Lecture 25

Classification & Cluster Analysis

STAT 8020 Statistical Methods II
November 24, 2020

Whitney Huang
Clemson University

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: $\mathrm{P}(Y = k_{th}\ \text{category} | \boldsymbol{X} = \boldsymbol{x})$

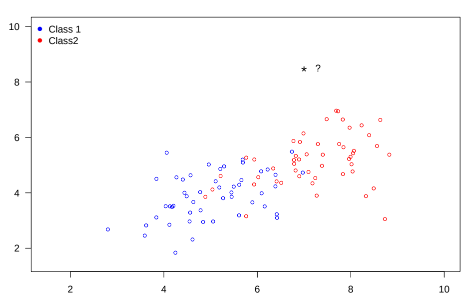
- In this lecture we will focus on binary linear classification

Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

Illustrating Example

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

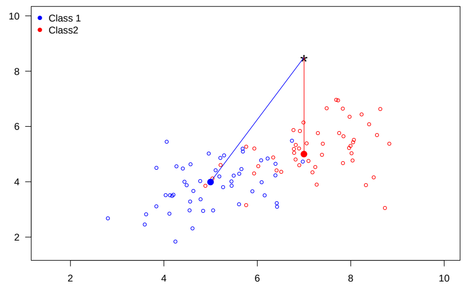


Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

Illustrating Example Cont'd

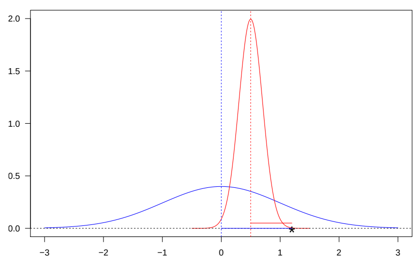
We could compute the distances from this new observation $\boldsymbol{z} = (z_1, z_2)$ to the groups, for example, $d_1 = \sqrt{(z_1 - \mu_{11})^2 + (z_2 - \mu_{12})^2}$, $d_2 = \sqrt{(z_1 - \mu_{21})^2 + (z_2 - \mu_{22})^2}$. We could assign $\boldsymbol{z}$ to the group with the smallest distance



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 $\boldsymbol{z}$ 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:

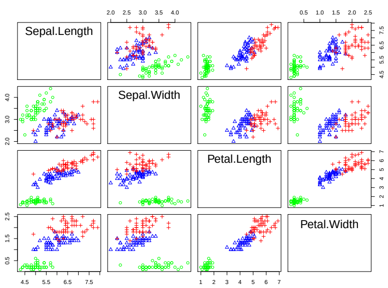
  Group 1 if $(\boldsymbol{z} - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_1) < (\boldsymbol{z} - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_2)$

All the classification methods above are equivalent

Notes

Example: Fisher's Iris Data

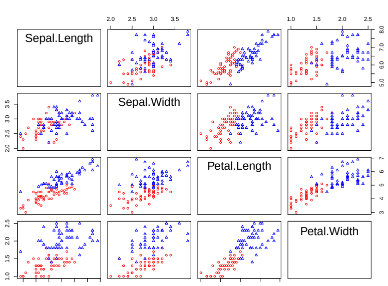
4 variables (sepal length and width and petal length and width), 3 species (setosa, versicolor, and virginica)



Notes

Fisher's Iris Data Cont'd

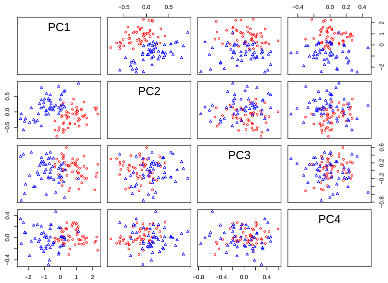
Let's focus on the latter two classes (versicolor, and virginica)



