Birla Institute of Technology and Science

Foundations of Data Science (CS F320)

1st Semester 2021-22

Lab Sheet #5: Naive Bayes Classifier

Naive Bayes Classifier

Naive Bayes classifier is a generative classifier with assumption that features are statistically independent.For *n* features $x_1,x_2,...,x_n$ and a class C_k The model can be expressed as:

$$P(C_K|X_1,X_2,....,X_n)\alpha P(C_K)\ \Pi^n_{i=1}\ P(X_i\ |\ C_i)$$

The naive Bayes functionality is supported in the e1071 package. Install and load the package by issuing install.packages(“e1071”) and library(e1071). This will make the naiveBayes() function available for us to use.

We would use the Titanic dataset available with the e1071 package to create our first NaiveBayes model. We need to expand the summarize dataset into individual rows before we can use it for model creation.

Laplace Smoother

An additional issue to be aware of - since naive Bayes uses the product of feature probabilities conditioned on each class, we run into a serious problem when new data includes a feature value that never occurs for one or more levels of a response class. What results is P $(x_i\ |\ C_k) = 0$ for this individual feature and this zero will ripple through the entire multiplication of all features and will always force the posterior probability to be zero for that class.

A solution to this problem involves using the *Laplace smoother*. The Laplace smoother adds a small number to each of the counts in the frequencies for each feature, which ensures that each feature has a nonzero probability of occurring for each class.

$$L_i = \frac{C_i + \mathrm{K}}{\mathrm{N} + \mathrm{K.c}}$$

C_i :- Count of ith tupples
K : Laplacian parameter (add small no.)
N: Total no. of tupples
c: no. of classes

Predicting survival

```
#Install the package
install.packages(“e1071”)

#Loading the library
library(e1071)

#Next load the Titanic dataset
data("Titanic")

#Save into a data frame and view it
Titanic_df=as.data.frame(Titanic)

#To retrieve the dimension of data frame
dim(Titanic_df)

#Creating data from table
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_dataset=Titanic_df[repeating_sequence,]

#We no longer need the frequency, drop the feature
Titanic_dataset$Freq=NULL

#To view the structure of dataset
str(Titanic_dataset$Survived)

#check the proportion of our data in class target
prop.table(table(Titanic_dataset$Survived))

#Fitting the Naive Bayes model
Naive_Bayes_Model=naiveBayes(Survived ~., data=Titanic_dataset)

#What does the model say? Print the model summary
Naive_Bayes_Model

#Prediction on the dataset
NB_Predictions=predict(Naive_Bayes_Model,Titanic_dataset)

#Design Confusion matrix
table(NB_Predictions,Titanic_dataset$Survived)
```

```
#Model evaluation in naive-bayes by using confusion matrix to check
accuracy

install.packages("listenv")
install.packages("caret")
library(caret)
library(listenv)

mat1 <- confusionMatrix(data = NB_Predictions, reference =
Titanic_dataset$Survived, positive = "No")
```

Exercise

1. Find the true positive rate (TPR) and false positive rate(FPR) for the Titanic survival classifier we modeled above.
2. Find out about Laplace smoothing and its need in Naïve Bayes classification. How can you control it using the naiveBayes() function?
3. Download the Census Income DataSet (from this link https://archive.ics.uci.edu/ml/datasets/census+income). Builda NaiveBayes Classifier to predict income level. You may only use categorical variables, and ignore the continuous variables. Check weather using laplace our model performed better or not.
4. Find out about numeric and factor variables.
5. Find out how data is stored in dataframes.