Report of the bigram.py file

Satenik Rafayelyan

December 6, 2019

1 Introduction

A bigram is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words. In this exercise, it is given that bigrams are created by using 26 words English alphabet.

There are 2 different data sets including possible bigrams. The first data includes 5493 distinct words and these words are classified as 0 and 1. Therefore, total number of 0's and 1's equal to the number of words. The second data includes 5493 different words and these are classified as 0,1 and 2. The total number of classifiers 0,1 and 2 equal to 6001.

To analyze the data, some preprocessing is required. Firstly, 26 words from English alphabet is created. Then, bigram is formed by matching two by two letters. By using this bigram, each word in the first and the second data is splitted into 2 letters and transformed into vectors based on whether bigram includes these two letters or not. The aim of the project is to make a classification and measure it with corresponding metrics.

2 Methods/Models

When considering the data set, it is obvious that analysing the data sets is supervised machine learning task since the data sets include inputs and the labels of those inputs. Since it is supervised machine learning problem and the response is categorical, we can consider classification. There are many classification algorithms in machine learning. In this project, logistic regression and k-nn classification algorithms are used as classifiers.

3 Experiments

First of all we defined a function which takes our data as an input and returns every two combination of letters. So in this case we have a big list with small lists inside, overall the length of the list is 16479. For example the first list is [('A', 'E'), ('E', 'P'), ('P', 'P'), ('P', 'K'), ('K', 'B'), ('B', 'V')],which corresponds to the first word 'AEPPKBV'. Then we should define a function which will make the elements of the list lower case and merge the elements in each tuple. This function will return a list with ['ae', 'ep', 'pp', 'pk', 'kb', 'bv'] as a first element. Then we create a list of English alphabet and take all possible combinations of that (keywords): ['aa', 'ab', 'ac', 'ad',...,'zz']. Then we continue our preprocessing by defining a new function, which takes as inputs our data and keywords. Here we iterate through our data and create a logic according to which if the the elements of the data and the element of the keyword are equal we have 1, otherwise 0. Therefore we gain a binary matrix with the shape (5493, 676) in the first case and (6001, 676) in the second case. Afterwards, for the first dataset we perform Logistic Regression and KNN. For Logistic Regression we have the best results with the following parameters penalty = 'l1', tol=0.01, C=0.9, where the penalty used to specify the norm used in the penalization, tol is the tolerance for stopping criteria and C is the inverse of regularization strength. For KNN taking the parameter n_equal to 10 which shows the number of neighbors is 10 gives the highest accuracy and precision. For the second data set the same parameters were used.

3.1 Data

Two different data sets are used in the analysing steps. Both two data are binary in terms of their features.

Mean is one of the characteristics which represent the data. However, because we are working with the binary data set, calculating the mean value is not possible. Therefore, instead of using mean value, mode can be used. For both data sets, the most repeated value is "by" with 776. The first data has 5493 observations and 676 features while the second one has 6001 observations and 676 features. The main differences between the first and the second data set is that their response variable. In the first one the response variable has only two categories which are 0,1 and in the second one response variable has three categories which are 0,1,2.

It is important to determine whether the data is balanced or not in terms of the number of observations in each category. To understand this condition for both data sets, we checked the number of observations in the categories.

In the first data set, number of observations in case of Y=1 is 1219 and for the case Y=0 it is 4274. There is a huge gap between these two categories in terms of number of observations. As a solution, some rows can be removed until the gap is eliminated or the category which has less observations can be generated. In that project, some rows are eliminated and at the end of this step both the number of Y=0 and the number of Y=1 equal to 1219.

In the second data set, the number of observations for the condition of Y=0 is 2002, Y=1 is 2000 and Y=2 is 1999. It is clear that the data can be considered as a balanced data. Therefore, no pre-processing is used for the second data.

3.2 Feature Extraction

Feature selection can be described as a task to detect the most relevant feature from the data. There are three different feature extraction methods such as filter methods, wrapper methods and embedded methods. In this assignment, L1 regularization method is used for feature selection. After applying L1 into the data, the number of selected features is 333 while the eliminated ones is 343. While the logistic regression and the K-NN regression algorithm were applying, feature extraction part is added by writing penalty="l1".

3.3 Awesome results

Figure 1: Results

Table 1: Accuracy Table

Data	Logistic Regression	K-NN(k=7)	K-NN(k=10)
data 1	0.74	0.67	0.74
data 2	0.65	0.5	0.49

Table 2: Precision Matrix

Data	Logistic Regression	K-NN(k=7)	K-NN(k=10)
data 1	0.71	0.68	0.78
data 2	0.65	0.49	0.51

4 Discussion

In this assignment, two different data are used for classification. As a classification algorithms, logistic regression and K-NN method are applied. Furthermore, to select the best model, L1 method is used while eliminating the features. Considering the accuracy table, it is clear that selecting logistic regression for both data is logical since the accuracy values in logistic regression are the higher.