**ML Assignment 3**

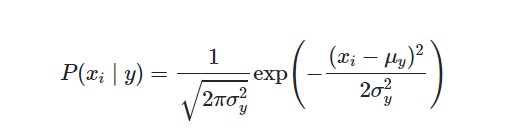
**Jugvi Mankad**

**21BDA49**

**1.Gaussian Naïve Bayes:**

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.

The likelihood of the features is assumed to be-



Sometimes assume variance

* is independent of Y (i.e., σi),
* or independent of Xi (i.e., σk)
* or both (i.e., σ)

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.

**2.Multinomial Naïve Bayes:**

Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often i.e. frequency.

For example, it 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.

**3.Complement Naïve Bayes:**

Complement Naive Bayes is particularly suited to work with imbalanced datasets. In complement Naive Bayes, 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 Naive Bayes.

**4.Bernoulli Naïve Bayes:**

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:

* True or False
* Spam or Ham
* Yes or No
* 0 or 1

**5.Categorical Naïve Bayes:**

Categorical distribution is a special case of the multinomial distribution, in that it gives the probabilities of potential outcomes of a single drawing rather than multiple drawings.

Categorical naïve bayes 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.

**6. Out-of-Core Naive Bayes Model Fitting:**

Out-of-core (or “external memory”) learning is a technique used to learn from data that cannot fit in a computer’s main memory (RAM).

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, MultinomialNB, BernoulliNB, and GaussianNB expose a partial\_fit method that can be used incrementally as done with other classifiers

C.

**Cosine Similarity:**

Cosine similarity is a metric, helpful in determining, how similar the data objects are irrespective of their size. We can measure the similarity between two sentences in Python using Cosine Similarity. In cosine similarity, data objects in a dataset are treated as a vector. The formula to find the cosine similarity between two vectors is –

Cos(x, y) = x . y / ||x|| \* ||y||

where,

x . y = product (dot) of the vectors ‘x’ and ‘y’.

||x|| and ||y|| = length of the two vectors ‘x’ and ‘y’.

||x|| \* ||y|| = cross product of the two vectors ‘x’ and ‘y’.

**Jaccard Similarity:**

The Jaccard Similarity Index is a measure of the similarity between two sets of data.

Developed by Paul Jaccard, the index ranges from 0 to 1. The closer to 1, the more similar the two sets of data.

The Jaccard similarity index is calculated as:

Jaccard Similarity = (number of observations in both sets) / (number in either set)

Or, written in notation form:

J(A, B) = |A∩B| / |A∪B|

If two datasets share the exact same members, their Jaccard Similarity Index will be 1. Conversely, if they have no members in common, then their similarity will be 0.

In cases where the vectors A and B are comprised 0s and 1s only, **Cosine Similarity** divides the number of common attributes by the product of A and B's distance from zero. Whereas in **Jaccard Similarity**, the number of common attributes is divided by the number of attributes that exists in at least one of the two objects.