ML ASSIGNMENT 6
IIT2018178, Manav

INTRODUCTION

We have to design all three Naïve Bayes classifier for filtering Spam and Ham (Normal) messages. and make a comparative study on the performance of all these three models.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes is of three types:
Multinomial Naive Bayes
Bernoulli Naive Bayes
Gaussian Naive Bayes

The basic naive bayes algorithm is based on Bayes Theorem and shown below.

$$P(c_i \mid x) = \frac{P(x \mid c_i) * P(c_i)}{\sum_j P(x \mid c_j) * P(c_j)}$$

The calculation of prior probabilities for each of the three models is the same, however the conditional probability of each model is calculated in different ways

$$Prior\ Probability(c) = \frac{No.of\ instances\ of\ class\ c}{Total\ No.of\ instances\ in\ the\ dataset}$$

MULTINOMIAL NAIVE BAYES

The conditional probability for each feature for each class is given as the numerical sum of that feature divided by the number of samples of that particular class + number of features of the class.

The accuracy of the Model was : 98.55334538878843 %

BERNOULLI NAIVE BAYES

The conditional probability for each feature for each class is given as the numerical sum of that feature divided by the number of samples of the particular class. Note that, all the features are converted to 1's and 0's from integers for this model.

The accuracy of the Model was : 95.02712477396021 %

GAUSSIAN NAIVE BAYES

In this model we obtain our training set, and calculate the mean and standard deviation for each feature for all classes. Then we use the test set to get a sample and use the mean and standard deviation of each class for each feature of the sample as a conditional probability. We use the gaussian probability function to calculate each conditional probability for each sample.

$$P(x_i \mid c) = \frac{1}{\sqrt{2 * \pi * sigma_{x_i,c}^2}} * \exp\left(-\frac{\left(x_i - mean_{x_i,c}\right)^2}{2 * sigma_{x_i,c}^2}\right)$$

'x' is one sample and x_i is one feature of the sample. $mean_{x_i,c}$ is the mean of that feature for that class. Similarly for $sigma_{x_i,c}$.

The accuracy of the model was : 87.9746835443038 %

COMPARISON

When we compare the above results, we see that Multinomial and Bernoulli both produce almost similar results with some variations, however they outperform the Gaussian Naive Bayes model.

Bernoulli models the presence/absence of a feature. Multinomial models the number of counts of a feature, and gaussian looks at the mean and standard variation of the training set and determines the cut off on that basis.

We note that, according to our obtained results Gaussian Naive Bayes is not an accurate model to be used here. As a Multinomial or Bernoulli's Naive Bayes outperforms it.