric variables which are coded as numbers but are really factor variables. Convert these variables, then define a training and test data set using the function `initial_split` from `rsample`.

```
attrition$JobLevel<-as.factor(attrition$JobLevel)
attrition$StockoptionLevel<-as.factor(attrition$StockOptionLevel)
set.seed(1)
split <- initial_split(attrition, prop = .7, strata = "Attrition")
train <- training(split)
test  <- testing(split)
prop.table(table(train$Attrition))
prop.table(table(test$Attrition))
```

Fit a naive Bayes Model to the test data using the following subset of variables:
`Attrition~Age+DailyRate+DistanceFromHome+HourlyRate+MonthlyIncome+MonthlyRate`

Analyse the output as with the Titanic data, using the test data. Notice that the specificity is especially poor. Repeat the model with all of the available variables. The sensitivity increases to 68%.[1]

Exercise 4 Comparison with LDA
The MASS library contains the function `lda()` to fit a linear discriminant analysis. Use this to fit a similar model to the attrition data, and compare the results with the Naive Bayes results.

Tips The `lda()` function gives a warning that there is collinearity in the predictor variables. This means that that two or more variables are linearly dependent. This has no effect on the predicted values.

The default output from `predict()` includes a matrix of probabilities and the classes use `predict()$class` for the classification matrix.

Exercises to do at home

Exercise 5 Conditional independence
Use the elementary definitions in probability theory of stochastic independence and conditional probability, to solve the following exercises.

(a) Show that

$$P(X_1|X_2) = P(X_1)$$

is equivalent to

$$P(X_1 \cap X_2) = P(X_1)P(X_2)$$

and

$$P(X_2|X_1) = P(X_2)$$

(b) Show that

$$P(X_1|Y, X_2) = P(X_1|Y)$$

is equivalent to

$$P(X_1 \cap X_2|Y) = P(X_1|Y)P(X_2|Y)$$

and

$$P(X_2|Y, X_1) = P(X_2|Y)$$

[1] A more comprehensive NB analysis of these data, using a different style of R-code can be found at the website: `http://uc-r.github.io/naive_bayes`.

Exercise 6 Conditional independence

A box contains two coins: a regular coin M_1 with Heads H and Tails T, and one trick two-headed coin M_2, i.e. $P(H|M_2) = 1$. A coin is chosen at random 50-50 from the box and it is tossed twice.

Define the following events.

- $A =$ First coin toss results in an H
- $B =$ Second coin toss results in an H

Calculate

(a) $P(A|M_1)$,

(b) $P(B|M_1)$,

(c) $P(A \cap B|M_1)$. You may assume that *if* we know that we are using a fair coin that the result of the second toss is independent from the first.

(d) $P(A|M_2)$,

(e) $P(B|M_2)$,

(f) $P(A \cap B|M_2)$.

(g) Use the law of total probability: $P(A) = P(A|M_1)P(M_1) + P(A|M_2)P(M_2)$ to calculate $P(A)$.

(h) Calculate $P(B)$, and

(i) Use the law of total probability: $P(A \cap B) = P(A \cap B|M_1)P(M_1) + P(A \cap B|M_2)P(M_2)$ to calculate $P(A \cap B)$.

Confirm that A and B are NOT independent, but they are conditionally independent given which coin M_1 or M_2 is used.