Shahjalal University of Science & Technology
School of Applied Science & Technology
Department of Computer Science & Engineering
Sylhet-3114

CSE 476

Machine Learning Lab Report

Submitted by
Nazmul Islam - 2014331034
Md. Shamsul Arafin Mahtab - 2014331063

Submitted to
Assistant Professor, Ayesha Tasnim

15 July, 2018

1 Fake News Detection

1.1 Datasets

The datasets of Fake new Detection is collected from the site kaggle.com regarding 4 columns and around 6000 observations as Crawler Id, Title, Text and Label. Label is the target attribute containing two classes FAKE or REAL. For training the datasets, the Text column is used.

1.2 Preprocessing and Feature

As we have to concern about time, we could not do any preprocessing. But the datasets is well preprocessed. So the accuracy did not come to low. The data sets is categorized and 33 percentage reserved as test. We have used sklearn library for the all the machine learning model and features.
After the preprocessed data loaded from pandas and trained data is vectorized using count vectorizer.

1.3 Applied Methods

- **Multinomial Naive Bayes**

  It is a supervised classification method assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
  Accuracy: 89.335 %

- **Support Vector Machine**

  It is also a supervised classification method. We plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
  Accuracy: 74.079 %

- **Decision Tree**

  In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.
  Accuracy: 81.731 %

- **Neural Network**

  This is a single layer neural network method where is each 3 attributes are 13 for simplicity. And the maximum iterations is 500.
  Accuracy: 91.153 %

2 Apartment Price Predictions

2.1 Datasets

The KC House Data consists of 19 house features and the price and the id columns with 21613 observations.

2.2 Preprocessing

The csv format data is loaded using pandas. One of the feature time is converted to 1 and 0 as the values were ambiguous. Then the train dataset was formed by dropping 'id' and 'price'. The feature price is the label and treated as output of the file. The file was then split into two sets to get the training and testing set.

2.3 Applied Methods

- **Linear Regression**

  It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.
  Accuracy: 73.203 %