CPH 636 — Spring 2009 — Dr. Charnigo

Lecture 7

Overview of discriminant analysis

A motivating problem. Suppose that 10% of the healthy adults in a certain town will develop heart disease in the next ten years. Suppose, moreover, that the distribution of systolic blood pressure (SBP) among those who will develop heart disease is normal with mean 150 and standard deviation 15, while the distribution of SBP among those who will not develop heart disease is normal with mean 130 and standard deviation 10. For convenience, we will use the shorthand $N(150, 15^2)$ and $N(130, 10^2)$ to represent these two normal distributions.

If a randomly selected healthy adult in this town has SBP = 139, is this someone for whom we would predict heart disease in the next ten years? Here, for simplicity, we are putting aside other prognostic factors like family history, tobacco use, and age.

First attempt at solving the motivating problem. Noting that 139 is closer to 130 than to 150, our first reaction may be to say that the 139 is more consistent with a $N(130, 10^2)$ distribution than a $N(150, 15^2)$ distribution, in which case we would not predict heart disease.

Unfortunately, this logic is flawed. There is more variation in the $N(150, 15^2)$ distribution than in the $N(130, 10^2)$ distribution. So, for instance, a 10-point deviation from the mean of 150 is not as extreme as a 10-point deviation from the mean of 130. But how does an 11-point deviation from the mean of 150 compare to a 9-point deviation from the mean of 130?

Second attempt at solving the motivating problem. Noting that 139 has a Z score of $(139 - 130)/10 = 0.90$ in relation to the $N(130, 10^2)$ distribution and a Z score of $(139 - 150)/15 = -0.73$ in relation to the $N(150, 15^2)$ distribution, our second reaction may be to say that the 139 is somewhat more consistent with the $N(150, 15^2)$ distribution than with the $N(130, 10^2)$ distribution, in which case we <u>would</u> predict heart disease.

Unfortunately, this logic also is flawed. There are many fewer people in the subpopulation described by the $N(150, 15^2)$ distribution than in the subpopulation described by the $N(130, 10^2)$ distribution. Only if 60% (rather than 10%) were in the first subpopulation would this approach to prediction work. [[In Written Assignment 4 you will prove the more general result that, with obvious notation, a comparison of Z scores is reasonable only when the membership fraction of the first subpopulation is $\sigma_1/(\sigma_1 + \sigma_2)$.]]

Neither of our first two attempts at solving the motivating problem has involved a calculation of the <u>probability</u> that the SBP of 139 originated from the $N(150, 15^2)$ distribution. Such a calculation would be of interest in its own right. For instance, imagine a person being told that he had a 35% chance of developing heart disease in the next ten years. He might be predicted not to have heart disease, but with a chance as high as 35% he might be seriously motivated to exercise more, eat less, or quit smoking.

Third time's the charm. We can compute the probability that the SBP of 139 originated from the $N(150, 15^2)$ distribution. This probability is

$$\frac{0.10 \times 15^{-1} \exp\left[-\frac{(139-150)^2}{(2\times 15^2)}\right]}{0.10 \times 15^{-1} \exp\left[-\frac{(139-150)^2}{(2\times 15^2)}\right] + 0.90 \times 10^{-1} \exp\left[-\frac{(139-130)^2}{(2\times 10^2)}\right]} = 0.078. \quad (1)$$

On this basis, the person with the SBP of 139 would not be unreasonably concerned about developing heart disease in the next ten years.

In fact, if $100p_1\%$ of individuals belong to a subpopulation described by the $N(\mu_1, \sigma_1^2)$ distribution and $100p_2\%$ of individuals belong to a subpopulation described by the $N(\mu_2, \sigma_2^2)$ distribution, then the probability that an observation x originated from the $N(\mu_1, \sigma_1^2)$ distribution is

$$\frac{p_1 \times \sigma_1^{-1} \exp\left[-\frac{(x-\mu_1)^2}{(2\times\sigma_1^2)}\right]}{p_1 \times \sigma_1^{-1} \exp\left[-\frac{(x-\mu_1)^2}{(2\times\sigma_1^2)}\right] + p_2 \times \sigma_2^{-1} \exp\left[-\frac{(x-\mu_2)^2}{(2\times\sigma_2^2)}\right]}. \tag{2}$$

[[This, too, you will prove in Written Assignment 4.]] In expression (1) we had $p_1 = 0.10$, $p_2 = 0.90$, $\sigma_1 = 15$, $\mu_1 = 150$, $\sigma_2 = 10$, $\mu_2 = 130$, and $x = 139$.

The general problem. The motivating problem required some effort, but the general problem is even more challenging.

Suppose that we have a categorical response variable Y. For convenience, let the categories be labeled $1, 2, \ldots, m$. Suppose that we have multiple explanatory variables $X_1, \ldots, X_k$. We want to predict a person's Y value (i.e., into which response category he falls) given his values on $X_1, \ldots, X_k$.

Discussion question. There are multiple reasons that the general problem is more challenging than the motivating problem. What are they?

A key assumption for the general problem. Let us assume that $X_1, \ldots, X_k$ follow a <u>multivariate normal distribution</u> among people with a given Y value (i.e., among people in a given response category). This implies that X_1 individually is normally distributed among people with a given Y value, that X_2 individually is normally distributed among people with a given Y value, and so forth. [[Actually, much more is implied. For any constants $c_1, \ldots, c_k$, the linear combination $c_1X_1 + \cdots + c_kX_k$ must be normally distributed among people with a given Y value.]]

Discussion question. How does the present scenario compare with that for binary logistic regression?

Strategy for addressing the general problem. We address the general problem by mimicking the approach used to resolve the motivating problem. That is, we come up with an analogue to expression (2) that applies when there are m (possibly more than 2) response categories and when there are k (possibly more than 1) explanatory variables.

Let μ_1 be a vector (i.e., a column of numbers) containing the means of $X_1, \ldots, X_k$ among individuals with $Y = 1$. Define μ_2 through μ_m analogously. Let Σ_1 be a matrix (i.e., an array of numbers) containing the variances and covariances of $X_1, \ldots, X_k$ among individuals with $Y = 1$. Define Σ_2 through Σ_m analogously. Let p_1 be the proportion of individuals with $Y = 1$. Define p_2 through p_m analogously.

Given realized values $x_1, \ldots, x_k$ for the explanatory variables $X_1, \ldots, X_k$, the probability that $Y = 1$ equals

$$\frac{p_1 \times \left(\sqrt{Det(\Sigma_1)}\right)^{-1} \exp[-(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)/2]}{\Sigma_{j=1}^{m} p_j \times \left(\sqrt{Det(\Sigma_j)}\right)^{-1} \exp[-(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)/2]}, \tag{3}$$

where x is a vector containing $x_1, \ldots, x_k$, Det denotes the determinant of a matrix, and $'$ denotes the transpose of a vector. Similarly, the probability that $Y = 2$ equals

$$\frac{p_2 \times \left(\sqrt{Det(\Sigma_2)}\right)^{-1} \exp[-(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)/2]}{\Sigma_{j=1}^{m} p_j \times \left(\sqrt{Det(\Sigma_j)}\right)^{-1} \exp[-(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)/2]}, \tag{4}$$

and if we wanted we could write out analogous expressions for the probabilities that $Y = 3$, $Y = 4$, and so forth.

Expressions (3) and (4) are more complicated than expression (2). Nonetheless, their basic structure reveals that what we are doing now is just a multidimensional version of what we did to resolve the motivating problem.

Of course, in practice we do not know $p_1, \ldots, p_m$, $\mu_1, \ldots, \mu_m$, or $\Sigma_1, \ldots, \Sigma_m$. Thus, we estimate these quantities using a training data set. Once we have done so, we are free to use expressions (3) and (4) – with the unknown quantities replaced by their estimates – to calculate the probability that an individual belongs to the first response category, the probability that an individual belongs to the second response category, and so forth.

Remarks. Some statisticians refer to $p_1, \ldots, p_m$ as <u>prior probabilities</u> because $p_1, \ldots, p_m$ can be interpreted as the probabilities of an individual belonging to the various response categories <u>before</u> we have any information about that individual's values on $X_1, \ldots, X_k$. These statisticians also refer to expressions (3) and (4) – as well as analogous expressions for $Y = 3$, $Y = 4$, and so forth – as <u>posterior probabilities</u>.

Linear and quadratic discriminants. Since the posterior probability that $Y = 1$ is proportional to

$$p_1 \times \left(\sqrt{Det(\Sigma_1)}\right)^{-1} \exp[-(x - \mu_1)'\Sigma_1^{-1}(x - \mu_1)/2],$$

the posterior probability that $Y = 2$ is proportional to

$$p_2 \times \left(\sqrt{Det(\Sigma_2)}\right)^{-1} \exp[-(x - \mu_2)'\Sigma_2^{-1}(x - \mu_2)/2],$$

and so forth, our prediction entails identifying the j among $1, \ldots, m$ for which

$$p_j \times \left(\sqrt{Det(\Sigma_j)}\right)^{-1} \exp[-(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)/2]$$

is largest. This is equivalent to identifying the j for which

$$D_j := \log[p_j] - \log[Det(\Sigma_j)]/2 - (x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)/2 \tag{5}$$

is largest. We refer to D_j as the discriminant function for the j^{th} response category. That is why making predictions in this manner is called discriminant analysis.

If $\Sigma_1 = \cdots = \Sigma_m$, then the discriminant functions are called linear. Otherwise, the discriminant functions are called quadratic. To understand this, suppose that $k = 2$. We can carve the x_1x_2-plane into regions where we predict that $Y = 1$, that $Y = 2$, and so forth. If the discriminant functions are linear, then the boundaries of these regions turn out to be straight lines. Otherwise, the boundaries turn out to be curves. The situation is not so easily visualized when $k > 2$. Nevertheless, linear discriminant functions yield boundaries that are governed by linear equations, while quadratic discriminant functions yield boundaries that are governed by quadratic equations. [[See Section 4.3 of Hastie et al for mathematical details.]]

Understanding the output from the DISCRIM macro

Illustrative example. We use the diabetes data sets supplied by Fernandez in {diabet1.sas7bdat} and {diabet2.sas7bdat}. A description is found on page 288 of the textbook. The response variable Y has $m = 3$ categories: normal (1), overt diabetic (2), and chemical diabetic (3). The explanatory variables X_1, X_2, X_3, X_4, X_5 represent relative weight, fasting plasma glucose level, test plasma glucose, plasma insulin during test, and steady-state plasma glucose level. We regard {diabet1.sas7bdat} as a training data set and {diabet2.sas7bdat} as a test data set, even though Fernandez refers to the latter as a validation data set.

Exploration. Discriminant analysis can be based on all available explanatory variables (in this example, X_1, X_2, X_3, X_4, X_5) or just some of them. An exploration of the training data set is useful for determining whether all available explanatory variables should be used and, if not, which ones should be removed. The output in {GFLIB.Diabet181.doc} describes such an exploration.

Page 1 displays scatter plots of X_1 against X_2, of X_1 against X_3, and so forth. The scatter plots use the symbols $1, 2, 3$ to identify the response category to which each person in the training data set belongs. We see, for instance, that X_2 and X_3 separate group 3 cleanly from groups 1 and 2. Thus, our initial reaction is that X_2 and X_3 may be useful in making predictions about Y.

Pages 2 through 6 show the results of a backward elimination procedure for discriminant analysis; this is conceptually similar to the backward elimination procedures for linear regression and logistic regression. [[See pp. 8-11 of {http://web.as.uky.edu/statistics/users/rjchar2/CPH930F06/L2930F06.pdf} and p. 6 of {http://.../CPH930F06/L7930F06.pdf}.]] Before seeing how the backward elimination procedure plays out in this example, we note on page 2 that 72 (51.1%) of the 141 people in the training data set have $Y = 1$, 36 (25.5%) have $Y = 2$, and 33 (23.4%) have $Y = 3$.

We ignore the boxes labeled "Multivariate Statistics", but the boxes labeled "Statistics for Removal" are important. We begin with all available explanatory variables. Then we remove variables with high p-values, one at a time, until all remaining variables have low p-values. On page 4 we see that X_5 has the largest p-value in Step 1. [[This p-value is computed from an auxiliary "analysis of covariance" model with X_5 as the response and X_1, X_2, X_3, X_4, Y as explanatory variables.]] Since this p-value (0.2820) exceeds Fernandez's cutoff of 0.15, X_5 is removed. Continuing, we see on page 5 that X_4 has the largest p-value among the remaining explanatory variables in Step 2. [[This p-value is computed from an auxiliary model with X_4 as the response and X_1, X_2, X_3, Y as explanatory variables.]] Yet, since this p-value (0.0156) is less than 0.15, X_4 is retained. The backward elimination procedure terminates with the suggestion that X_1, X_2, X_3, X_4 be used in the discriminant analysis.

Pages 7 through 13 show the results of a stepwise selection procedure for discriminant analysis, while pages 14 through 20 show the results of a forward selection procedure. For this example, all three procedures suggest that X_1, X_2, X_3, X_4 be used in the discriminant analysis.

In general, the three procedures do not always make the same suggestion. Even when they do make the same suggestion, we may want to consider other possibilities. For instance, another possibility is to use all available explanatory variables. As with linear regression and logistic regression, we can make a final choice about explanatory variables based either on validation data set performance or on a selection criterion derived from training data set performance.

Discriminant analysis. For now we proceed with discriminant analysis using only X_1, X_2, X_3, X_4. The results are presented in {GFLIB.Diabet183.doc} and are described below.

Assessing multivariate normality. Pages 1 through 17 all pertain to the assessment of multivariate normality, but attention may be confined to pages 4, 10, and 16. Page 4 presents a quantile-quantile plot for group 1, page 10 presents a quantile-quantile plot for group 2, and page 16 presents a quantile-quantile plot for group 3. These quantile-quantile plots may be regarded as analogues to normal probability plots that accommodate multiple variables simultaneously.

In each quantile-quantile plot, a configuration resembling a straight line suggests that X_1, X_2, X_3, X_4 have an approximate multivariate normal distribution within the corresponding group. Of course, one's interpretation of "resembling a straight line" is subjective. I favor a liberal interpretation, more liberal than for a normal probability plot. [[The reason is that the axes for a quantile-quantile plot are on a different scale than those for a normal probability plot. This different scale tends to highlight minor departures from multivariate normality, more than minor departures from univariate normality are highlighted in a normal probability plot.]]

Each plot is accompanied by p-values for multivariate analogues to skewness and kurtosis. Low p-values suggest departures from multivariate normality within the corresponding group.

Data summarization. Page 18 identifies the numbers of individuals in the training data set belonging to each response category. The fractions of such individuals are the estimates of p_1, p_2, and p_3.

Page 19 presents descriptive statistics on X_1, X_2, X_3, X_4 overall and within each response category.

Linear or quadratic discriminants. Page 21 tests $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3$ against the complementary alternative hypothesis. As the p-value is (far) less than Fernandez's cutoff of 0.10, the null hypothesis is rejected. Thus, we will not assume that $\Sigma_1 = \Sigma_2 = \Sigma_3$ in our discriminant analysis. In other words, we will use quadratic discriminants instead of linear discriminants.

Canonical variables. Let Z_1, Z_2, Z_3, Z_4 denote standardized versions of X_1, X_2, X_3, X_4. Page 26 identifies

$$C_1 := 0.2516Z_1 - 0.4104Z_2 + 3.2274Z_3 + 0.0426Z_4$$

as an optimal linear combination of Z_1, Z_2, Z_3, Z_4 in the sense of separating the groups. Page 26 also identifies

$$C_2 := 0.5628Z_1 - 2.5000Z_2 + 2.4103Z_3 + 0.4699Z_4$$

as an optimal linear combination in the sense of separating the groups subject to the constraint that C_2 be uncorrelated with C_1. We refer to C_1 and C_2 as canonical variables. In general, the number of canonical variables is the smaller of $(m - 1)$ and k.

Additional output, to be considered later, will clarify how C_1 and C_2 are separating the groups.

Discriminant analysis results for the training data. Pages 27 and 28 present results for the training data set. These results rely on cross validation, a statistical technique that compensates for the optimism inherent to assessments made with the training data. Thus, if we don't have a validation data set, we can use these results to decide which variables to include in the discriminant analysis. [[The idea behind cross validation is to compare the actual response of subject i in the training data set to what would be predicted for subject i if his/her information had not been used to estimate parameters. In fact, we have seen cross validation earlier in the semester, without explicitly labeling it as such. Do you remember where?]]

Page 27 identifies five individuals in the training data set who were incorrectly classified by quadratic discriminants. For instance, subject #5 was really normal but was classified as overt diabetic because he had an estimated posterior probability of 0.8731 for being overt diabetic compared to an estimated posterior probability of 0.1242 for being normal.

Page 28 summarizes how the people in each group were classified. We had an error rate of $3/72 = 0.0417$ in group 1, an error rate of $1/36 = 0.0278$ in group 2, and an error rate of $1/33 = 0.0303$ in group 3. To obtain a single number summary, we take a weighted average of these error rates. The weights are the estimates of the prior probabilities p_1, p_2, and p_3 (i.e., the observed fractions of people in group 1, group 2, and group 3 in the training data set). With these weights, the single number summary – called the prior-weighted error rate – is one minus the correct classification rate. In this example, the prior-weighted error rate is

$$0.0417 \times 0.5106 + 0.0278 \times 0.2553 + 0.0303 \times 0.2340 = 0.0355.$$

If we were to repeat the discriminant analysis using all available explanatory variables, then we would compare the new prior-weighted error rate to 0.0355. If the new prior-weighted error rate were less than 0.0355, then using all available explanatory variables would be deemed preferable to using only X_1, X_2, X_3, X_4.

Discriminant analysis results for the test data. The discriminant functions can be applied not just to classify individuals in the training data set but also to classify individuals in any other data set with the same variables. If this is the validation data set, then the purpose is to decide which variables should be included in the discriminant analysis. If this is the test data set, then the purpose is to formally evaluate the predictive capabilities of the discriminant functions.

In our example, there is no validation data set. However, there is a test data set. Pages 30 through 35 list the predictions for each and every person in the test data set. Estimated posterior probabilities are also included.

Page 36 reveals a prior-weighted error rate of 0.0422 for the test data set.

Page 37 explicitly identifies the six individuals in the test data set who were incorrectly classified.

Some graphical displays and related output on canonical variables. Pages 38 and 39 show box plots of C_1 and C_2 by response category in the training data set. We see that people in group 3 tend to have high values of C_1, people in group 2 tend to have medium values of C_1, and people in group 1 tend to have low values of C_1.

Pages 40 and 41 identify subjects in the training data set with the lowest and highest values of C_1.

Page 42 shows a scatterplot of C_2 against C_1 in the training data set. The plotting symbols identify group membership. This scatterplot and the aforementioned box plots clarify how C_1 and C_2 separate the groups.