**ASSIGNMENT 03**

**Name: Gayatri Krishna**

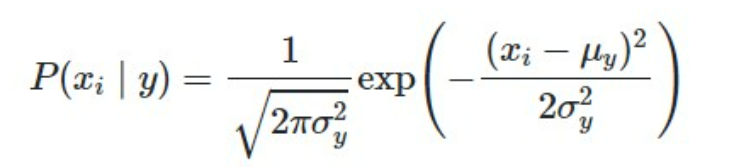
**Reg No: 21BDA16**

* 1. Write the difference between the following:

The simplest and fastest classification approach known, Naive Bayes, is best suited for dealing with massive amounts of data. The Naive Bayes classifier has proven to be effective in a variety of applications, including spam filtering, text classification, sentiment analysis, and recommender systems. The Bayes theory of probability is used to create predictions about unknown classes.

1. **Gaussian Naïve Bayes:**

When working with continuous data, one common assumption is that the continuous values associated with each class follow a normal (or Gaussian) distribution. The features' likelihood is considered to be-



1. **Multinomial Naïve Bayes:**

For discrete data like word counts, multinomial classification works best. It disregards the absence of the features. As a result, if the frequency is 0, the likelihood of that feature occurring is also 0. Multinomial naive Bayes ignores that feature.

1. **Complement Naïve Bayes:**

This method is nearly identical to the Multinomial, with the exception that we now count the occurrences of a word in the complement to the class. With unstable data, Multinomial Naive Bayes does not perform well.

1. **Bernoulli Naïve Bayes:**

When characteristics are binary, the Bernoulli formula is similar to the multinomial one. If you have discrete features in 1s and 0s that signify the existence or absence of a feature, you can use them instead of the frequency of the word. In that instance, the features will be binary, and Bernoulli Naive Bayes will be used.

1. **Categorical Naïve Bayes:**

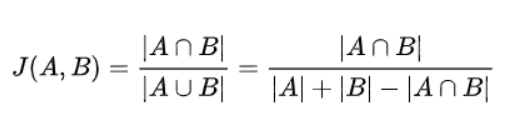
The categorical Naive Bayes classifier is well suited to classification with categorically distributed discrete features. Each feature's categories are chosen from a categorical distribution.

1. **Out-of-core Naïve Bayes model fitting:**

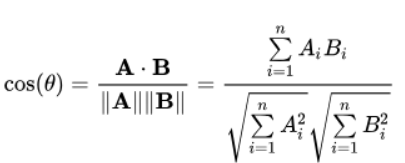
Various nave Bayes methods can be used to solve large-scale classification issues where the entire training data is too vast to fit in memory. As a result, most types supply a partial fit technique for this problem, which may be used progressively like other classifiers, as seen in Out-of-core classification of text documents.

What is Jaccard and Cosine Similarity?

* The Jaccard similarity coefficient is another name for the Jaccard similarity index. It determines how similar two sets are. The percentage range is 0 to 100%. The higher the percentage, the more similar the two words are.



* Cosine similarity calculates the cosine of the angle formed by two vectors in their dot product space. Their similarity is one if the angle is zero; the larger the angle, the less similar they are.



* In comparison to Jaccard similarity and the joint of Cosine and Jaccard similarity, the results of cosine similarity have the highest value.
* When analysing text similarity, Jaccard similarity is good for circumstances where duplication is unimportant, whereas cosine similarity is good for cases where duplication is important.
  1. Define which text preprocessing and text transformation steps did you use for the above.