Math 70 Homework 7

Alex Craig

Problem 1

Instructions

A bank wants to develop a rule for granting or denying a credit card application. Over the years, the bank collected the data on 1996 applicants - some of them failed a minimum payment and some did not, see the data `creditpr.csv`. The first column indicates `failed = 1`, not `failed = 0`; the second column contains monthly paycheck; and the third column contains months at work.

(a) Develop and formulate the rule to grant credit card applications in the form $M < a + bP$, where M is months at work and P is paycheck.

(b) Plot the data points using red for those who failed the minimum payment and green otherwise (use `legend`) and display the discrimination line.

(c) Compute and display the total empirical and theoretical misclassification probability.

(d) Compute the area under the respective binormal ROC curve and compare it with the total misclassification error from LDA. Hint: adopt the R function `mah`.'

Solution

(a) If we assume that the populations of the two groups are normally distributed with equal covariance matrices, then the optimal rule is given by the linear discriminant rule. Let $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Omega})$ be the population for credit card payers, and $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Omega})$ be the random variable for credit card non-payers.

The linear discrimination rule is given by by finding a plane $(\mathbf{z} - \mathbf{s})^T \mathbf{n} = 0$ where $\mathbf{s}$ is called the translation vector and $\mathbf{n}$ is called the normal vector. A given point $\mathbf{z}$ is classified as $\mathbf{x}$ if $(\mathbf{z} - \mathbf{s})^T \mathbf{n} > 0$. The translation and normal vectors are given by:

$$\mathbf{s} = \frac{1}{2}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y), \quad \mathbf{n} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)$$

Now, we may computer our linear discrimination rule. Our populations of credit card payers and non payers will both have mean values of months at work (M), and paycheck (P). We may then define their means as follows:

$$\boldsymbol{\mu}_x = \begin{bmatrix} E[M_x] \\ E[P_x] \end{bmatrix}, \quad \boldsymbol{\mu}_y = \begin{bmatrix} E[M_y] \\ E[P_y] \end{bmatrix}$$

We may also define the populations' covariance matrices as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} cov(M_x, M_x) & cov(M_x, P_x) \\ cov(P_x, M_x) & cov(P_x, P_x) \end{bmatrix} = \begin{bmatrix} cov(M_y, M_y) & cov(M_y, P_y) \\ cov(P_y, M_y) & cov(P_y, P_y) \end{bmatrix} = \begin{bmatrix} \sigma_M^2 & \sigma_{MP} \\ \sigma_{PM} & \sigma_P^2 \end{bmatrix}$$

For simplicity, we will assume that our covariance matrix $\boldsymbol{\Omega}$ is diagonal, i.e., $cov(M, P) = cov(P, M) = 0$, and that variances of M and P are equal, $\sigma_M^2 = \sigma_P^2 = \sigma^2$. This simplifies $\boldsymbol{\Omega}$ to:

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\mathbf{\Omega}^{-1}$ is then given by:

$$\mathbf{\Omega}^{-1} = \frac{1}{\sigma^2}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

Substituting these values into our formulas for $\mathbf{s}$ and $\mathbf{n}$, we get:

$$\mathbf{s} = \frac{1}{2}\begin{bmatrix}M_x + M_y\\ P_x + P_y\end{bmatrix}, \quad \mathbf{n} = \frac{1}{\sigma^2}\begin{bmatrix}M_x - M_y\\ P_x - P_y\end{bmatrix}$$

So, our linear discriminant rule for classifying an applicant $\mathbf{z}$ with months M_z and paycheck P_z as group $\mathbf{x}$ (credit card payer) is:

$$\begin{bmatrix}M_z - \frac{1}{2}(E[M_x] + E[M_y])\\ P_z - \frac{1}{2}(E[P_x] + E[P_y])\end{bmatrix}^T \frac{1}{\sigma^2}\begin{bmatrix}E[M_x] - E[M_y]\\ E[P_x] - E[P_y]\end{bmatrix} > 0$$

This simplifies to:

$$\frac{1}{\sigma^2}[(M_z - \frac{1}{2}(E[M_x] + E[M_y]))(E[M_x] - E[M_y]) + (P_z - \frac{1}{2}(E[P_x] + E[P_y]))(E[P_x] - E[P_y])] > 0$$

Because variance is always positive, $\frac{1}{\sigma^2}$ is positive and we can thus simplify and expand to:

$$M_z E[M_x] - M_z E[M_y] - \frac{E[M_x]}{2}(E[M_x] + E[M_y]) + \frac{E[M_y]}{2}(E[M_x] + E[M_y]) + \dots$$

$$\dots P_z E[P_x] - P_z E[P_y] - \frac{E[P_x]}{2}(E[P_x] + E[P_y]) + \frac{E[P_y]}{2}(E[P_x] + E[P_y]) > 0$$

Define α as:

$$\alpha = -\frac{E[M_x]}{2}(E[M_x] + E[M_y]) + \frac{E[M_y]}{2}(E[M_x] + E[M_y]) - \frac{E[P_x]}{2}(E[P_x] + E[P_y]) + \frac{E[P_y]}{2}(E[P_x] + E[P_y])$$

Our linear discriminant rule is then:

$$M_z(E[M_x] - E[M_y]) + P_z(E[P_x] - E[P_y]) + \alpha > 0$$

$$-M_z(E[M_x] - E[M_y]) - P_z(E[P_x] - E[P_y]) - \alpha < 0$$

$$\Rightarrow M_z(E[M_x] - E[M_y]) > P_z(E[P_x] - E[P_y]) + \alpha$$

$$\Rightarrow M_z < P_z\frac{E[P_x] - E[P_y]}{E[M_y] - E[M_x]} + \frac{\alpha}{E[M_y] - E[M_x]}$$

Let $b = \frac{E[P_x] - E[P_y]}{E[M_y] - E[M_x]}$ and $a = \frac{\alpha}{E[M_y] - E[M_x]}$. We then have our linear discriminant rule:

$$M_z < a + bP_z$$

Problem 2

Instructions

Logistic regression has two predictors x_1 and x_2 Develop algorithms for testing the null hypothesis that the two slope coefficients are the same via the Wald and likelihood ratio test. Formulate algorithms as step-by-step computations with the rule for accepting or rejecting the null hypothesis.