Lecture 12

Discrimination and Classification

Readings: Zelterman, 2015, Chapters 10

DSA 8070 Multivariate Analysis
November 1 - November 5, 2021

Whitney Huang
Clemson University

Notes

Agenda

Notes

Discrimination and Classification

Main objectives behind discrimination and classification are:

- Discrimination: separate distinct "populations" of observations
- Classification: classify new objectives into pre-defined populations

Examples

- Given measurements on the concentrations of five elements in bullet lead, find combinations of those concentrations that best describe bullets made by Cascade, Federal, Winchester and Remington ⇒ discrimination
- Using information on prisoners eligible for parole (good behavior, history of drug use, job skills, etc) can we successfully allocate a prisoners elegible for parole into two groups: those who will commit another crime or those who will not commit another crime? ⇒ classification

Notes

Classification

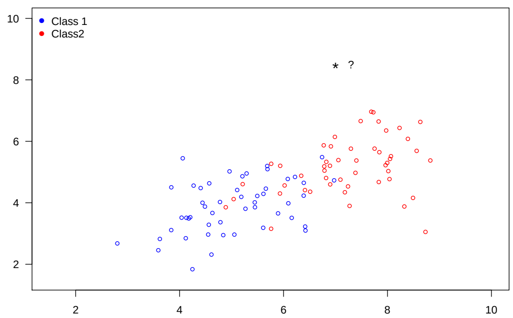
- **Data:**
$$\{\boldsymbol{X}_i, Y_i\}_{i=1}^n,$$
where Y_i is the class information for the i_{th} observation $\Rightarrow Y$ is a qualitative variable
- Classification aims to classify a new observation (or several new observations) into one of those classes

  Quantity of interest: $\mathrm{P}(Y = k_{th}$ category$|\boldsymbol{X} = \boldsymbol{x})$
- In this lecture we will focus on binary linear classification

Notes

Toy Example

Wish to classify a new observation $\boldsymbol{x}_i = (x_{1i}, x_{2i})$, denoted by $(*)$, into one of the two groups (class 1 or class 2)



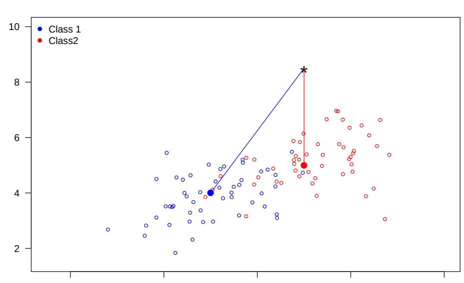
Notes

Toy Example Cont'd

We can compute the distances from this new observation $\boldsymbol{x} = (x_1, x_2)$ to the groups, for example,

$$d_1 = \sqrt{(x_1 - \mu_{11})^2 + (x_2 - \mu_{12})^2},$$
$$d_2 = \sqrt{(x_1 - \mu_{21})^2 + (x_2 - \mu_{22})^2}.$$

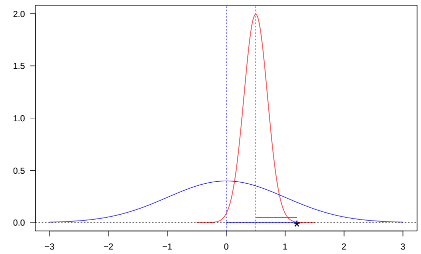
We can assign $\boldsymbol{x}$ to the group with the smallest distance



Notes

Variance Corrected Distance

In this one-dimensional example, $d_1 = |x - \mu_1| > |x - \mu_2|$. Does that mean x is "closer" to group 2 (red) than group 1 (blue)?



> We should take the "spread" of each group into account. $\tilde{d}_1 = |x - \mu_1|/\sigma_1 < \tilde{d}_2 = |x - \mu_2|/\sigma_2$

Notes

General Covariance Adjusted Distance: Mahalanobis Distance

The Mahalanobis distance [Mahalanobis, 1936] is a measure of the distance between a point $\boldsymbol{x}$ and a multivariate distribution of $\boldsymbol{X}$:

$$D_M(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})},$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\boldsymbol{X}$

One can use the Mahalanobis distance, by computing the Mahalanobis distance between an observations $\boldsymbol{x}_i$ and the "center" of the k_{th} population $\boldsymbol{\mu}_k$, to carry out classification

Notes

Binary Classification with Multivariate Normal Populations

Assume $\boldsymbol{X}_1 \sim \mathrm{MVN}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\boldsymbol{X}_2 \sim \mathrm{MVN}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, that is, $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$

- Maximum Likelihood of group membership:

  Group 1 if $\ell(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) > \ell(\boldsymbol{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

- Linear Discriminant Function:

  Group 1 if $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > 0$

- Minimize Mahalanobis distance:

  Group 1 if $(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) < (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)$

> All the methods above are equivalent

Notes

Priors and Misclassification Costs

In addition to the observed characteristics of units $\{\boldsymbol{x}_i\}_{i=1}^n$, other features of classification rules are:

- Prior probability:

  If one population is more prevalent than the other, chances are higher that a new unit came from the larger population. Stronger evidence would be needed to allocate the unit to the population with the smaller prior probability.

