

# DATA ANALYTICS AND BUSINESS INTELLIGENCE (8696/8697)

## CLUSTER ANALYSIS

Graham.Williams@togaware.com

Data Scientist  
Australian Taxation Office

Adjunct Professor, University of Canberra

Graham.Williams@togaware.com  
<http://datamining.togaware.com>



# OVERVIEW

- 1 CLUSTER ANALYSIS
  - Introduction
  - Requirements
- 2 MEASURING SIMILARITY
  - Distances
  - Data Types
- 3 ALGORITHMS
  - Measure Cluster Quality
  - Cluster Methods
  - K-Means



# OVERVIEW

- 1 CLUSTER ANALYSIS
  - Introduction
  - Requirements
- 2 MEASURING SIMILARITY
  - Distances
  - Data Types
- 3 ALGORITHMS
  - Measure Cluster Quality
  - Cluster Methods
  - K-Means



# WHAT IS CLUSTER ANALYSIS?

- 1 How do we understand the behaviour of an individual?
  - “Paint everyone with the same brush;”
  - Treat everyone as an individual.
- 2 How do we understand our fellow human beings in the world?
  - Through understanding every individual in the world?
- 3 We categorise, for *good or bad*, observations into groups:
  - Socio-economic groups: “the poor”, “the rich”;
  - Political: a lefty, a new right;
  - Racial, religious, geographical, . . . ;
- 4 We find that to understand our world we generally talk about groups, not individuals. But computers don’t need to—they increasingly have the power to build an understanding of the individual, for *better or worse*—Facebook, gmail, . . . .



# WHAT IS CLUSTER ANALYSIS?

- 1 How do we understand the behaviour of an individual?
  - “Paint everyone with the same brush;”
  - Treat everyone as an individual.
- 2 How do we understand our fellow human beings in the world?
  - Through understanding every individual in the world?
- 3 We categorise, for *good or bad*, observations into groups:
  - Socio-economic groups: “the poor”, “the rich”;
  - Political: a lefty, a new right;
  - Racial, religious, geographical, . . . ;
- 4 We find that to understand our world we generally talk about groups, not individuals. But computers don’t need to—they increasingly have the power to build an understanding of the individual, for *better or worse*—Facebook, gmail, . . . .



# WHAT IS CLUSTER ANALYSIS?

- 1 How do we understand the behaviour of an individual?
  - “Paint everyone with the same brush;”
  - Treat everyone as an individual.
- 2 How do we understand our fellow human beings in the world?
  - Through understanding every individual in the world?
- 3 We categorise, for *good or bad*, observations into groups:
  - Socio-economic groups: “the poor”, “the rich”;
  - Political: a lefty, a new right;
  - Racial, religious, geographical, . . . ;
- 4 We find that to understand our world we generally talk about groups, not individuals. But computers don't need to—they increasingly have the power to build an understanding of the individual, for *better or worse*—Facebook, gmail, . . . .



# WHAT IS CLUSTER ANALYSIS?

- 1 How do we understand the behaviour of an individual?
  - “Paint everyone with the same brush;”
  - Treat everyone as an individual.
- 2 How do we understand our fellow human beings in the world?
  - Through understanding every individual in the world?
- 3 We categorise, for *good or bad*, observations into groups:
  - Socio-economic groups: “the poor”, “the rich”;
  - Political: a lefty, a new right;
  - Racial, religious, geographical, . . . ;
- 4 We find that to understand our world we generally talk about groups, not individuals. But computers don’t need to—they increasingly have the power to build an understanding of the individual, for *better or worse*—Facebook, gmail, . . . .



# SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.





## SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.



## SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.



## SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.



## SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.



## SPECIFIC EXAMPLES—WHERE CLUSTERS HELP

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Spatial Data Analysis:** Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Fraud:** Identifying groups of individuals that, as a group, are very different to the other groups.
- **WWW:** Document classification, question categorisation, and web log data to discover similar access patterns.
- **City Planning:** Identify services for households according to their house type, value, and geographic location.



