Lab 3 Assignment

Hendrick Cardoso 21BDA20

Write the difference between the following:

i Gaussian Naive Bayes

 Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.

ii Multinomial Naive Bayes

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.

iii 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. Hence is called Complement Naive Bayes.

iv Bernoulli Naive Bayes

 It implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions. There may be multiple features but each one is assumed to be a binary-valued variable.

v. Categorical Naive Bayes

 It is suitable for classification with discrete features that are categorically distributed. The categories of each feature are drawn from a categorical distribution.

vi. 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, 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](https://scikit-learn.org/stable/auto_examples/applications/plot_out_of_core_classification.html#sphx-glr-auto-examples-applications-plot-out-of-core-classification-py).

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

Reference: <https://scikit-learn.org/stable/modules/naive_bayes.html>

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. Jaccard similarity 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.

Reference: <https://www.sciencedirect.com/topics/computer-science/cosine-similarity>