Lecture 38
Statistical Classification

STAT 8020 Statistical Methods II
December 2, 2019

Whitney Huang
Clemson University

Notes

Agenda

1. An Overview of Multivariate Analysis
2. Classification Problems
3. Linear Discriminant Analysis & Logistic Regression

Statistical Classification
CLEMSON UNIVERSITY
An Overview of Multivariate Analysis
Classification Problems
Linear Discriminant Analysis & Logistic Regression

Notes

An Overview of Multivariate Analysis

- In many studies, observations are collected on several variables on each experimental/observational unit
- Multivariate analysis is a collection of statistical methods for analyzing these multivariate data sets
- **Common Objectives**
  - Dimensionality reduction
  - Classification
  - Grouping (Clustering)

Statistical Classification
CLEMSON UNIVERSITY
An Overview of Multivariate Analysis
Classification Problems
Linear Discriminant Analysis & Logistic Regression

Notes

Multivariate Data

We display a multivariate data that contains n units on p variables using a matrix

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \cdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{pmatrix}$$

Summary Statistics

- Mean Vector: $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T$
- Variance-Covariance Matrix: $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{p}$, where $\sigma_{ii} = \mathrm{Var}(X_i), \quad i = 1, \cdots, p$ and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i \neq j$

Notes

Classification and Discriminant Analysis

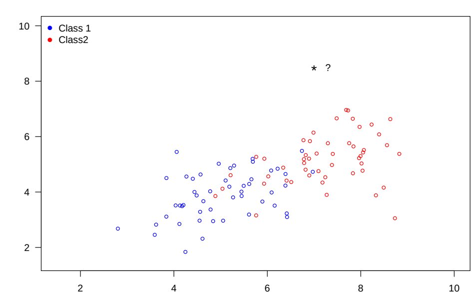
- **Data:**
$$\{\boldsymbol{X}_i, Y_i\}_{i=1}^{n},$$
where Y_i is the class information for the i_{th} observation $\Rightarrow Y$ is a qualitative variable
- Classification aims to classify a new observation (or several new observations) into one of those classes

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

Notes

Illustrating Example

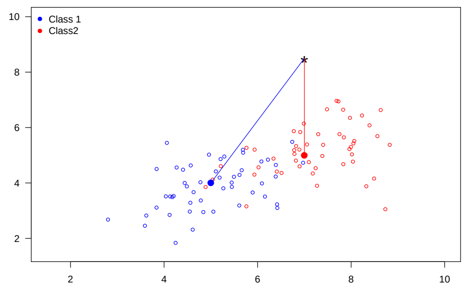
Wish to classify a new observation $z(*)$ into one of the two groups (class 1 or class 2)



Notes

Illustrating Example Cont'd

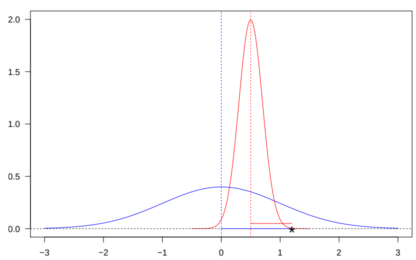
We could compute the distances from this new observation $\boldsymbol{z} = (z_1, z_2)$ to the groups, for example, $d_1 = \sqrt{(z_1 - \mu_{11})^2 + (z_2 - \mu_{12})^2}$, $d_2 = \sqrt{(z_1 - \mu_{21})^2 + (z_2 - \mu_{22})^2}$. We could assign $\boldsymbol{z}$ to the group with the smallest distance



Notes

Variance Corrected Distance

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



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

Notes

General Covariance Adjusted Distance: Mahalanobis Distance

The Mahalanobis distance is a measure of the distance between a point z and a distribution F:

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

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

Notes

Binary Classification

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

- Maximum Likelihood of group membership:

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

- Linear Discriminant Function:

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

- Minimize Mahalanobis distance:

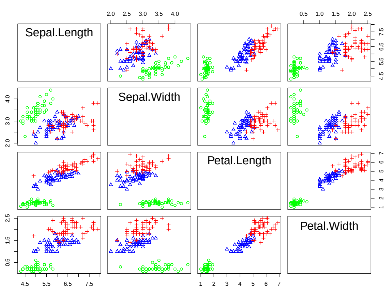
  Group 1 if $(\boldsymbol{z} - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_1) < (\boldsymbol{z} - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_2)$