# WHAT IS CLUSTER ANALYSIS?

- **Cluster:** a collection of observations
  - Similar to one another within the same cluster
  - Dissimilar to the observations in other clusters
- **Cluster analysis**
  - Grouping a set of data observations into classes
- Clustering is **unsupervised classification**: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



# WHAT IS CLUSTER ANALYSIS?

- **Cluster:** a collection of observations
  - Similar to one another within the same cluster
  - Dissimilar to the observations in other clusters
- **Cluster analysis**
  - Grouping a set of data observations into classes
- Clustering is **unsupervised classification**: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



# WHAT IS CLUSTER ANALYSIS?

- **Cluster:** a collection of observations
  - Similar to one another within the same cluster
  - Dissimilar to the observations in other clusters
- **Cluster analysis**
  - Grouping a set of data observations into classes
- Clustering is **unsupervised classification**: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms





# WHAT IS CLUSTER ANALYSIS?

- **Cluster:** a collection of observations
  - Similar to one another within the same cluster
  - Dissimilar to the observations in other clusters
- **Cluster analysis**
  - Grouping a set of data observations into classes
- Clustering is **unsupervised classification**: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



# WHAT IS GOOD CLUSTERING?

- High Quality:
  - high intra-class similarity
  - low inter-class similarity
- The Quality depends on:
  - similarity measure
  - algorithm for searching
- *Depends on the opinion of the user, and the algorithm's ability to discover hidden patterns that are of interest to the user.*



# WHAT IS GOOD CLUSTERING?

- High Quality:
  - high intra-class similarity
  - low inter-class similarity
- The Quality depends on:
  - similarity measure
  - algorithm for searching
- *Depends on the opinion of the user, and the algorithm's ability to discover hidden patterns that are of interest to the user.*



# WHAT IS GOOD CLUSTERING?

- High Quality:
  - high intra-class similarity
  - low inter-class similarity
- The Quality depends on:
  - similarity measure
  - algorithm for searching
- *Depends on the opinion of the user, and the algorithm's ability to discover hidden patterns that are of interest to the user.*



# CLUSTERING CAVEATS

Clustering may not be the best way to discover interesting groups in a data set. Often visualisation methods work well, allowing the human expert to identify useful groups. However, as the data set sizes increase to millions of observations, this becomes impractical and clusters help to partition the data so that we can deal with smaller groups.

Different algorithms, and even multiple runs of the one algorithm, will deliver different clusterings.



# CLUSTERING CAVEATS

Clustering may not be the best way to discover interesting groups in a data set. Often visualisation methods work well, allowing the human expert to identify useful groups. However, as the data set sizes increase to millions of observations, this becomes impractical and clusters help to partition the data so that we can deal with smaller groups.

Different algorithms, and even multiple runs of the one algorithm, will deliver different clusterings.



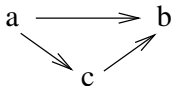
# OVERVIEW

- 1 CLUSTER ANALYSIS
  - Introduction
  - Requirements
- 2 MEASURING SIMILARITY
  - Distances
  - Data Types
- 3 ALGORITHMS
  - Measure Cluster Quality
  - Cluster Methods
  - K-Means



# SIMILARITY AND DISSIMILARITY BETWEEN OBSERVATIONS

- “Distance” measures the **dissimilarity** between two data observations  $x = (x_1, x_2, \dots, x_n)$  and  $y = (y_1, y_2, \dots, y_n)$ .
- A distance measure should satisfy the following requirements:
  - $d(x, y) \geq 0$  distance is non-negative
  - $d(x, y) = 0$  distance to itself is 0
  - $d(x, y) = d(y, x)$  distance is symmetric
  - $d(x, y) \leq d(x, z) + d(z, y)$  triangular inequality

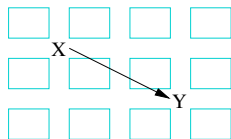




# EUCLIDEAN DISTANCE

- The straight line distance between two points;
- Default measure based on numeric values;
- The distance calculation we learned in school

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

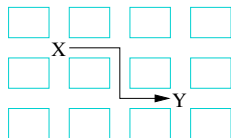


- $x$  and  $y$  are the observations;
- $n$  variables;
- $x_i$  is the value of variable  $i$  for observation  $x$ ;
- Similarly for  $y_i$ .