- Costs of misclassification:

  It may be more costly to misclassify a seriously ill subject as healthy than to misclassify a healthy subject as being ill.

Notes

Classification Regions

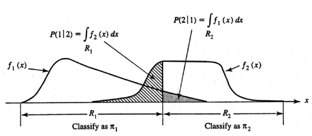
Ignore for now the prior probabilities of each population and the potentially different misclassification costs.

- The probability of misclassifying an object into π_2 when it belongs in π_1 is

$$P(2|1) = \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_2|\pi_1)$$

- The probability of misclassifying an object into π_1 when it belongs in π_2 is

$$P(1|2) = \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_1|\pi_2)$$



Source: Figure 11.3 from Applied Multivariate Statistical Analysis, 6th Ed (Johnson & Wichern). Visualization is for $p = 1$ variable.

Notes

Probability and Expected Cost of Misclassification

Let p_1 and p_2 denote the prior probabilities of π_1, π_2, and $c(1|2), c(2|1)$ be the costs of nisclassification:

- Then probabilities of the four possible outcomes are:

$$\begin{aligned}
\mathbb{P}(\text{correctly classified as } \pi_1) &= \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_1|\pi_1)\mathbb{P}(\pi_1) = P(1|1)p_1 \\
\mathbb{P}(\text{incorrectly classified as } \pi_1) &= \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_1|\pi_2)\mathbb{P}(\pi_2) = P(1|2)p_2 \\
\mathbb{P}(\text{correctly classified as } \pi_2) &= \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_2|\pi_2)\mathbb{P}(\pi_2) = P(2|2)p_2 \\
\mathbb{P}(\text{incorrectly classified as } \pi_2) &= \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_2|\pi_1)\mathbb{P}(\pi_1) = P(2|1)p_1
\end{aligned}$$

- Classification rules are often evaluated in terms of the expected cost of misclassification (ECM):

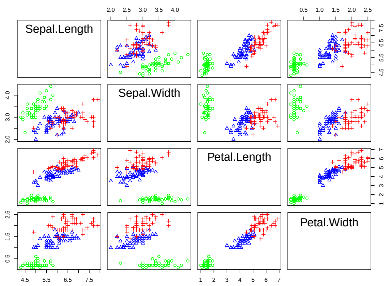
$$\text{ECM} = c(2|1)p(2|1)p_1 + c(1|2)P(1|2)P(1|2)p_2,$$

  and we seek rules that minimize the ECM

Notes

Example: Fisher's Iris Data

4 variables (sepal length and width and petal length and width), 3 species (setosa, versicolor, and virginica)

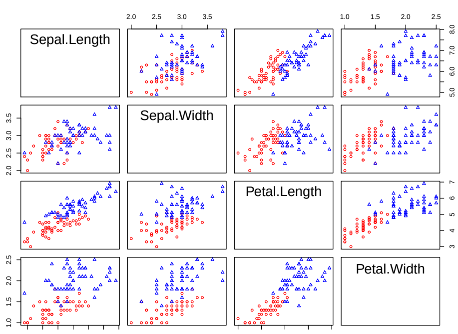


Task: Classify flowers into different species based on lengths and widths of sepal and petal

Notes

Fisher's Iris Data Cont'd

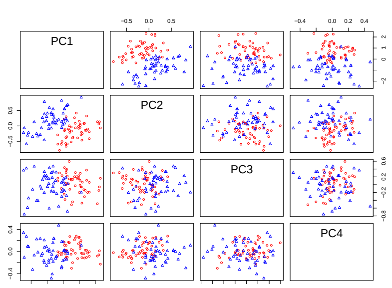
Let's focus on the latter two classes (versicolor, and virginica)