All the classification methods above are equivalent

Notes

Example: Fisher's Iris Data

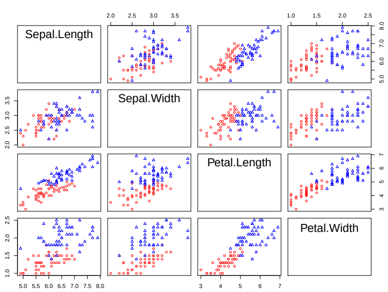
4 variables (sepal length and width and petal length and width), 3 species (setosa, versicolor, and virginica)



Notes

Fisher's Iris Data Cont'd

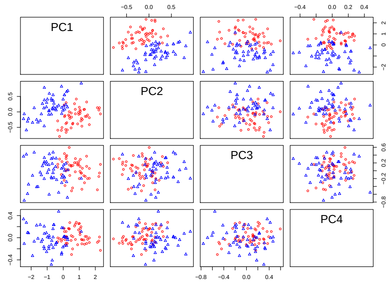
Let's focus on the latter two classes (versicolor, and virginica)



Notes

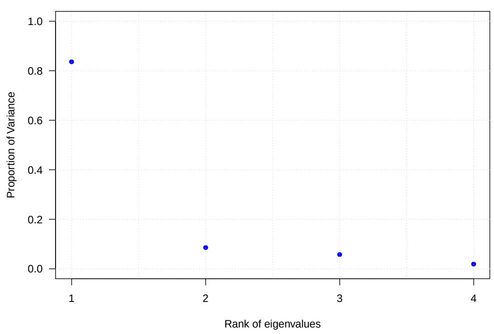
Fisher's iris Data Cont'd

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



Notes

Screen Plot



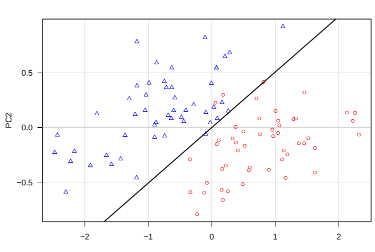
Statistical Classification

An Overview of Multivariate Analysis
Classification Problems
Linear Discriminant Analysis & Logistic Regression

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

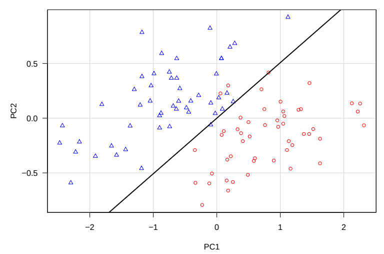


Statistical Classification

An Overview of Multivariate Analysis
Classification Problems
Linear Discriminant Analysis & Logistic Regression

Notes

Classification Performance Evaluation



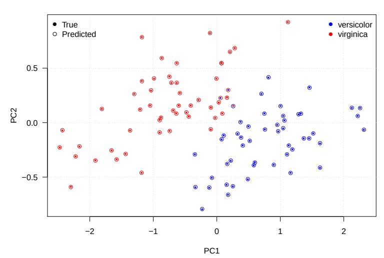
fit.LDA

	versicolor	virginica
versicolor	47	3
virginica	1	49

Notes

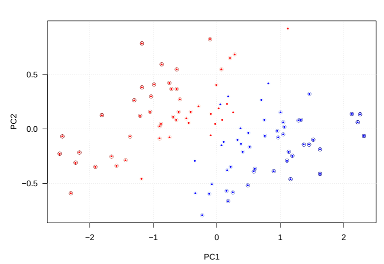
Logistic Regression Classifier

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



Notes

Logistic Regression Classifier Cont'd



logisticPred

	versicolor	virginica
versicolor	48	2
virginica	1	49

Notes

Quadratic Discriminant Analysis

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

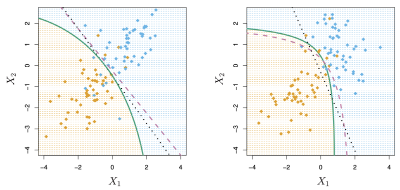


Figure: Figure courtesy of An Introduction of Statistical Learning by G. James et al. pp. 150

Notes

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

Notes