# MANHATTAN DISTANCE

- The distance walking the streets of Manhattan.

$$\text{dist}(x, y) = \sum_{i=1}^n |x_i - y_i|$$



# MINKOWSKI DISTANCE

A general measure of distance:

$$d(x, y) = \sqrt[q]{(|x_1 - y_1|^q + |x_2 - y_2|^q + \dots + |x_n - y_n|^q)}$$

- If  $q = 2$ ,  $d$  is the **Euclidean distance**.
- If  $q = 1$ ,  $d$  is the **Manhattan distance**.
- A variation is the weighted distance variables have different importance:

$$d(x, y) = \sqrt[q]{(w_1|x_1 - y_1|^q + w_2|x_2 - y_2|^q + \dots + w_n|x_n - y_n|^q)}$$



## ISSUE OF SCALE

- Scale plays havoc on distance measures!
- Compare age and income:
  - $x = (25\text{years}, \$35,000)$
  - $y = (55\text{years}, \$125,000)$
- Calculate differences:
  - Difference in age: 30
  - Difference in income: 90000
- How can we add them together: 90030?
- Normalise using a z-score:  $x'_i = \frac{x_i - \text{mean}(X)}{\text{stdev}(X)}$
- Now  $|-0.4 - 0.6| + |-0.1 - 0.7| = 1.0 + 0.8 = 1.8$
- The difference in age is more significant than in income.



# TYPE OF DATA IN CLUSTERING ANALYSIS

A distance measure works well for numeric variables, but what about other types of variables?

The types of variables include:

- Numeric, interval-scaled variables
- Binary variables
- Categorical, ordinal, and ratio variables
- Variables of mixed types



# INTERVAL-SCALED VARIABLES

*Interval-scaled variables are continuous variables of a roughly linear scale.*

- The Euclidean distance or some other instance of the Minkowski distance can be used.
- Before applying the distance measure, the variables need to be normalized:
  - Variables with larger ranges (e.g., *income*) will overwhelm variables with smaller ranges (e.g., *age*):  
 $\$50,000 - \$40,000 = 10,000$  versus  $50\text{years} - 40\text{years} = 10$
  - Variation of *z-score normalisation*:

$$v' = \frac{v - m}{s}$$

where  $m$  is the **mean** and  $s$  is the **mean absolute deviation**  
(c.f. **stdev**: robust to outliers and retains the outliers)



## BINARY VARIABLES

*Binary variables have just two possible values: 0 and 1.*

We consider as a group all of the binary variables and count for observation  $x_i$  and  $x_j$  the number of times they both have 0, 1, or (0, 1) or (1, 0) to build a contingency table:

	1	0	sum
1	a	b	a+b
0	c	d	c+d
sum	a+c	b+d	n

- Simple matching coefficient (symmetric variable):

$$d(x_i, x_j) = \frac{b + c}{a + b + c + d}$$

- Jaccard coefficient  
(asymmetric: 1 is more important - e.g. diseases):

$$d(x_i, x_j) = \frac{b + c}{a + b + c}$$



# CATEGORICAL VARIABLES

- A generalisation of the binary variable in that it can take more than 2 levels, e.g., red, yellow, blue, green.
- Method 1: Simple matching

$$d(x, y) = \frac{n - p}{n}$$

where  $p$  is the number of matched categorical variables and  $n$  is the total number of variables.

- Method 2: Convert each level into a binary variable, creating many new binary variables.





# VARIABLES OF MIXED TYPES

- A dataset may contain all types of variables: interval, binary, categorical.
- Use a weighted formula to combine the different normalised (to  $[0, 1]$ ) distances, where the weights are used to express the relative importance of the variables:

$$d(x_i, x_j) = \sum_k w_k d_{ij}^{A_k}$$

where  $w_k$  is the weight of variable  $A_k$ ,  $d_{ij}^{A_k}$  is the dissimilarity of the  $i$ th observation and the  $j$ th observation on variable  $A_k$ .  $d_{ij}^{A_k}$  is normalized to  $[0, 1]$



# OVERVIEW

- 1 CLUSTER ANALYSIS
  - Introduction
  - Requirements
- 2 MEASURING SIMILARITY
  - Distances
  - Data Types
- 3 ALGORITHMS
  - Measure Cluster Quality
  - Cluster Methods
  - K-Means



# MEASURE CLUSTER QUALITY

The goal is to obtain:

- high intra-class similarity
- low inter-class similarity

We can sum the square of the distances between the observations within a cluster, and total this up for each of the clusters — this is the **within sum of squares** measure — the intra-class similarity.