Notes

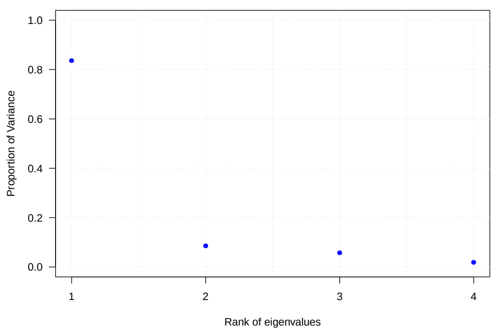
Fisher's iris Data Cont'd

To further simplify the matter, let's focus on the first two PCs of $\boldsymbol{X}$



Notes

Screen Plot



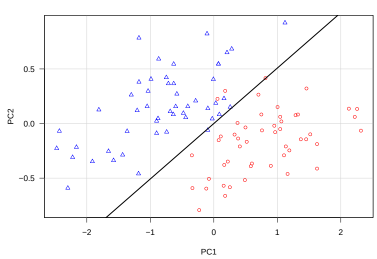
Notes

Linear Discriminant Analysis

Main idea: Use Bayes rule to compute

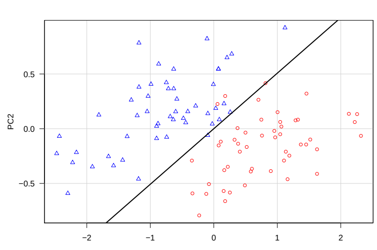
$$\mathrm{P}(Y=k|\boldsymbol{X}=\boldsymbol{x}) = \frac{\mathrm{P}(Y=k)\mathrm{P}(\boldsymbol{X}=\boldsymbol{x}|Y=k)}{\mathrm{P}(\boldsymbol{X}=\boldsymbol{x})} = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{k=1}^{K}\pi_k f_k(\boldsymbol{x})}$$

Assuming $f_k(\boldsymbol{x}) \sim \mathrm{MVN}(\boldsymbol{\mu}_k, \Sigma), \quad k = 1, \cdots, K$ and use $\hat{\pi}_k = \frac{n_k}{n} \Rightarrow$ it turns out the resulting classifier is linear in $\boldsymbol{X}$



Notes

Classification Performance Evaluation

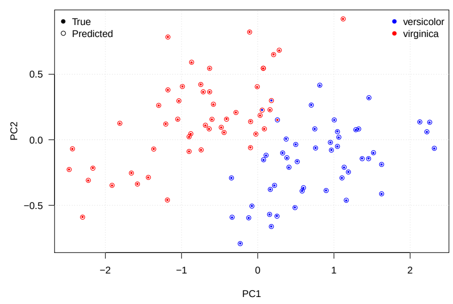


fit.LDA	versicolor	virginica
versicolor	47	3
virginica	1	49

Notes

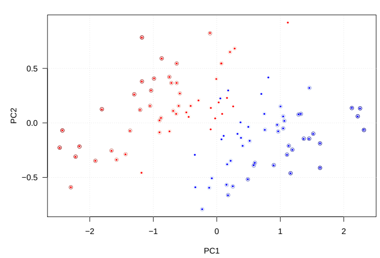
Logistic Regression Classifier

Main idea: Model the logit $\log\left(\frac{\mathrm{P}(Y=1)}{1-\mathrm{P}(Y=1)}\right)$ as a linear function in $\boldsymbol{x}$



Notes

Logistic Regression Classifier Cont'd




```
           logisticPred
            versicolor virginica
versicolor          48         2
virginica            1        49
```


Notes

Linear Discriminant Analysis Versus Logistic Regression

For a binary classification problem, one can show that both linear discriminant analysis (LDA) and logistic regression are linear classifiers. The difference is in how the parameters are estimated:

- Logistic regression uses the conditional likelihood based on $\mathrm{P}(Y|\boldsymbol{X}=\boldsymbol{x})$
- LDA uses the full likelihood based on multivariate normal assumption on $\boldsymbol{X}$
- Despite these differences, in practice the results are often very similar

Notes

Quadratic Discriminant Analysis

In linear discriminant analysis, we **assume** $\{f_k(\boldsymbol{x})\}_{k=1}^{K}$ are normal densities and $\boldsymbol{\Sigma_1 = \Sigma_2}$, therefore we obtain a linear classifier. What if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2 \Rightarrow$ we get quadratic discriminant analysis

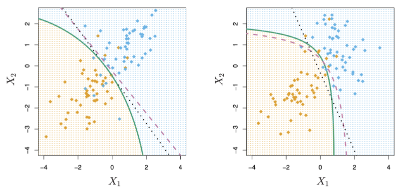


Figure: Figure courtesy of An Introduction of Statistical Learning by G. James et al. pp. 150

Notes

Notes

Notes