Objective: Study some techniques for Naive Bayes Text Classification

Research paper- https://drive.google.com/file/d/1N-hYMNwy18MuiH_HnGQogrXqna8_d9cN/view?usp=sharing

Team members- Krishna Paswan, Depanshu Sahu, and Chirag Garg. (Grp – 3)

Introduction:

The Naive Bayes classifier is considered a simple yet effective machine learning model, despite relying on an unrealistic independence assumption. It discusses the limitations of the traditional multivariate Bernoulli Naive Bayes model, which does not consider term frequencies. To address these limitations, the Multinomial model is proposed as an alternative. However, the Multinomial model faces challenges in estimating word probabilities and handling rare categories. The paper introduces a per-document length normalization approach using a multivariate Poisson model and a weight-enhancing method to improve performance in rare categories and address parameter reliability.

Algorithm:

The algorithm assumes that all attributes (features) of examples are independent given the class context, although this assumption has been found to perform well in various domains.

PROPOSED METHODS FOR IMPROVING NAÏVE BAYES TEXT CLASSIFIER -

The paper presents methods to improve the Naive Bayes text classifier by introducing the multivariate Poisson model as a more flexible alternative to the traditional multinomial model which can be expanded by adopting the various methods to estimate Poisson mean and average numbers of times a term appears in a document.

PARAMETER ESTIMATION USING NORMALIZED TERM FREQUENCIES-

Parameter estimation using normalized term frequencies and per-document length normalization is proposed as an approach to improve the multinomial model. The feature weighting approach is suggested as an alternative to feature selection, addressing practical issues and drawbacks associated with binary features.

CONSIDERATIONS OF FEATURE WEIGHTS-

Determining the appropriate number of features for each category while maintaining balance is a crucial task. Additionally, selected features are often treated equally without considering their relative importance as a cutoff criterion. To address these challenges and mitigate the drawbacks associated with binary features, the proposed approach suggests a feature weighting method instead of feature selection.

The analysis reveals that all variants of the proposed classifiers exhibit a significant performance improvement compared to the traditional multinomial Naive Bayes classifier, particularly when applied to the specified collection. Notably, the proposed classifier achieves notably higher macroF1 performance, even when the normalization parameter is set to 1.0. This finding suggests that term frequencies are not normalized based on the lengths of documents in which they appear.

Results:

EFFECT OF PER-DOCUMENT LENGTH NORMALIZATION -

The findings reveal that per-document length normalization proves to be highly effective, especially in scenarios where training documents are scarce. All per-length normalization methods exhibit significantly superior performance when compared to the traditional multinomial text classifier.

EFFECT OF FEATURE WEIGHTS-

The performance of the weight-enhanced Poisson classifier is evaluated, focusing on its effect on macroF1 performance. The results demonstrate a substantial improvement in macroF1 performance compared to the pure Poisson classifier. However, the microF1 performances yield less impressive outcomes, indicating that the weight-enhanced method primarily benefits rare categories rather than overall classification performance.

Conclusion:

Based on the experimental results from two distinct collections, the proposed model emerges as a highly valuable tool for constructing probabilistic text classifiers, requiring minimal additional time and space compared to traditional multinomial classifiers. In scenarios where training documents are scarce, enhancing classification weights offers a compelling alternative to the traditional Naive Bayes classifier. Although the proposed Naive Bayes text classifier does not surpass SVM classifiers, its simplicity, efficiency, and ability for incremental learning make it well-suited for practical systems such as spam filtering or adaptive news-alert systems.