Classification: Linear Discriminant Analysis

- Discriminant analysis uses sample information about individuals that are *known* to belong to one of several populations for the purposes of *classification*.
- Based on the variable values for these individuals (called the *training data*) whose population memberships are known, we determine a *rule* to classify new individuals (called the *test data*) whose population memberships are *unknown*.
- We can collect variable values for a new individual and use the data and this *classification rule* to classify the new individual (i.e., predict which population it belongs to).
- When there are two populations, Fisher's Linear Discriminant Function specifies a linear function of the q variables that best separates the sampled individuals into two groups.
- The classification rule is then based on that linear combination of the variables.

Examples of Classification

- Admissions officials in colleges attempt to use data on applicants (GPA, SAT scores, etc.) to classify them into two groups: those who successfully graduate and those who fail to graduate.
- They can use the data on previous years' applicants (who are known to have either graduated or not) to build a classification rule.
- Archaeologists and zooarchaeologists attempt to classify animal remains into groups (such as male/female) based on measurements on bones.
- Marketing experts can use data on past potential customers to classify future individuals as likely buyers or unlikely buyers.
- Lifetime data can only be measured by destroying an object. We may want to classify an object as defective or not (without destroying it) using certain preliminary measurements, after obtaining a small sample of objects that have been seen to be defective or not.

Mathematical Details of Fisher's Linear Discriminant Analysis (LDA)

- We assume that we are classifying individuals into one of two known groups based on their values of the variables $x_1, x_2, \ldots, x_q$.
- We have the data (on the q variables) for n_1 individuals that are known to belong to Group 1 and for n_2 individuals that are known to belong to Group 2.
- We will use the linear combination of these variables:

$$z = a_1x_1 + a_2x_2 + \cdots + a_qx_q$$

  that maximizes the ratio of between-group variance of z to within-group variance of z for the observed sample of individuals whose groups are known.
- See geometric interpretation in $q = 2$ dimensions.

Mathematical Details of LDA (continued)

- That is, we choose $\mathbf{a} = (a_1, \ldots, a_q)'$ to maximize

$$V = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}},$$

  where $\mathbf{S}$ is the pooled within-group sample covariance matrix, and $\mathbf{B}$ is the covariance matrix of the group sample means (see p. 143 for formulas).

- For two groups, the choice of $\mathbf{a}$ that maximizes V is

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

The LDA Classification Rule

- Let $z = \mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + \cdots + a_qx_q$ be the discriminant score for any individual.
- Also let $\bar{z}_1$ be the sample mean of the discriminant scores for the individuals known to come from Group 1, and let $\bar{z}_2$ be the mean discriminant score for the individuals known to come from Group 2.
- *Case I: Suppose* $\bar{z}_1 > \bar{z}_2$*.* For a new individual with discriminant score z^*, classify this individual into Group 1 if $z^* > (\bar{z}_1 + \bar{z}_2)/2$; otherwise classify it into Group 2.
- *Case II: Suppose* $\bar{z}_1 < \bar{z}_2$*.* For a new individual with discriminant score z^*, classify this individual into Group 1 if $z^* < (\bar{z}_1 + \bar{z}_2)/2$; otherwise classify it into Group 2.
- The LDA method is implemented in `R` by the `lda` function in the `MASS` package.

Remarks about the LDA Classification Rule

- This LDA approach assumes the data come from one of two multivariate normal populations, each with the same covariance matrix.
- This LDA rule is equivalent to classifying an individual having data vector $\mathbf{x}$ into Group 1 if and only if $MVN(\mathbf{x}, \bar{\boldsymbol{x}}_1, \mathbf{S}) > MVN(\mathbf{x}, \bar{\boldsymbol{x}}_2, \mathbf{S})$, where MVN represents the multivariate normal density function.
- This rule is appropriate only if the prior probabilities of the individual being in each group are equal.
- If the prior probabilities p_1 and p_2, are not equal, then $z^* - (\bar{z}_1 + \bar{z}_2)/2$ would be compared to $\ln(p_2/p_1)$ rather than to 0 as before.
- Also, the rule assumes the *cost of misclassification* is the same whether an individual is misclassified into Group 1 or misclassified into Group 2.