Notes

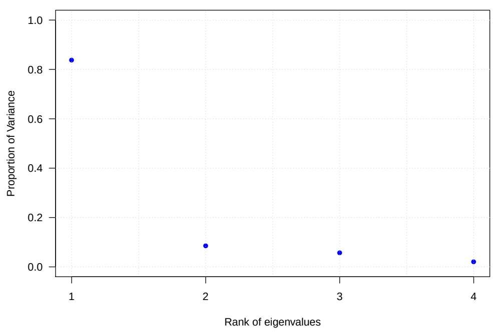
Fisher's iris Data Cont'd

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



Notes

Screen Plot

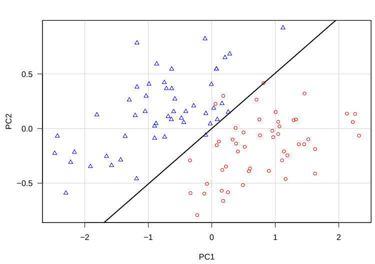


Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

Linear Discriminant Analysis

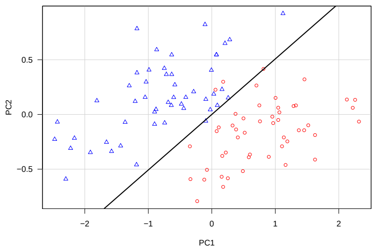
Main idea: Use Bayes rule to compute $\mathrm{P}(Y=k|\boldsymbol{X}=\boldsymbol{x}) = \frac{\mathrm{P}(Y=k)\mathrm{P}(\boldsymbol{X}=\boldsymbol{x}|Y=k)}{\mathrm{P}(\boldsymbol{X}=\boldsymbol{x})} = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k f_k(\boldsymbol{x})}$.
Assuming $f_k(\boldsymbol{x}) \sim \mathrm{MVN}(\boldsymbol{\mu}_k, \Sigma), \quad k = 1, \cdots, K$. Use $\hat{\pi}_k = \frac{n_k}{n} \Rightarrow$ it turns out the resulting classifier is linear in $\boldsymbol{X}$



Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

Classification Performance Evaluation



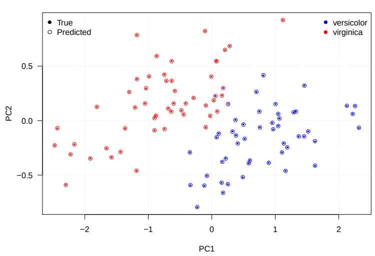
fit.LDA

	versicolor	virginica
versicolor	47	3
virginica	1	49

Notes

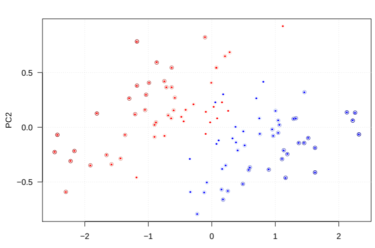
Logistic Regression Classifier

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

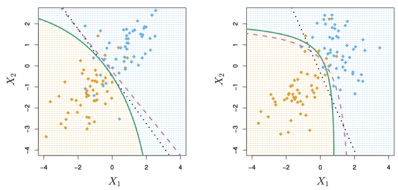


Figure: Figure courtesy of An Introduction of Statistical Learning by G. James et al. pp. 150

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

What is Cluster Analysis?

- Cluster: a collection of data objects
  - “Similar” to one another within the same cluster
  - “Dissimilar” to the objects in other clusters
- Cluster analysis: Grouping a set of data objects into clusters
- Clustering is unsupervised classification, unlike classification, there is no predefined classes, and the number of clusters is usually unknown

Notes

Some Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Notes

What Is Good Clustering?

Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

- A good clustering method will produce clusters with
  - high within-class similarity
  - low between-class similarity
- The quality of a clustering result depends on both the similarity measure used and its implementation
- The performance of a clustering method is measured by its ability to discover the hidden patterns

Notes

Major Clustering Approaches

Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

- **Partitioning algorithm:** partition the observations into a pre-specified number of clusters, for example, k-means clustering
- **Hierarchy algorithm:** Construct a hierarchical decomposition of the observations to build a hierarchy of clusters, for example, hierarchical agglomerative clustering
- **Model-based Clustering:** A model is hypothesized for each of the clusters, for example, Gaussian mixture models

Notes

Partitioning Algorithm

Let $C_1, \cdots, C_K$ denote sets containing the indices of the observations $\{\boldsymbol{x}_i\}_{i=1}^n$ in each cluster. These sets satisfy two properties:

- $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \cdots, n\} \Rightarrow$ each observation belongs to at least one of the K clusters
- $C_k \cap C_{k'} = \emptyset \, \forall k \neq k' \Rightarrow$ no observation belongs to more than one cluster

