Birla Institute of Technology and Science, Pilani

BITS-F464: Machine Learning
2^{nd} Semester 2018-19

Labsheet-04: Naive Bayes

1 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:

$$P(C_k|x_1, x_2, ..., x_n) \propto P(C_k)\Pi_{i=1}^{n} P(x_i|C_k) \tag{1}$$

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` data set available with the `e1071` package to create our first Naive Bayes model. We need to expand the summarized data set into individual rows before we can use it for model creation.

1.1 Predicting survival

```
  # Load the Titanic dataset
> dataset("Titanic")
> print(Titanic)

  # Create a data frame
> Titanic_dataframe = as.data.frame(Titanic)
> print(Titanic_dataframe)

  # Create the data set by repeating each row required number of times
> sequence = rep.int(seq_len(nrow(Titanic_df)), Titanic_df$Freq)
> Titanic_dataset=Titanic_df[sequence,]

  # The frequency column is not needed now
> Titanic_dataset$Freq=NULL

  # Build the model
> model=naiveBayes(Survived ~., data=Titanic_dataset)
> print(model)

  # Create a confusion matrix
> predictions = predict(model,Titanic_dataset)
> table(predictions, Titanic_dataset$Survived)
```

1.1.1 Explore

Issue the following commands to learn more: `?as.data.frame, ?rep.int, ?naiveBayes`

2 Question-01

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 Naive Bayes classification. How can you control it using the naiveBayes() function?
3. Find out about numeric and factor variables.
4. Find out how data is stored in data frames.

3 Question-02

Download the Census Income Data Set (from this link [1]). Build a Naive Bayes Classifier to predict income level. You may only use categorical variables, and ignore the continuous variables.

[1] https://archive.ics.uci.edu/ml/datasets/Census+Income