Assignment Lab03

Ashly Thampi

21BDA19

a.Write the difference between the following:

1. **Gaussian Naive Bayes.**

Gaussian Naive Bayes 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 co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

<https://iq.opengenus.org/gaussian-naive-bayes/#:~:text=Gaussian%20Naive%20Bayes%20supports%20continuous,(independent%20dimensions)%20between%20dimensions>.

1. **Multinomial Naive Bayes.**

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language 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 tag's likelihood for a given sample and outputs the tag with the greatest chance

<https://www.upgrad.com/blog/multinomial-naive-bayes-explained/#:~:text=The%20Multinomial%20Naive%20Bayes%20algorithm%20is%20a%20Bayesian%20learning%20approach,tag%20with%20the%20greatest%20chance>.

1. **Complement Naive Bayes**

Complement Naive Bayes is somewhat a modification of the standard Multinomial Naive Bayes algorithm. Multinomial Naive Bayes is not able to do very well with unstable data. Imbalanced data sets are instances where the number of instances belonging to a particular class is greater than the number of instances belonging to different classes. This implies the spread of the examples is not even. This kind of data can be difficult to analyse as models can easily overfit this data to benefit a class with a larger instance.

<https://www.javatpoint.com/complement-naive-bayes-algorithm>

1. **Bernoulli Naive Bayes.**

Bernoulli Naive Bayes is one of the variants of the Naive Bayes algorithm in machine learning. It is very useful to be used when the dataset is in a binary distribution where the output label is present or absent. The main advantage of this algorithm is that it only accepts features in the form of binary values such as:

1. True or False
2. Spam or Ham
3. Yes or No
4. 0 or 1
5. **Categorical Naive Bayes**

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

1. **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, we expose a partial fit method that can be used incrementally as done with other classifiers as demonstrated in Out-of-core classification of text documents.**

**Contrary to the fit method, the first call to partial fit needs to be passed the list of all the expected class labels**

What is Jaccard **and** Cosine Similarity?

* Jaccard similarity takes only unique set of words for each sentence / document while cosine similarity takes total length of the vectors.

<https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50#:~:text=Jaccard%20similarity%20takes%20only%20unique,term%20frequency%20or%20tf%2Didf>