Generalizations to More than Two Groups

- When we have three or more groups, the classification rule can be generalized.
- If there are k (multivariate normal) populations, then we classify an individual having data vector $\mathbf{x}$ into Group i if and only if $MVN(\mathbf{x}, \bar{\boldsymbol{x}}_i, \mathbf{S}) > MVN(\mathbf{x}, \bar{\boldsymbol{x}}_j, \mathbf{S})$ for all $j \neq i$, where MVN represents the multivariate normal density function.
- For three groups, this is equivalent to basing the classification on a series of pairwise rules for choosing between Groups 1 and 2, between Groups 1 and 3, and between Groups 2 and 3 (see pp. 150-151 for details).
- This rule could actually be further generalized to the case in which the prior probabilities of being in each group are not equal, and in which the populations have some known non-normal distributions:
- Classify an individual having data vector $\mathbf{x}$ into Group i if and only if $p_i f_i(\mathbf{x}) > p_j f_j(\mathbf{x})$ for all $j \neq i$, where $f_j(\cdot)$ is the j-th of the density functions.

Other Types of Discriminant Analysis

- When the covariance matrices for Populations 1 and 2 are not believed to be equal, then *quadratic discriminant analysis* (QDA) is more appropriate than LDA.
- QDA is more flexible than LDA, but it can often *overfit* the observed data, creating a rule that classifies the known observations nearly perfectly but that is not as generalizable to future observations.
- QDA is implemented by the `R` function `qda` in the `MASS` package.
- A compromise between LDA and QDA is *regularized discriminant analysis*.
- When the populations are not close to multivariate normal, an alternative to LDA is *logistic discrimination*.

Judging the Performance of the Discriminant Function

- To judge how well the discriminant rule is classifying, we could calculate the "plug-in" misclassification rate, which is simply the proportion of the "known" individuals that would be misclassified if we used the rule to classify them.
- Since the rule has been derived using those known individuals, the "plug-in" rate typically is too optimistic — it underestimates the rate at which the rule would misclassify *new* individuals.
- A better approach is the *cross-validation* (or "leave-one-out") method.
- This uses all but one of the "known" individuals to derive a classification rule and then, based on that rule, classifies the other individual.
- This is done (separately) for all the known individuals, and the misclassification rate is the proportion of those classifications that are incorrect.
- The `R` function `predict` can help produce these classification rates.

Using Regression Methods for Classification

- In classification problems, we use one or more numerical variables to *predict* the "category" of an individual with respect to some categorical variable.
- Since a major purpose of regression is prediction, could we use regression techniques to classify individuals?
- One option: Code the categories as numerical values (1, 2, 3, . . .) and use linear regression to predict the category based on the numerical explanatory variables.
- Problem: We must specify an ordering of the categories, which may be arbitrary.
- It is easier with binary (two-category) response data, as the ordering is not as important.

Logistic Regression

- A better regression approach is to use a method designed for categorical responses: logistic regression.
- This works best when there are only two categories: We can code them as $Y = 0$ or $Y = 1$.
- A logistic regression predicts the probability that $Y = 1$, given a value of some explanatory variable X:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- This model (unlike linear regression) will always produce a predicted probability between 0 and 1.
- The parameters β_0 and β_1 are estimated from the training data using the *maximum likelihood* method.
- This can be done easily in `R` using the `glm` function.

Classifications Based on Logistic Regression

- Assuming two groups, a simple classification rule is to predict that $Y = 1$ for any individual whose predicted probability that $Y = 1$ is greater than 0.5.
- If $P(Y = 1) \leq 0.5$, we would predict $Y = 0$ for that individual.
- If the number of individuals in the population having $Y = 1$ is believed to differ from the number having $Y = 0$, then we could adjust our cutoff value away from 0.5.
- One option: Predict $Y = 1$ if $P(Y = 1) > p^*$, where p^* is chosen to minimize the misclassification rate for the individuals in the training set.

Multiple Logistic Regression

- If we have several numerical variables measured on each individual, we can generalize our logistic regression model to:

$$P(Y = 1|X_1, \ldots, X_q) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_q X_q}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_q X_q}}$$

- Z-tests about the regression coefficients assess the importance of each individual explanatory variable on the classification.

- The logistic model can be extended to situations where we are classifying into more than 2 categories, but it is more common to use other classification methods (like discriminant analysis) in those cases.