Discrimination and Classification

• Goals:

**Discrimination:** finding the features that *separate* known groups in a multivariate sample.

**Classification:** developing a rule to *allocate* a new object into one of a number of known groups.

• A classification rule is based on the features that separate the groups, so the goals overlap.
Simplest Case: Two Groups

- Both discrimination and classification depend on multivariate observation $X$.

- Basic setup:
  - Consider each group a population: $\pi_1$ and $\pi_2$.
  - The density of $X$ in $\pi_g$ is $f_g(x)$, $g = 1, 2$. 
• Alternatively:

  – A single population with a group indicator $G$.
  – The density of $X$ given $G = g$ is $f_g(x)$.
  – $P(G = 1) = p_1$, $P(G = 2) = p_2 = 1 - p_1$.

• $p_1$ and $p_2$ are the prior probabilities of the groups.
• Example: ownership of a riding mower.
• Classification problem is to predict whether a homeowner owns a riding mower, given the Income and Lot Size.

• Rule must give an answer for any $x$, so it defines two regions $R_1$ and $R_2$.

• Classification errors:

\[ P(2|1) = P(X \in R_2|G = 1) = \int_{R_2} f_1(x)dx, \]

\[ P(1|2) = P(X \in R_1|G = 2) = \int_{R_1} f_2(x)dx. \]
• Unconditionally:

\[
P \text{ (correctly classified as } \pi_1) = P(1|1)p_1
\]

\[
P \text{ (missclassified as } \pi_1) = P(1|2)p_2
\]

\[
P \text{ (correctly classified as } \pi_2) = P(2|2)p_2
\]

\[
P \text{ (missclassified as } \pi_2) = P(2|1)p_1
\]

• The two types of classification error may have different costs:

\[
\begin{array}{ccc}
\text{Classify as:} & \pi_1 & \pi_2 \\
\hline
\text{True population:} & \pi_1 & 0 & c(2|1) \\
& \pi_2 & c(1|2) & 0
\end{array}
\]
• **Expected Cost of Misclassification:**

\[
ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2.
\]

• The classification region that minimizes ECM is

\[
R_1 : \frac{f_1(x)}{f_2(x)} \geq \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right).
\]

• Note: these regions depend on the ratios of:
  
  – the densities;
  
  – the costs of misclassification;
  
  – the prior probabilities.
• Special cases:

  – *Total Probability of Misclassification*:

    \[ \text{TPM} = P(\text{misclassify either way}) = P(2|1)p_1 + P(1|2)p_2 \]

    is ECM with \(c(1|2) = c(2|1) = 1\).

  – Posterior mode: allocate to the population with the higher posterior probability; by Bayes’s rule:

    \[
    P(\pi_1|x_0) = \frac{p_1 f_1(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)} \\
    P(\pi_2|x_0) = \frac{p_2 f_2(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}
    \]

    \(\Rightarrow\) same rule as minimizing TPM.
Multivariate Normal Populations

- If $f_1(x)$ and $f_2(x)$ are multivariate normal with the same $\Sigma$ and means $\mu_1$ and $\mu_2$, then the minimum-ECM allocation rule is based on a linear function of $x$.

- Basic result is

$$\ln \left[ \frac{f_1(x)}{f_2(x)} \right] = (\mu_1 - \mu_2)' \Sigma^{-1} \left[ x - \frac{1}{2} (\mu_1 + \mu_2) \right]$$

- Optimal region is

$$R_1 : \ (\mu_1 - \mu_2)' \Sigma^{-1} \left[ x - \frac{1}{2} (\mu_1 + \mu_2) \right] \geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right].$$
• Typically, $\Sigma$, $\mu_1$ and $\mu_2$ are unknown, but replaced with sample estimates, based on training data.

• Example: Hemophilia A carriers.
  
  – Women subjects in two groups: Non-carriers and Obligatory carriers.
  
  – Variables:
    
    \[
    X_1 = \log_{10}(\text{AHF activity}) \\
    X_2 = \log_{10}(\text{AHF-like antigen})
    \]

• SAS proc discrim program and output.
Evaluating Classification Functions

• The Optimum Error Rate is the minimum TPM.

• For multivariate normal populations,

\[ \text{OER} = 1 - \Phi \left( \frac{\Delta}{2} \right), \]

where \( \Delta^2 \) is the Generalized Squared Distance between groups:

\[ \Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2). \]

• For the hemophilia example, SAS reports \( \Delta^2 = 4.57431 \), whence \( \text{OER} = 0.1424 \).
• The OER calculation is the *estimated* error rate for allocation using the *true* parameters—good in large samples, but dubious in small samples.

• Sample-based evaluation:

  – *Apparent Error Rate* = fraction of training data that are misclassified; reported by SAS as “Resubstitution Summary”; APER = 0.1389 for the example; under-estimates error rate, because the data used to develop the rule are then used to evaluate it;

  – Lachenbruch’s holdout procedure: omit one observation at a time, recalculate the classification rule, and classify the omitted observation; reported by SAS as “Cross-validation Summary”, 0.1556 for the example.
Quadratic Classification

- If $f_1(x)$ and $f_2(x)$ are multivariate normal with different covariances $\Sigma_1$ and $\Sigma_2$ and means $\mu_1$ and $\mu_2$, then the minimum-ECM allocation rule is based on a quadratic function of $x$.

- Classification regions are now bounded by conic sections (generally elliptical or hyperbolic).

- No simple explicit representation.
• In `proc discrim`, use `pool = no` on the `proc` statement.

• For the AHF example, the resubstitution Total error rate is 0.1389 and the cross-validated Total error rate is 0.1722.

• The cross-validated Total error rate is higher than for linear classification, suggesting that the estimation cost of the additional parameters in the two $\Sigma$’s outweighs their value in improving the fit.
Nonparametric Classification

- One alternative to the multivariate normal assumption is classification based on nonparametric density estimates of $f_g(x)$.

- In SAS, use `method = npar` on the `proc discrim` statement. You have the choice of nearest neighbor and kernel estimation:
  - for nearest neighbor estimation, use `k = ...` to specify the number of neighbors;
  - for kernel estimation, use `kernel = ...` to name a kernel, and `r = ...` to specify the radius (bandwidth).