Notes on Linear Discriminant Analysis

Earl Patrick B. Macalam

December 11, 2020

1 Why Linear Discriminant Analysis?

1. When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.

2. If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.

3. Linear discriminant analysis is popular when we have more than two response classes.

2 Linear Discriminant Analysis

- Uses Baye's Theorem for Classification.

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

  Where π_k refers to the base probability of each class k observed in your training data. In Bayes' Theorem this is called the prior probability

$$\pi_k = \frac{n_k}{n}$$

  The $f(x)$ above is the estimated probability of x belonging to the class. A Gaussian distribution function is used for $f(x)$.

- Estimating μ_k:

$$\mu_k = \frac{1}{nk} \cdot \sum x$$

  Where μ_k is the mean value of x for the class k, n_k is the number of instances with class k.

- Estimating σ^2:
$$\sigma^2 = \frac{1}{n-K} \cdot \sum (x-\mu)^2$$
Where σ^2 is the variance across all inputs x, n is the number of instances, K is the number of classes and mu is the mean for input x.

- LDA makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made.

3 Assumptions

1. That your data is Gaussian, that each variable is is shaped like a bell curve when plotted.

2. That each attribute has the same variance, that values of each variable vary around the mean by the same amount on average.

4 Data Preparation

1. **Classification Problems**. This might go without saying, but LDA is intended for classification problems where the output variable is categorical. LDA supports both binary and multi-class classification.

2. **Gaussian Distribution**. The standard implementation of the model assumes a Gaussian distribution of the input variables. Consider reviewing the univariate distributions of each attribute and using transforms to make them more Gaussian-looking (e.g. log and root for exponential distributions and Box-Cox for skewed distributions).

3. **Remove Outliers**. Consider removing outliers from your data. These can skew the basic statistics used to separate classes in LDA such the mean and the standard deviation.

4. **Same Variance**. LDA assumes that each input variable has the same variance. It is almost always a good idea to standardize your data before using LDA so that it has a mean of 0 and a standard deviation of 1.

5 Extensions to LDA

1. **Quadratic Discriminant Analysis (QDA)**. Each class uses its own estimate of variance (or covariance when there are multiple input variables).

2. **Flexible Discriminant Analysis (FDA)**. Where non-linear combinations of inputs is used such as splines.

3. **Regularized Discriminant Analysis (RDA)**. Introduces regularization into the estimate of the variance (actually covariance), moderating the influence of different variables on LDA.