STUDY ON COMPARISON OF MACHINE LEARNING MODEL PERFORMANCES ON SENTIMENT TEXT ANALYSIS AND CLOSENESS OF PREDICTED DISTRIBUTIONS

CHATA.AI

NAME : VIVEK SURESH RAJ

DATE : 02-SEP-2021

Keywords

Naive bayes

Gaussian naïve bayes

Conditional probability

K-Nearest Neighbors

Clustering

Abstract

A Naïve bayes classifier algorithm based on conditional probability is applied over the multi-label classification dataset for text sentimental analysis. A strong comparison over the dataset is shown with K-nearest neighbor algorithm. The predicted labels were observed and strong similarity check is performed to determine the closeness of the predicted vectors. The probability distribution of the models were subjected to pairwise similarity score measure for cross-verification of the predicted distribution.

1. INTRODUCTION:

The given datasets were split into train and test data. The feature attributes of the dataset were '*review'*. The target label contained the *'ratings'* of the sentences ranging from 1 to 5. The target column was grouped according to the labels to performed an exploratory data analysis on the dataset.

Fig.1. Pie chart represenation with highest number exploded. Fig.2. Scattered representation of ratings.

The preprocessing steps for the text includes cleaning data inluding Stemming, stopwords removal and lowercasing of the texts with train and test data. The corpus was later converted to certain standard vector/token representation.

INTERPRETATIONS:

The reviews with the ratings are as follows according to the analysis in the *'train_data'*.

- Rating -1 : 7028 reviews
- Rating -2 : 7031 reviews
- Rating -3: 6971 reviews
- Rating -4 : 6997 reviews
- Rating -5: 6977 reviews.

2. METHOD:

2.1 Why Gaussian Naïve Bayes algorithm?

The naïve bayes algorithm is employed because of the independency towards the target labels. Since in the given we have target as multi-class labels which should be treated independent of each other, a sophisticated naïve bayes classifier based on conditional probability is used.

P(rating | reviews) = P(reviews | ratings) * P(rating) / P(review)

Gaussian: Since because of the normal distribution, it is easy to predict the future rating within the predicted bell-curve distribution for the test data.

```
Test review: Not at all what expected.   Our mountain view was of the garage.  Asked to move but no rooms available and they said this was
considered mountain view since it wasn't ocean view.   This was a business trip so if I was paying for this I would have been livid.  Woul
d suggest if you book this place guarantee your view for the prices they charge.  No storage in bathroom at all.  Razor barely fit.  The t
hree drawers provided were all this tiny and no space on the sink.    There coffee table is the size of a postage stamp so don't plan on p
utting anything there.  The rooms that they have for hotel guests are small.  Beware also the rooms have doorbells so kids and some guests
like to ring these at all hours day and night.  We did complain so they had security do extra rounds.  Why a…

prob_of_predicted_rating: [1. 0. 0. 0. 0.]
```

Fig.3. Sample of the predicted rating from distribution (GaussianNB prediction)

2.1. Why KNN – Knearest neighbor?

Another interesting and advanced algorithm is the K-nearest neighbor. It is being used after keen analysis over the train_data. Since, the target has multiple classes, it is better to choose a model that could separate the classes with Euclidean distance in a continuous space. Upon, observing from the above-mentioned perspective, KNN was a better advanced choice to solve sentimental text classification.

The same sample as observed with naïve bayes algorithm is subjected to KNN algorithm. The output probability distribution is found to show similar rating as corresponding to the previous method.

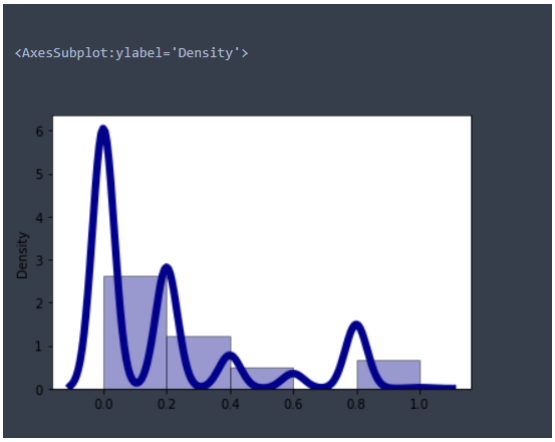
```
Test review: Not at all what expected.   Our mountain view was of the garage.  Asked to move but no rooms available and they said this was
considered mountain view since it wasn't ocean view.   This was a business trip so if I was paying for this I would have been livid.  Woul
d suggest if you book this place guarantee your view for the prices they charge.  No storage in bathroom at all.  Razor barely fit.  The t
hree drawers provided were all this tiny and no space on the sink.    There coffee table is the size of a postage stamp so don't plan on p
utting anything there.  The rooms that they have for hotel guests are small.  Beware also the rooms have doorbells so kids and some guests
like to ring these at all hours day and night.  We did complain so they had security do extra rounds.  Why a…

prob_of_predicted_rating KNN: [0.8 0.2 0.  0.  0. ]
```

Fig.4. Sample of the predicted rating from distribution (KNN prediction)

3. EXAMING THE CLOSNESS OF THE PREDICTED DISTRIBUTIONS:

3.1.PROBABILITY DISTRIBUTION:

The predicted rating of the two models were subjected to normal distributions. Upon observing the distributions of the classes of two models, they had the dense distribution around the same ratings (**test_data ratings corresponds to '1' in both model distributions**).



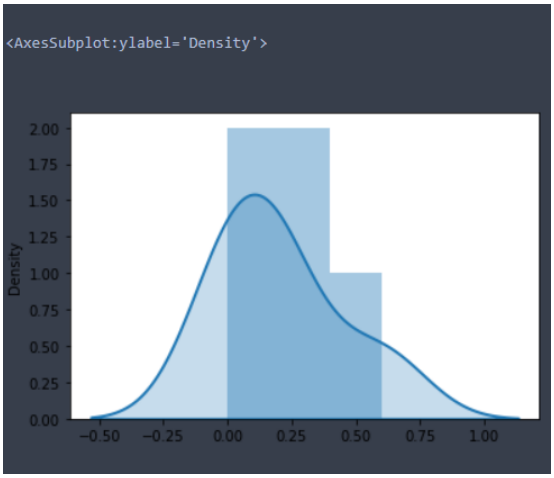


Fig.5. KNN distribution with rating=1 as max. Fig.6. NB predicted distribution with rating=1 as max

3.2.CLOSENESS OF DISTRIBUTION INDEX:

A complete verification of the closeness of the distribution by two models were studied. Cosine similarity metric is used to further analyse the closeness of the vector fitting the data elements in vector spaces.

Cos = a . b / ||a|| . ||b||

Upon final observation made with between random sample of the distributions, the scores were found to be **0.97**. Interestingly both the distributions has peak at rating=1.

```
Closeness score between 2 distributions showing max score at rating 1: [0.9701425  0.24253563 0.         0.         0.        ]

Individual distributions confirming the same rating :
 NB 0.6
 KNN [0.8 0.2 0.  0.  0. ]
```

Fig.7. Closeness of the distributions.

4. DISCUSSIONS:

The study could further be extended by approach with ensemble learning methods for collective analysis over dataset. It also be approached by using deep learning method with embedding representations and dense layers.

5. CONCLUSION:

This study has addressed and analysed the sentiment classification of textual data with two of the models. Further the closeness of the predicted distributions was studied.