An Empirical Study of Linear Separability on Authorship Attribution Feature Spaces

John Noecker Jr. // Duquesne University, Pittsburgh PA // jnoecker@gmail.com

Patrick Juola // Duquesne University, Pittsburgh PA // juola@mathcs.duq.edu

In the field of authorship attribution through statistical analysis of text documents, a variety of methods have been proposed with varying levels of success. The Ad-hoc Authorship Attribution Competition (AAAC), an experiment in authorship attribution held as part of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, provides us with results from a variety of different analysis techniques, but does not provide us with any information on why one technique works while another does not. In order to determine the characteristics of a good analysis technique, we need additional insight into the structure of the feature spaces induced by these authorship attribution methods. We will study this structure through an empirical comparison of the performance of two analysis methods, LDA and SVM, in order to determine which common authorship attribution feature spaces, if any, are linearly separable.

Whether or not a space is linearly separable can have important consequences on performing classification within that space. An n-dimensional space containing two classes of points is said to be linearly separable if there exists an n-1 dimensional hyperplane which separates the classes. A linearly separable space has the advantage that simpler classification algorithms will work for classifying points within that space. A simple distance metric may be sufficient to distinguish between two classes in a linearly separable space, while more complex methods like support vector machines or neural networks are necessary to capture nonlinear class boundaries. Linear Discriminant Analysis (LDA) is a technique for performing classification by constructing a separating hyperplane which maximizes the margins between the classes. The margins of a hyperplane are the distances from the hyperplane to the hypothesized class boundaries. Thus, LDA is effective only when the feature space is linearly separable. Support vector machines (SVM) work in a similar way, by constructing a hyperplane and maximizing the margins between the hyperplane and the classes. However, support vector machines use a kernel function to implicitly map the feature space into a higher-dimensional space in which the data is more likely to be linearly separable. Thus, SVMs are effective on both spaces which are linearly separable and those which are not. We therefore propose that if LDA performs comparatively with SVMs, the feature space is likely to be linearly separable and the kernel function may be unnecessary baggage. If SVMs considerably outperform LDA, it seems likely that the original feature space is not linearly separable. This work is an outgrowth from a discovery made at the 2008 CLSP Summer Workshop on Human Language Technology at the Johns Hopkins University, where an experiment in how various analysis methods performed at the task of speaker verification provided valuable insight into the structure of the speaker feature space, allowing for a more efficient analysis. As a result of this research, new methods of speaker verification were developed which are considerably faster than traditional methods and have only a moderate decrease in performance. We hope to find similar results for author spaces, or to find some explanation of the difference between speaker and author spaces.

The task of authorship attribution is to assign authorship labels to a set of documents by performing statistical analysis on another set of documents for which the authorship labels are known. One common method is to extract statistics about each author's use of function words and use these statistics to compute the likelihood that a given author composed an unlabeled work. A function word is a short, common word that holds little or no meaning itself, and thus is independent of nuisances like the genre or length of the documents. The function words extracted from each document are the features. These features can be represented numerically, as a matrix of relative occurrences within a document, for example, and the set of

these matrices across a collection of documents induces a feature space. Statistical analyses like LDA and SVM can then be performed on this feature space in order to model class boundaries. LDA will generate a separating hyperplane within the feature space, while SVMs are capable of modeling more complex boundaries. Unlabeled documents may then be embedded into this space, and a label is assigned based on where these documents are mapped in relation to the class models. Hence, the accuracy of the assigned labels depends entirely on the assumptions of the statistical techniques used to model the authors. Since LDA makes an assumption of linear separability that is not made by SVM, the relative performance of these analysis methods can be viewed as an affirmation or rejection of this assumption.

We will use JGAAP (The Java Graphical Authorship Attribution Program - *www.jgaap.com*), a freely available Java program for performing authorship attribution, to evaluate the relative performance of LDA and SVM on a variety of feature sets. The AAAC corpus provides texts from a wide variety of different genres, languages and document lengths, and will serve as our standard test corpus for the experiments. By using such a varied corpus, we can make both general statements about certain feature sets as well as statements about a feature set within a certain category such as language or genre. After performing processing to remove punctuation, capitalization and other nuisances from the text, we will explore the performance of LDA and SVM on the AAAC corpus for various types of feature spaces. These feature spaces will include function word spaces, the space of all words across all documents, character-level spaces and some word or character n-gram features. Although numerous studies have applied SVM or LDA to the problem of authorship attribution, these studies do not use a single unified corpus or a standardized feature set, making comparison impossible. We will also use the radial basis kernel function for the SVMs, which will ensure that they are able to model complex, nonlinear class boundaries. If LDA performs comparatively to or better than SVMs for a given feature space, we can conclude empirically that the assumption of linear separability was valid. If SVMs perform significantly better than LDA for a given feature set, it seems likely that the assumption of linear separability was not valid, as the additional freedom provided by the SVM allowed for a better authorship model.

We expect that at least some of the feature spaces will be linearly separable, and hope to find feature spaces that are both linearly separable and useful in performing authorship attribution. If instead we find that SVM vastly outperforms LDA and none of the feature spaces are linearly separable, we plan to examine why the authorship attribution factor spaces are so different from those used in speaker verification, which the 2008 CLSP Workshop at the Johns Hopkins University results suggest are linearly separable. Regardless of the outcome of these experiments, we hope to provide the scholarly community with new insight on the structure of the feature spaces used for authorship attribution. This insight will allow us to create more efficient and accurate analysis methods which are custom-tailored to the task of authorship attribution. Even if no overall trend of linear separability is observed, we expect that we may be able to raise questions about the linear separability of features of one specific kind of document. Perhaps English texts are not linearly separable but Latin texts are. The widely varied nature of the AAAC corpus makes such observations possible, and we hope this will lead to a variety of interesting questions for later investigation.

By comparing the relative performance of Linear Discriminant Analysis and Support Vector Machines with a Radial Basis Kernel on documents from a variety genres, languages and lengths, we will empirically determine which feature spaces commonly used in authorship attribution are linearly separable. We will determine a feature space to be linearly separable if the performance of LDA is comparable to the SVM performance in that feature space. We claim that this is a valid empirical argument because LDA makes an assumption of linear separability which is not present in the SVM. In the absence of linearly separable feature spaces, we will attempt to glean some knowledge of the underlying structure and to explain why author spaces are not linearly separable while the speaker factor spaces used in speaker identification are. We will attempt to draw conclusions about the general distribution of authors within feature spaces composed of various combinations of function words as well as word and characters and N-grams constructed from characters and words. A knowledge of the structure of these feature spaces will assist the scholarly community in choosing the appropriate analysis methods for a given feature set, resulting in more accurate and efficient statistical analysis.