Comparison of LDA and SPRT using Biological Datasets

Colin Brown
Manchester Community College

Brittany Nkounkou
University of Connecticut
Statistical classification is a method of machine learning where items are grouped based on quantitative information that is known about the items and based on a training set of data with known statuses.

Given a training data set \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) to produce a classifier for this data set such that \( h : x \to y \) is able to map any object \( x \in \chi \) to its true status \( y \in \gamma \).
The problem then becomes which algorithm should be used to determine the distinct probability. K-nearest neighbor is one of the most commonly used algorithms because it is a simple algorithm. Another common method is to use logistic regression which tries to fit the probability of data being in a particular class by fitting the data to a logistic curve. Our project uses Linear Discriminant Analysis which relies on some Bayesian principles that are used when the covariances in a dispersion matrix are equal to 0, which implies independence.

\[ P(\text{parkinsons}|x) = f(x; \theta) \] where \( x \) is the vector input to the function and \( \theta \) is a parameter. This being such it is easy to determine the opposite status as well \( P(\text{noparkinsons}|x) = 1 - P(\text{parkinsons}) \).
Methodology

Classification Algorithms
- Linear Discriminant Analysis (LDA)
- Sequential Probability Ratio Test (SPRT)
- Modified SPRT (MSPRT)
- Component Sampling
- Leave-One-Out Cross-Validation
The likelihood that instance $\mathbf{x}$ is in class $g$

$$
\Pr(\mathbf{X} = \mathbf{x} | G = g) \propto \exp \left( -\frac{1}{2} (\mathbf{x} - \mu_g)^T \Sigma^{-1} (\mathbf{x} - \mu_g) \right)
$$

where $\mathbf{x}$ is the component vector of an instance belonging to class $g$ with mean vector $\mu_g$ and dispersion matrix $\Sigma$ common to all classes. The parameters $\mu_g$’s and $\Sigma$ are estimated with a training dataset.

Note: All components of the vectors are used in the computation.
For binary classification, compute the likelihood ratio:

\[ \Lambda = \frac{\Pr(X = x | G = 1)}{\Pr(X = x | G = 2)} \]

The instance \( x \) is assigned to class 1 if \( \Lambda > 1 \) or class 2 if \( \Lambda < 1 \).
Sequential Probability Ratio Test

The likelihood ratio is computed one component at a time:

\[
\Lambda = \frac{\Pr(X = x|G = 1)}{\Pr(X = x|G = 2)} = \prod_{i=1}^{p} \frac{\phi\left(\frac{x_i - \mu_{1i}}{\sigma_i}\right)}{\phi\left(\frac{x_i - \mu_{2i}}{\sigma_i}\right)}
\]

where \( \phi(\cdot) \) is the pdf of the standard normal distribution, \( \mu_{1i} \)'s and \( \mu_{2i} \)'s are components of \( \mu_1 \) and \( \mu_2 \), respectively, and \( \sigma_i \)'s are the standard deviations of the \( x_i \)'s.
For easier computation, we equivalently have:

$$\log \Lambda = \sum_{i=1}^{p} \log \left( \frac{\phi \left( \frac{x_i - \mu_{1i}}{\sigma_i} \right)}{\phi \left( \frac{x_i - \mu_{2i}}{\sigma_i} \right)} \right) = \sum_{i=1}^{p} z_i$$

where \( \log \Lambda \) is iteratively computed until it falls outside of range \((a,b)\).

We assign \(\mathbf{x}\) to class 1 if \(\log \Lambda < a\) or to class 2 if \(\log \Lambda > b\).
The decision boundaries $a$ and $b$ are computed from inputs $a$ and $\beta$, which are the desired error rates for class 1 and class 2, respectively, such that $0 < a, \beta < 1$. The decision boundaries are then given by:

$$a = \log \left( \frac{1 - \beta}{\alpha} \right)$$

$$b = \log \left( \frac{\beta}{1 - \alpha} \right)$$
SPRT cont’d.

What if $\log \Lambda$ is still inside the range $(a,b)$ after sampling the last component? In this case, truncation is utilized.

The decision boundary is truncated to the midpoint between $a$ and $b$. That is, instance $\mathbf{x}$ is assigned to class 1 if $\log \Lambda < \frac{1}{2}(a+b)$ or to class 2 if $\log \Lambda > \frac{1}{2}(a+b)$.