We can sum the squares of the distances between the observations in different clusters — this is the **between sum of squares** — the inter-class similarity.



# REQUIREMENTS OF CLUSTERING IN DATA MINING

- Scalability—many observations and variables
- Different variable types—limit to numerics
- Clusters with arbitrary shape
- Minimal domain knowledge required
- Can cope with noise and outliers
- Insensitive to order of input records
- High dimensionality—curse of dimensionality



# MAJOR CLUSTERING APPROACHES

- **Partitioning algorithms** (`kmeans`, `pam`, `clara`, `fanny`): Construct various partitions and then evaluate them by some criterion. A fixed number of clusters,  $k$ , is generated. Start with an initial (perhaps random) cluster.
- **Hierarchical algorithms**: (`hclust`, `agnes`, `diana`) Create a hierarchical decomposition of the set of observations using some criterion
- **Density-based algorithms**: based on connectivity and density functions
- **Grid-based algorithms**: based on a multiple-level granularity structure
- **Model-based algorithms**: (`mclust` for mixture of Gaussians) A model is hypothesized for each of the clusters and the idea is to find the best fit of that model



# BASIC PARTITIONING ALGORITHM

- Partition dataset  $D$  of  $n$  observations into  $k$  clusters
  - Given  $k$ , find  $k$  clusters that optimises partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  
- Heuristic methods: k-means and k-medoids algorithms
  - k-means: Each cluster represented by center of the cluster
  - k-medoids or PAM (partition around medoids): Each cluster represented by one of the observations in the cluster

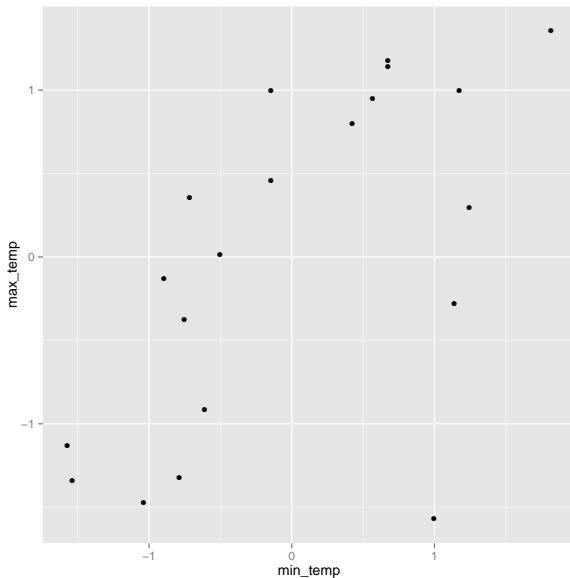


# THE K-MEANS CLUSTERING METHOD

- Given  $k$ , the k-means algorithm is implemented as:
  - 1 Identify  $k$  random means.
  - 2 Assign each observation to the nearest mean.
  - 3 Recalculate the new means from the new clustering.
  - 4 Repeat from 2 until no observations change clustering.
- Example iterations follow:
  - Identify two clusters;
  - Using weather dataset from Rattle.