For instance, if the i_{th} observation (i.e. $\boldsymbol{x}_i$) is in the k_{th} cluster, then $i \in C_k$

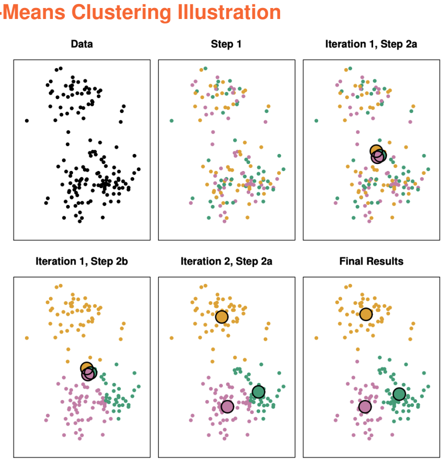
Notes

The k-Means Algorithm

- **Step 0:** Choose the number of clusters K
- **Step 1:** Randomly assign a cluster (from 1 to K), to each of the observations. These serve as the initial cluster assignmemts
- **Step 2:** Iterate until the cluster assignment stop changing
  - For each of the K cluster, compute the cluster centroid. The k_{th} cluster centroid is the mean vector of the observations in the k_{th} cluster
  - Assign each observations to the cluster whose centroid is closest in terms of Euclidean distance

Notes

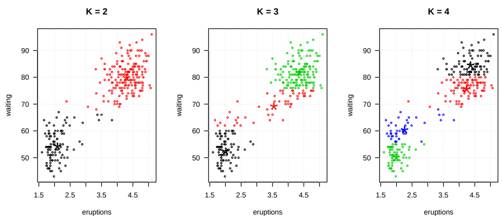
k-Means Clustering Illustration



Notes

K-Means Clustering in R

```
kmean3.faithful <- kmeans(x = faithful, centers = 3)
```



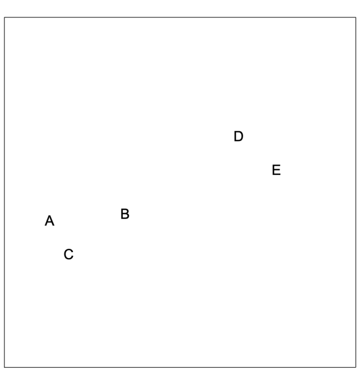
Notes

Hierarchical Clustering

- k-means clustering requires us to pre-specify the number of clusters K
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K
- Agglomerative clustering: This is a “bottom-up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy

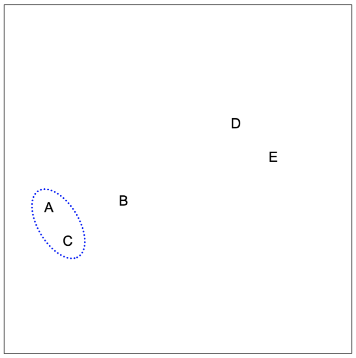
Notes

Hierarchical Agglomerative Clustering Illustration



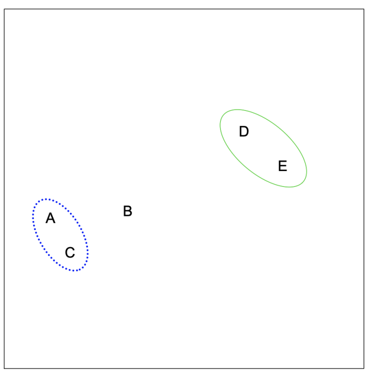
Notes

Hierarchical Agglomerative Clustering Illustration



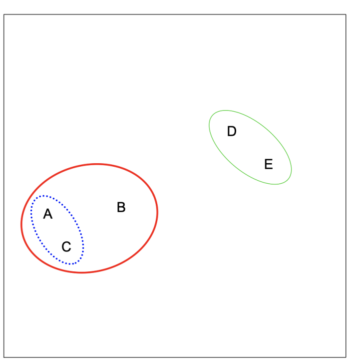
Notes

Hierarchical Agglomerative Clustering Illustration



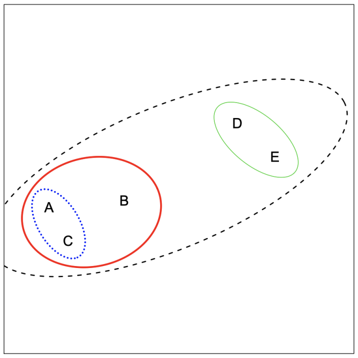
Notes

Hierarchical Agglomerative Clustering Illustration



Notes

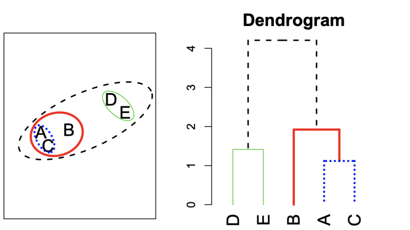
Hierarchical Agglomerative Clustering Illustration



Notes

Hierarchical Agglomerative Clustering Algorithm

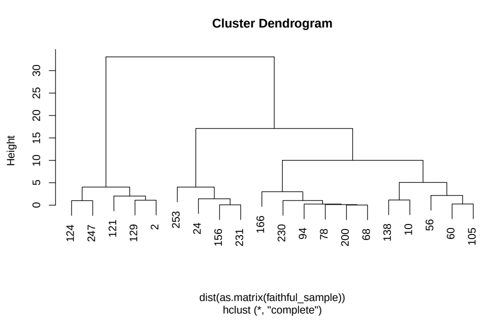
1. Start with each observation in its own cluster
2. Identify the closest two clusters and merge them
3. Repeat
4. Ends when all observations are in a single cluster



Notes

Hierarchical Agglomerative Clustering in R

```
hc.faithful <- hclust(dist(faithful_sample))
plot(hc.faithful)
```



Notes

Model-based clustering

- One disadvantage of hierarchical clustering and k-means is that they are largely heuristic and not based on formal statistical models. Formal inference is not possible
- Model-based clustering is an alternative:
  - Sample observations arise from a mixture distribution of two or more components
  - Each component (cluster) is described by a probability distribution and has an associated probability in the mixture.
  - In Gaussian mixture models, we assume each cluster follows a multivariate normal distribution
  - Therefore, in Gaussian mixture models, the model for clustering is a mixture of multivariate normal distributions

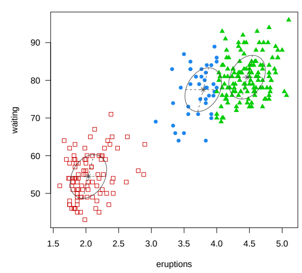
Notes

Fitting a Gaussian Mixture Model in R

```
library(mclust)
```

```
## Package 'mclust' version 5.4.5
## Type 'citation("mclust")' for citing this R package in publications.
```

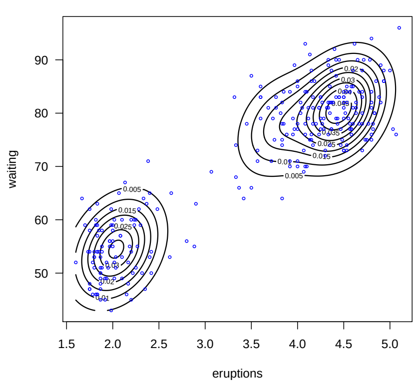
```
BIC <- mclustBIC(faithful)
model1 <- Mclust(faithful, x = BIC)
```



Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

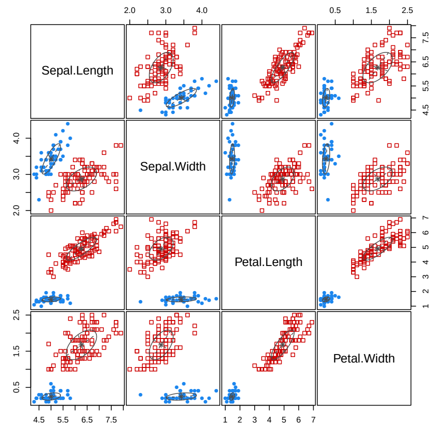
Fitting a Gaussian Mixture Model in R Cond't



Classification & Cluster Analysis
CLEMSON UNIVERSITY
Classification Problems
Linear Discriminant Analysis & Logistic Regression
An Overview of Cluster Analysis
The K-Means Algorithm
Hierarchical Clustering
Model-based clustering

Notes

Model-Based Clustering Analysis for Iris Data



Notes

Notes

Notes