**DEON S**

**21BDA57**

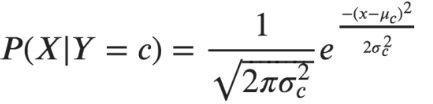
1. **Write the difference between the following**

a.

* **Gaussian Naive Bayes**

Naive Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naive Bayes is the extension of Naive Bayes. Gaussian Naive Bayes is simplest to implement as you will need to calculate the mean and standard deviation for the training data.

That is Gaussian distribution. To compute this, we need the mean and variance of X.



* **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.

* **Complement Naive Bayes**

Complement Naive Bayes is an adaptation of the standard Multinomial Naive Bayes algorithm. Multinomial Naive Bayes does not perform very well on imbalanced datasets. Imbalanced datasets are datasets where the number of examples of some class is higher than the number of examples belonging to other classes. This means that the distribution of examples is not uniform.

* **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.

* **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.

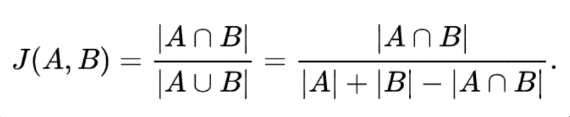
* **Out-of-core Naive Bayes model fitting**

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.

**c. What is Jaccard and Cosine similarity?**

**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.



**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.

