Application of Sequential Linear Discriminant Analysis to Biological Data
Colin Brown, Manchester Community College and Brittany Nkounkou, University of Connecticut

Classification algorithms are used to assign an input item to a certain class, based on quantitative information of one or more variables. They are determined by observing the same quantitative measurements of several instances whose class assignments are previously known. Classification algorithms can be applied to biological data, for example, to determine whether or not a tissue sample is cancerous, based on different measurements.

LDA works by considering all of the variable values of the instance to be classified, to determine the likelihood that it should be in each class. The instance is then labeled as the class with the largest likelihood. SPRT, on the other hand, observes each variable of the instance one at a time. At each iteration, the algorithm computes the product of the likelihoods of the current and previous variables observed. If the value of the current product is outside of a specified range, then iteration stops and the instance is labeled to a class. If not, the algorithm moves on to the next variable. The minimum and maximum of the range for each variable are determined by alpha and beta, the preferred error rates for each class. If there are no more variables to observe, or the truncation limit has been reached (the maximum number of variables desired to be observed) then the instance’s class assignment is determined by whether the final likelihood product is in the upper or lower half of the specified range.

We plan to implement different classification algorithms, including Linear Discriminant Analysis (LDA) and Sequential Probability Ratio Test (SPRT), and compare their respective accuracies. SPRT, which is essentially Sequential LDA, takes three adjustable parameters: an error rate for each of the two classes and an input specifying truncation. In addition to comparing SPRT to LDA, we’ll compare the accuracy of SPRT by varying each of its three variables. We’ll also observe the effects of ordering the observation of variables in SPRT, for example, in decreasing variance order. Finally, we aim to extend binary SPRT to handle multi-class problems.

We’ll implement the LDA and SPRT algorithms in Python, an interpreted programming language. One method of analyzing their accuracies is by exercising leave-one-out cross-validation. This is a very common technique used to estimate the performance of a predictive function, such as a classification algorithm. This method takes all but one instance of the preliminary data to train the function. It then performs the classification algorithm on the instance left out. The result is then compared to the previously known class label of that instance. This procedure is repeated for each instance. The accuracy of the algorithm is then given by the percentage of instances that were classified correctly.

We’ll train our classification algorithms using a Parkinson’s dataset [1] created by Max Little of the University of Oxford. We may research other larger datasets in order to observe any effect the size of the dataset may have on the performance of our algorithms. In addition to real datasets, we’ll also generate simulated datasets, and compare their results to those of the real datasets. We have already implemented LDA and SPRT with variables in decreasing variance order. We performed leave-one-out cross-validation on both of them, using real data [1] and different values of alpha and beta for SPRT. The following results were obtained:

LDA: 0.687179487179

SPRT RESULTS:
alpha=beta=0.4: 0.74358974359
alpha=beta=0.3: 0.748717948718
alpha=beta=0.2: 0.789743589744
alpha=beta=0.1: 0.728205128205

References
1. Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering