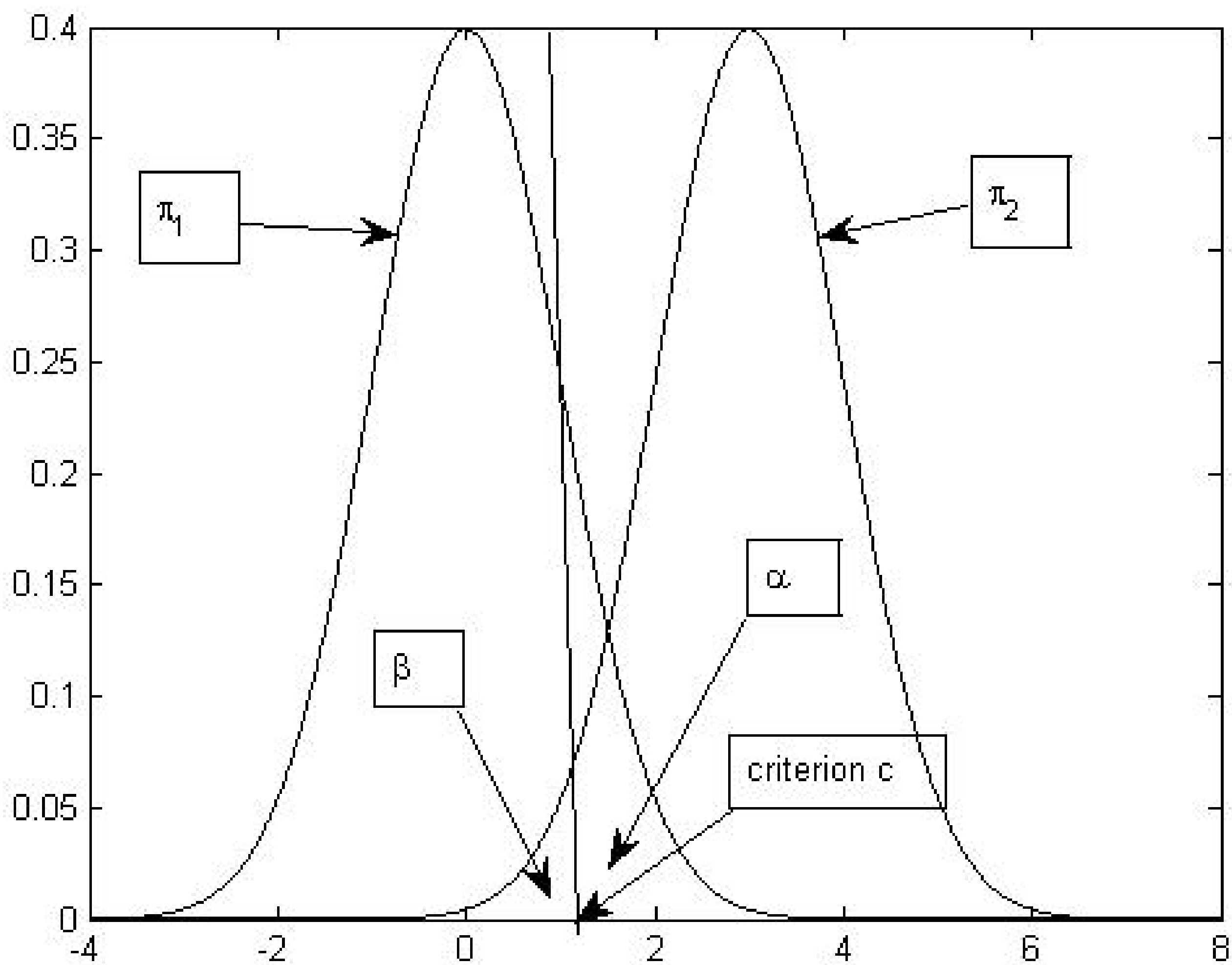
Version: September 30, 2010

Notes on Discrimination and Classification

The term "discrimination" (in a nonpejorative statistical sense) refers to the task of discrimination among groups through linear combinations of variables that maximize some criterion, usually F-ratios. The term "classification" refers to the task of allocating observations to existing groups, typically to minimize the cost and/or probability of misclassification. These two topics are intertwined, but it is most convenient to start with the topic of classification.

In the picture to follow, we have two populations, called π_1 and π_2; π_1 is characterized by a normal distribution with mean μ_1, and variance σ_X^2 (the density is denoted by $f_1(x)$); π_2 is characterized by a normal distribution with mean μ_2, and (common) variance σ_X^2 (the density is denoted by $f_2(x)$). I have an observation, say x_0, and wish to decide where it should go, either to π_1 or π_2. Assuming implicitly that $\mu_1 \leq \mu_2$, we choose a criterion point, c, and allocate to π_1 if $x_0 \leq c$, and to π_2 if $> c$. The probabilities of misclassification can be given in the following chart (and in the figure):

		True State	
		π_1	π_2
Decision	π_1	$1 - \alpha$	β
	π_2	α	$1 - \beta$



0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-2
0
2
4
6
8
π_1
π_2
α
β
criterion c

If I want to choose c so that $\alpha + \beta$ is smallest, I would select the point at which the densities are equal. A more complicated way of saying this decision rule is to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq 1$; if < 1, then allocate to π_2. Suppose now that the prior probabilities of being drawn from π_1 and π_2 are p_1 and p_2, where $p_1 + p_2 = 1$. I wish to choose c so the Total Probability of Misclassification (TPM) is minimized, i.e., $p_1\alpha + p_2\beta$. The rule would be to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq p_2/p_1$; if $< p_2/p_1$, then allocate to π_2. Finally, if we include costs of misclassification, $c(1|2)$ (for assigning to π_1 when actually coming from π_2), and $c(2|1)$ (for assigning to π_2 when actually coming from π_1), we can choose c to minimize the Expected Cost of Misclassification (ECM), $c(2|1)p_1\alpha + c(1|2)p_1\beta$, with the associated rule of allocating to π_1 if $f_1(x_0)/f_2(x_0) \geq (c(1|2)/c(2|1))(p_2/p_1)$; if $< (c(1|2)/c(2|1))(p_2/p_1)$, then allocate to π_2.

Using logs, the last rule can be restated: allocate to π_1 if $\log(f_1(x_0)/f_2(x_0)) \geq \log((c(1|2)/c(2|1))(p_2/p_1))$. The left-hand-side is equal to $(\mu_1 - \mu_2)(\sigma_X^2)^{-1}x_0 - (1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2)$, so the rule can be restated further: allocate to π_1 if

$$x_0 \leq \{(1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2)$$
$$- \log((c(1|2)/c(2|1))(p_2/p_1))\}\{\frac{\sigma_X^2}{-(\mu_1 - \mu_2)}\}$$

or

$$x_0 \leq \{(1/2)(\mu_1+\mu_2) - \log((c(1|2)/c(2|1))(p_2/p_1))\}\{\frac{\sigma_X^2}{(\mu_2 - \mu_1)}\} = c\,.$$

If the costs of misclassification are equal (i.e., $c(1|2) = c(2|1)$), then the allocation rule is based on classification functions: allocate to π_1 if

$$[\frac{\mu_1}{\sigma_X^2}x_0 - (1/2)\frac{\mu_1^2}{\sigma_X^2} + \log(p_1)] - [\frac{\mu_2}{\sigma_X^2}x_0 - (1/2)\frac{\mu_2^2}{\sigma_X^2} + \log(p_2)] \geq 0 \ .$$