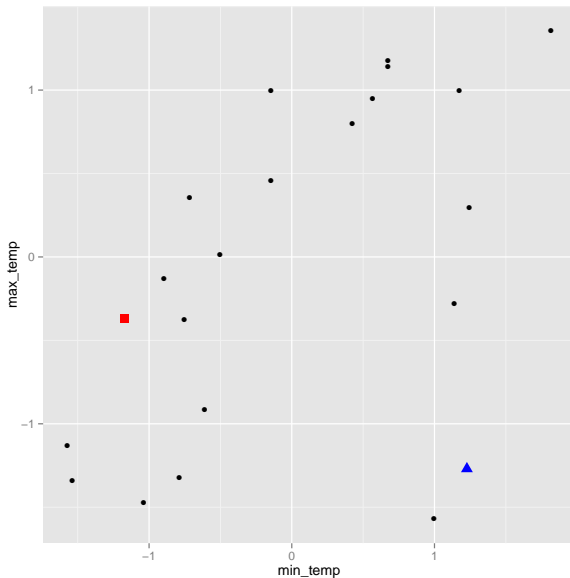


# K-MEANS: ORIGINAL DATA

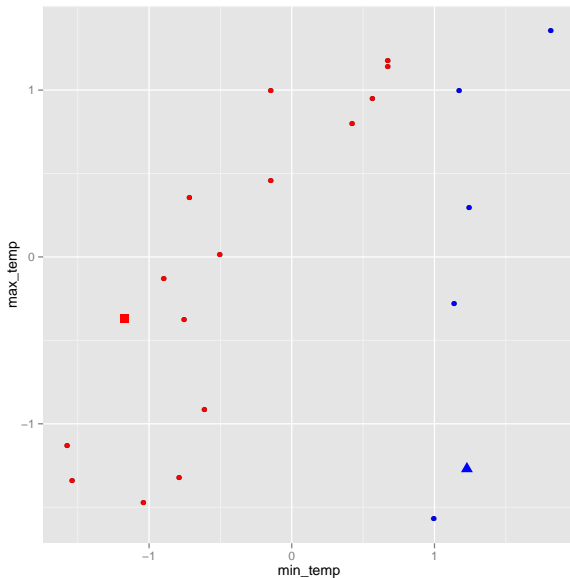




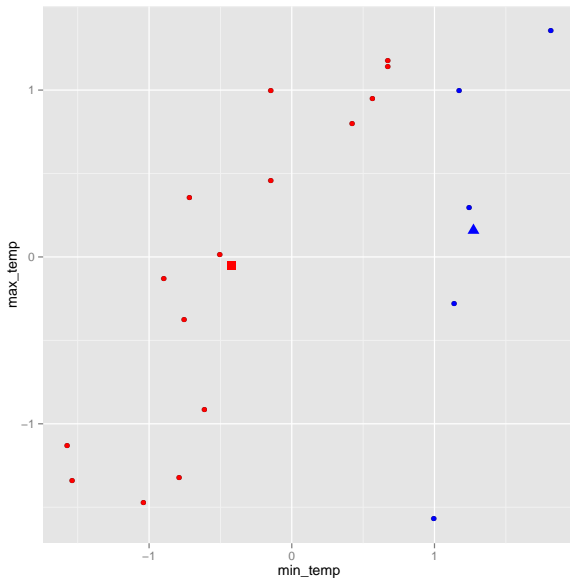
# K-MEANS: INITIALISE THE SEARCH



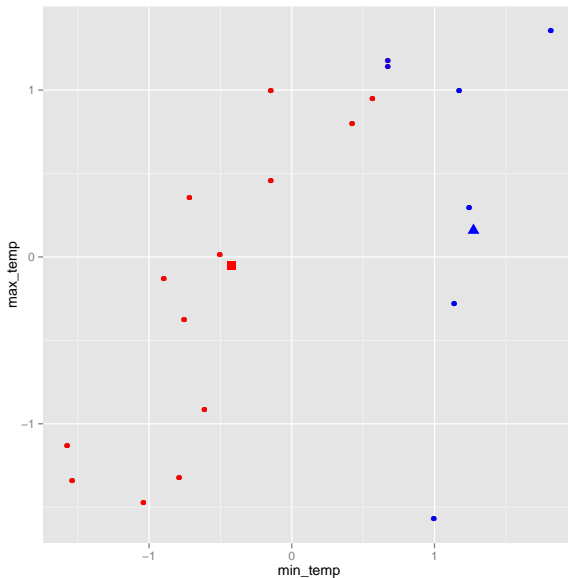
# K-MEANS: IDENTIFY CLOSEST MEANS



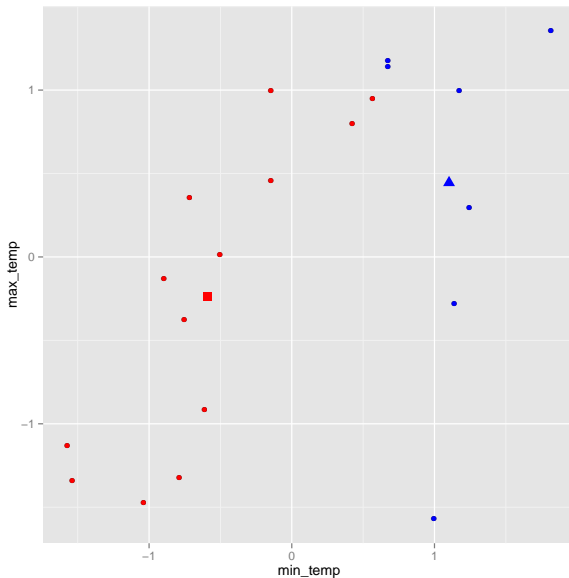
# K-MEANS: RECALCULATE MEANS



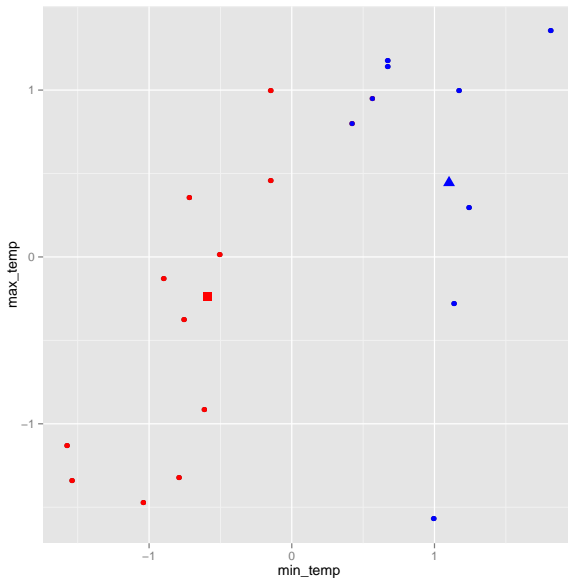
# K-MEANS: IDENTIFY NEW CLOSEST MEANS



# K-MEANS: RECALCULATE NEW MEANS

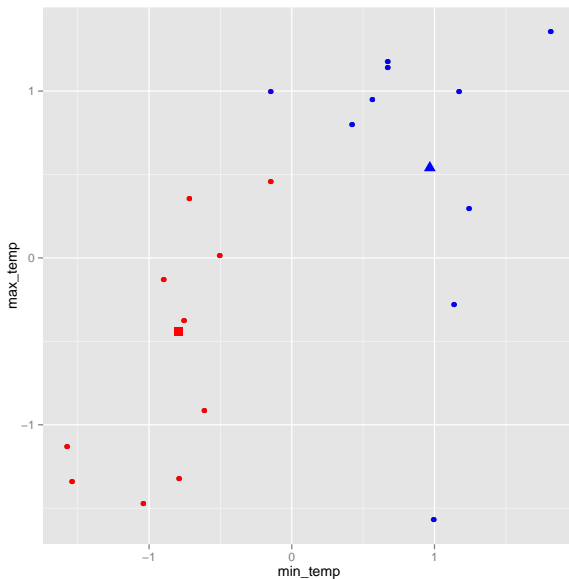


# K-MEANS: IDENTIFY NEW NEW CLOSEST MEANS





# K-MEANS: IDENTIFY CLOSEST MEANS AGAIN





# COMMENTS ON K-MEANS

- Strengths
  - Relatively efficient:  $O(tkn)$ , where  $n$  is the number observations,  $k$  is the number of clusters, and  $t$  is the number iterations. Normally,  $k, t \ll n$ .



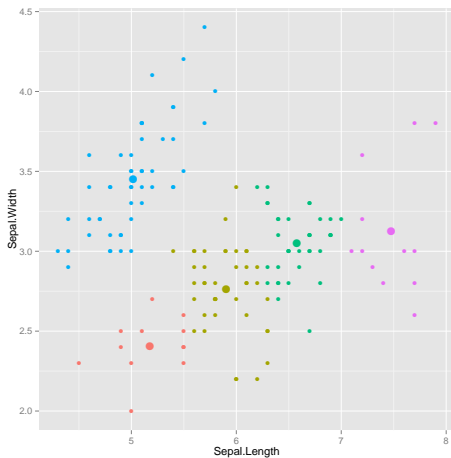
# COMMENTS ON K-MEANS

- Weaknesses
  - Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms
  - Applicable only when the mean is defined—what about categorical data?
  - Need to specify  $k$ , the number of clusters, in advance. There are algorithms that can help with this.
  - Unable to handle noisy data and outliers.
  - Not suitable for non-convex clusters.



