ALAN THOMAS

21BDA08

(a). Write the difference between the following:

i). Gaussian Naive Bayes

Gaussian Naive Bayes is built on the assumption of a normal distribution of probabilities. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell-shaped curve which is symmetric about the mean of the feature values.

ii). Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

iii). Complement Naive Bayes

Complement Naive Bayes approach is almost the same as the Multinomial Naive Bayes, though now we count the occurrences of a word in the complement to the class. Multinomial Naive Bayes is not able to do very well with unstable data.

iv). Bernoulli Naive Bayes

Bernoulli Naive Bayes only takes binary values. The most general example is where we check each value will be a word or not that appears in the document.

Bernoulli distribution has two mutually exclusive outcomes: P(X=1)=p or P(X=0)=1-p. In Bernoulli theorem we can have multiple features but each one is assumed to be binary valued variable i.e., Boolean. Therefore, this class requires samples to be represented as binary-valued feature vectors. In case, any other kind of data is provided, then Bernoulli instance may binarize its input.

v). Categorical Naive Bayes

It is suitable for classification with discrete features which assumes categorical distribution for each feature. The features should be encoded using label encoding techniques such that each category would be mapped to a unique number.

vi). Out-of-core naive Bayes model fitting

Naive Bayes models can be used to tackle large scale classification problems for which the full training set might not fit in memory. To handle this, MultinomialNB, BernoulliNB and GaussianNB expose a partial\_fit method that can be used incrementally as done with other classifiers. All naive bayes classifiers support sample weighting.

The partial\_fit method call of naive Bayes models introduces some computational overhead. It is recommended to use data chunk sizes that are as large as possible, that is as the available RAM allows.

(b). Define which text pre-processing and text transformation steps did you use for the above.

I used no text or transformation steps for above.

(c). Create a step-by-step tutorial to build a text classifier using Naive Bayes algorithm in a Jupyter Notebook. Define each step and evaluate the model.

**Jaccard Similarity:**

It is the method of measuring similarity between two sets, which means that when you apply two sets as documents, the size of the intersection between two documents is divided by the size of the union. In other words, the similarity measure for two datasets in the range 0 or 1, the closer you are to 1, the more similar the two populations are.

**Cosine Similarity:**

Cosine similarity refers to the similarity that can be obtained by vectorizing a document and using the cosine angle between two vectors. If the two vectors have exactly the same orientation, they will have a value of 1, and if they have an angle of 90 degree, they will have a value of -1 if they have opposite directions of 0 or 180 degrees. If the value is closer to 1, the higher the similarity between the two documents.