Moving toward the multivariate framework, suppose population π_1 is characterized by a $p \times 1$ vector of random variables, $\mathbf{X} \sim \text{MVN}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$; population π_2 is characterized by a $p \times 1$ vector of random variables, $\mathbf{X} \sim \text{MVN}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$. We have a similar allocation rule as in the univariate case: allocate to π_1 if $\mathbf{a}\mathbf{x}_0 - \mathbf{a}[(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})/2] \geq (c(1|2)/c(2|1))(p_2/p_1)$, where

$$\mathbf{a} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1} \ .$$

Or, if the misclassification costs are equal, allocate to π_1 if $\mathbf{a}\mathbf{x}_0 - \mathbf{a}[(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})/2] \geq [\log(p_2) - \log(p_1)]$. In effect, we define regions of classification, say R_1 and R_2; if an observation falls into region R_i, it is allocated to group i, for $i = 1, 2$ There are a number of ways of restating this last rule (assuming equal misclassification costs, this is choosing to minimize the Total Probability of Misclassification (TPM)):

A) Evaluate the classification functions for both groups and assign according to which is higher: allocate to π_1 if

$$[\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - (1/2)\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) + \log(p_1)] -$$
$$[\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - (1/2)\boldsymbol{\mu}_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \log(p_2)] \geq 0 \ .$$

B) Define the posterior probability of being in group i, for $i = 1, 2$, $P(\pi_i|\mathbf{x}_0)$ as $(f_i p_i)/(f_1 p_1 + f_2 p_2)$. We allocate to the group with the largest posterior probability.

C) We can restate our allocation rule according to Mahalanobis distances: define the squared Mahalanobis distance of $\mathbf{x}_0$ to $\mu_i, i = 1, 2$, as

$$(\mathbf{x}_0 - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_i) \ .$$

Allocate to π_i for the largest quantity of the form:

$$-(1/2)[(\mathbf{x}_0 - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_i)] + \log(p_i) \ .$$

When the covariance matrices are not equal in the two populations (i.e., $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$), the allocation rules get a little more complicated. The classification rules are now called "quadratic", and may produce regions of allocation that may not be contiguous. This is a little strange, but it can be done, and we can still split the allocation rule into two classification functions (assuming, as usual, equal costs of misclassification):

Assign to π_1 if

$$-(1/2)\mathbf{x}_0'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x}_0 + (\boldsymbol{\mu_1}'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu_2}'\boldsymbol{\Sigma}_1^{-1})\mathbf{x}_0 \ - \ k \geq$$

$$\log((c(1|2)/c(2|1))(p_2/p_1)) \ ,$$

where

$$k = (1/2)\log(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}) + (1/2)(\boldsymbol{\mu_1}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu_2}) \ .$$

Moving to the sample, we could just use estimated quantities and hope our rule does well — we use $\mathbf{S}_{pooled}$, assuming equal covariance matrices in the two populations, and sample means, $\hat{\boldsymbol{\mu}}_\mathbf{1}$ and $\hat{\boldsymbol{\mu}}_\mathbf{2}$. In fact, we can come up with the misclassification table based on the given sample and how they allocate the given n observations to the two groups:

		Group	
		π_1	π_2
Decision	π_1	a	b
	π_2	c	d
		n_1	n_2

The apparent error rate (APR) is $(b+c)/n$, which is overly optimistic because it is optimized with respect to *this* sample. To cross-validate, we could use a "hold out one-at-a-time" strategy (i.e., a sample reuse procedure commonly referred to as the "jackknife"):

		Group	
		π_1	π_2
Decision	π_1	a^*	b^*
	π_2	c^*	d^*
		n_1	n_2

To estimate the actual error rate (AER), we would use $(b^* + c^*)/n$.

Suppose we have g groups; p_i is the a priori probability of group i, $1 \leq i \leq g$; $c(k|i)$ is the cost of classifying an i as a k. The decision rule that minimizes the expected cost of misclassification (ECM) is: allocate $\mathbf{x}_0$ to population π_k, $1 \leq k \leq g$, if

$$\sum_{i=1; i \neq k}^{g} p_i f_i(\mathbf{x}_0) c(k|i)$$

is smallest.