# KMEANS IN R

```
k <- 5  
model <- kmeans(iris[1:2], k)
```



# KMEANS ITERATIONS

Experiment with the iterations at this week's tutorial:

```
k <- 5
set.seed(42)

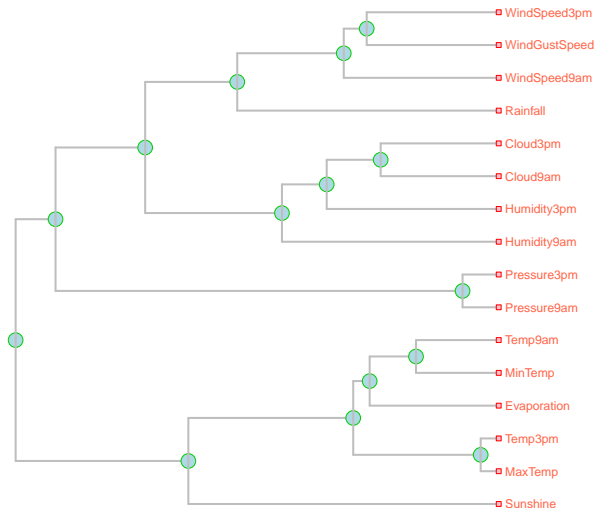
iris.kmeans <- kmeans(iris[1:2], k, iter.max=1)
plot(iris[1:2], col=iris.kmeans$cluster)
points(iris.kmeans$centers, pch=19, cex=1.5, col=1:k)

x11() # or windows() to start a new plot
iris.kmeans <- kmeans(iris[1:2], k, iter.max=2)
plot(iris[1:2], col=iris.kmeans$cluster)
points(iris.kmeans$centers, pch=19, cex=1.5, col=1:k)

x11() # or windows() to start a new plot
iris.kmeans <- kmeans(iris[1:2], k, iter.max=3)
plot(iris[1:2], col=iris.kmeans$cluster)
points(iris.kmeans$centers, pch=19, cex=1.5, col=1:k)
```



# RATTLE: HIERARCHICAL VARIABLE CLUSTER



# OVERVIEW

- 1 CLUSTER ANALYSIS
  - Introduction
  - Requirements
- 2 MEASURING SIMILARITY
  - Distances
  - Data Types
- 3 ALGORITHMS
  - Measure Cluster Quality
  - Cluster Methods
  - K-Means



# MODELLING FRAMEWORK

**Language** Set of means

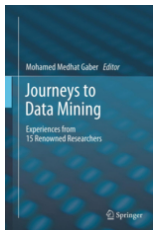
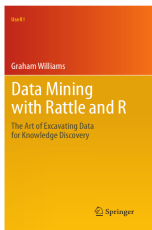
**Measure** Minimise intra cluster distance,  
maximise inter cluster distance

**Search** Random assignment, calculate mean, reassign, . . .



# RESOURCES AND REFERENCES

- **OnePageR:** <http://onepager.togaware.com>:
  - Using animate to watch the k-means movie;
  - Selecting  $k$ ;
  - Visualising clusters.





# SUMMARY

- Cluster analysis is unsupervised learning.
- Useful for partitioning a very large population, perhaps for data mining each sub-population separately.
- Often more effective under expert guidance.



# SUMMARY

- Cluster analysis is unsupervised learning.
- Useful for partitioning a very large population, perhaps for data mining each sub-population separately.
- Often more effective under expert guidance.



# SUMMARY

- Cluster analysis is unsupervised learning.
- Useful for partitioning a very large population, perhaps for data mining each sub-population separately.
- Often more effective under expert guidance.



# SUMMARY

- Cluster analysis is unsupervised learning.
- Useful for partitioning a very large population, perhaps for data mining each sub-population separately.
- Often more effective under expert guidance.