Truncation may also be desired at any integer $k$ such that $1 < k \leq p$ (i.e. to reduce run time).
Decision boundaries $a$ and $b$ are no longer constant. The decision boundaries at the $i$th component examined are given by:

$$a_i = a \left(1 - \frac{i}{k}\right)^{r_1} \quad b_i = b \left(1 - \frac{i}{k}\right)^{r_2}$$

where $0 < r_1, r_2 \leq 1$, $a > 0$, $b > 0$, and $k$ is the truncation specification. When running our algorithms, we simply set $r_1 = r_2 = 1$ each time. As $i \to k$, $a_i$ and $b_i \to 0$, thus ensuring that a decision will be made by the $k$th component.
How should the components be ordered when sampling in SPRT and MSPRT?

- Decreasing variance order
- Random order

Since the algorithms with components in random order were not pre-deterministic, their results were averaged over 100 runs.
Leave-One-Out Cross-Validation

- Exclude one instance
- Use the remaining instances as training data for the algorithm
- Predict the classification of the excluded instance
- Compare the result to its actual classification
- Repeat with all instances of the dataset
- Algorithm accuracy = percentage of instances predicted correctly
Binary Biological Datasets

Parkinson’s Disease Dataset
- 195 instances
- 22 components

Colon Cancer Dataset
- 63 instances
- 500 components

Breast Cancer Dataset
- 683 instances
- 10 components
SPRT ordered and MSPRT ordered were run with $\alpha=\beta$ from 0.01 to 0.40 with a step size of 0.01. The accuracies for SPRT random and MSPRT random were averaged over 100 runs each at $\alpha=\beta=0.175$ and $\alpha=\beta=0.32$. The highest accuracy for the Parkinson’s disease dataset was 0.7128, generated by the MSPRT random function at $\alpha=\beta=0.175$ and by the SPRT random function at $\alpha=\beta=0.32$. 
SPRT ordered and MSPRT ordered were run with $\alpha=\beta$ from 0.01 to 0.40 with a step size of 0.01. The accuracies for SPRT random and MSPRT random were averaged over 100 runs each at $\alpha=\beta=0.12$ and $\alpha=\beta=0.22$. The highest accuracy for the colon cancer dataset was 0.8871, generated by the MSPRT ordered function at $\alpha=\beta=0.22$. 

Colon Cancer Results
SPRT ordered and MSPRT ordered were run with $\alpha=\beta$ from 0.01 to 0.40 with a step size of 0.01. The accuracies for SPRT random and MSPRT random were averaged over 100 runs each at $\alpha=\beta=0.07$ and $\alpha=\beta=0.09$. The highest accuracy for the Parkinson’s disease dataset was 0.9648, generated by the MSPRT ordered function at $\alpha=\beta=0.07$ and by the SPRT ordered function at $\alpha=\beta=0.09$. 
Discussion of Results

- With the Parkinson’s dataset we had high accuracy using both SPRT random and MSPRT random, while for most $\alpha$ and $\beta$ values SPRT ordered was less than the LDA method. At its highest point MSPRT ordered only reached the LDA accuracy.

- In the Colon Cancer dataset, MSPRT ordered reached the maximum accuracy rate in the set, while SPRT ordered was just under its accuracy rate. However, in the case of MSPRT and SPRT random, they only reached the LDA accuracy rate and MSPRT random fell below the accuracy rating of LDA.

- Finally, in the Breast Cancer dataset, the accuracy of SPRT and MSPRT ordered both reached peaks in accuracy above the LDA accuracy rate at different $\alpha$ and $\beta$ values, but both reached the same peak in accuracy. SPRT and MSPRT random both fell short of the LDA accuracy rating, MSPRT random getting lower accuracies both times the tests were ran.
Possible Conclusions we can draw:

- The results of the datasets might suggest that in sets that have high accuracy using the LDA method, have a greater accuracy rating from SPRT and MSPRT ordered.

- Another conclusion that can possibly be rendered is that in datasets with low accuracy MSPRT random and SPRT random have higher accuracy than a standard LDA.
Another thing that can be noticed in the datasets is that SPRT and MSPRT both random and ordered may have some correlation.

In lower accuracy datasets such as the Parkinson’s set, MSPRT ordered has a lower accuracy rating than SPRT ordered, however in higher to mid-high accuracy datasets such as the breast and colon cancer datasets, MSPRT ordered has somewhat higher accuracy ratings than SPRT ordered. However with MSPRT and SPRT random, with high to mid-high accuracy datasets, the accuracy of MSPRT random seems to always be lower than the SPRT random.

The conclusion that can be learned from all of this though is that accuracy ratings of various testing methods is largely dependent on the dataset that you are using and the accuracy ratings and correlation between status and other variables.
References

- 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