ML ASSIGNMENT- LAB03

B) a) Write the difference between the following:

i) Gaussian Naïve Bayes

It supports continuous valued features and models each as conforming to a gaussian(normal) distribution. An approach to create a simple model is to assume that the data is described by a gaussian distribution with no covariance between the dimensions. The gaussian probability density function can be used to make predictions by substituting the parameters with the new input value of the variable and as a result, the gaussian function will give an estimate for the new input values probability.

ii) Multinomial Naïve Bayes

The multinomial naïve bayes algorithm is a Bayesian learning approach popular in Natural Learning Processing(NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the bayes theorem. It calculates each tags likelihood for a given sample and outputs the tag with the greatest chance. It normally requires integer feature counts. It is widely used for assigning documents to classes based on the statistical analysis of their contents. It provides an alternative to the “heavy” AI-based semantic analysis and drastically simplifies textual data classification.

iii) Complement Naïve Bayes

It is somewhat an adaptation of the standard multinomial naïve bayes algorithm. Complement naïve bayes is particularly suited to work with imbalanced datasets. Here, instead of calculating the probability of an item belonging to a certain class, we calculate the probability of the item belonging to all the classes. This is the literal meaning of the word, complement and hence is called complement naïve bayes.

CNB is used

when the dataset is imbalanced : Multinomial and gaussian naïve bayes may give a low accuracy. However, CNB will perform quite well and will give relatively higher accuracy.

For text classification tasks: CNB outperforms both gaussian naïve bayes and multinomial naïve bayes in text classification tasks.

iv) Bernoulli Naïve Bayes

This implements the naïve bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions i.e there may be multiple features but each one is assumed to be binary valued(Bernoulli, Boolean) variable. Therefore, this class requires samples to be represented as binary valued feature vectors; if handed any other kind of data, a bernoulliNB instance may binarize its input( depending on the binarize parameter).

In the case of text classification, word occurrence vectors (rather than word count vectors) may be used to train and use this classifier. BernoulliNB might perform better on some datasets, especially those with shorter documents. It is advisable to evaluate both models, if time permits.

v) Categorical naïve bayes

The categorical naïve bayes classifier is suitable for classification with discrete features that are categorically distributed. The categories of each feature are drawn from a categorical distribution. It implements the categorical naïve bayes algorithm for categorically distributed data. It assumes that each feature, which is described by the index i , has its own categorical distribution. For each feature i in the training set X, categoricalNB estimates a categorical distribution for each feature i of X conditioned on the class y. The index set of the samples is defined as J={1,……,m} with m as the number of samples.

vi) Out-of-core naïve 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 case, multinomial naïve Bayes, Bernoulli naïve Bayes, and Gaussian naïve Bayes expose a partial fit method that can be used incrementally as done with other classifiers as demonstrated in out of the core classification of text documents. All naïve Bayes classifiers support sample weighting.

b) Define which text preprocessing and text transformation steps did you use for above.

Text preprocessing is traditionallhy an important step for NLP tasks. It transforms text into a more digestible form so that machine learning algorithms can perform better. The dataset is preprocessed in order to check missing values, noisy data and other inconsistencies before executing it to the algorithm.

Techniques are expand contractions, lower case, remove punctuations, remove stopwords, rephrase text, remove extra spaces.

c) What is Jaccard and Cosine Similarity?

Jaccard similarity is a common proximity measurement used to compute the similarity between two objects , such as two text documents. It takes only unique set of words for each sentence / document. It can be used to find the similarity between two asymmetric binary vectors or to find the similarity between two sets.

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It takes total length of the vectors.