There are, again, alternative ways of stating this allocation rule; we will assume for convenience that the costs of misclassification are equal:

Allocate to group k if the posterior probability,

$$P(\pi_k|\mathbf{x}_0) = \frac{p_k f_k(\mathbf{x}_0)}{\Sigma_{i=1}^{g} p_i f_i(\mathbf{x}_0)} \ ,$$

is largest.

If in population k, $\mathbf{X} \sim \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, we allocate to group k if $\log(p_k f_k(\mathbf{x}_0)) =$

$$-(1/2)\log(|\boldsymbol{\Sigma}_k|) - (1/2)(\mathbf{x}_0 - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_k) + \log(p_i) + \text{constant} \ ,$$

is largest.

If all the $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k, then we allocate to π_k if

$$\boldsymbol{\mu}_k'\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_0 - (1/2)\boldsymbol{\mu}_k'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \log(p_k) \ ,$$

is largest.

It is interesting that we can do this in a pairwise way as well: allocate to π_k if

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_0 - (1/2)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) \geq \log(p_i/p_k) \ ,$$

for all $i = 1, \ldots, g$.

0.0.1 Discriminant Analysis

Suppose we have a one-way analysis-of-variance (ANOVA) layout with J groups (n_j subjects in group j, $1 \leq j \leq J$), and p measurements on each subject. If x_{ijk} denotes person i, in group j, and the observation of variable k ($1 \leq i \leq n_j$; $1 \leq j \leq J$; $1 \leq k \leq p$), then define the Between-Sum-of-Squares matrix

$$\mathbf{B}_{p \times p} = \{\sum_{j=1}^{J} n_j(\bar{x}_{\cdot jk} - \bar{x}_{\cdot\cdot k})(\bar{x}_{\cdot jk'} - \bar{x}_{\cdot\cdot k'})\}_{p \times p}$$

and the Within-Sum-of-Squares matrix

$$\mathbf{W}_{p \times p} = \{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ijk} - \bar{x}_{\cdot jk})(x_{ijk'} - \bar{x}_{\cdot jk'})\}_{p \times p}$$

For the matrix product $\mathbf{W}^{-1}\mathbf{B}$, let $\lambda_1, \ldots, \lambda_T \geq 0$ be the eigenvectors ($T = \min(p, J-1)$, and $\mathbf{p}_1, \ldots, \mathbf{p}_T$ the corresponding normalized eigenvectors. Then, the linear combination

$$\mathbf{p}'_k \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

is called the k^{th} *discriminant function*. It has the valuable property of maximizing the univariate F-ratio subject to being uncorrelated with the earlier linear combinations.

There are a number of points to make about (Fisher's) Linear Discriminant Functions:

A) Typically, we define a sample pooled variance-covariance matrix, $\mathbf{S}_{pooled}$, as $(\frac{1}{n-J})\mathbf{W}$. And generally, the eigenvalues are scaled so that $\mathbf{p}'_k \mathbf{S}_{pooled} \mathbf{p}_k = 1$.

B) When $J = 2$, the eigenvector, $\mathbf{p}_1'$, is equal to $(\hat{\boldsymbol{\mu}}_{\mathbf{1}} - \hat{\boldsymbol{\mu}}_{\mathbf{2}})'\mathbf{S}_{pooled}$. This set of weights maximized the square of the t ratio in a two-group separation problem (i.e., discriminating between the two groups). We also maximize the square of the effect size for this linear combination; the maximum for such an effect size is

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' ,$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the sample centroids in groups 1 and 2 for the p variables. Finally, if we define $Y = 1$ if an observation falls into group 1, and $= 0$ if in group 2, the set of weights in $\mathbf{p}_1'$ is proportional to the regression coefficient in predicting Y from $X_1, \ldots, X_p$.

C) The classification rule based on Mahalanobis distance (and assuming equal prior probabilities and equal misclassification values), could be restated equivalently using plain Euclidean distances in discriminate function space.