JAMES COOK UNIVERSITY
AUSTRALIA

19-B-MA5821-ONL-EXT-SP85

Advanced Statistical Methods for Data Scientists

Week-5

Presented by
Zilani (JCU)
mohammed.zilani@jcu.edu.au
mgzilani15@gmail.com

Main Focus:

- Week4 Quiz
- Basic concept of Tree
- Bagging
- Random Forests
- Clustering

Week4 Quiz

Question 1

10 Points

Which of the following is **not true** about logistic regression?

Choose at least one correct answer

- (A) A prediction estimate for the target variables is formed from a simple linear combination of the inputs *Correct answer*
- (B) The model predicts the probability of a particular level of the target variable at the given values of the input variables
- (C) The logit link function is one of the most common ways to make predictions, because it makes it easy to interpret the model
- (D) Two ways to interpret a logistic regression model are an odds ratio and a doubling amount

Week4 Quiz

Question 2

10 Points

Which of the following statements applies to decreasing a model's degrees of freedom and preventing model overfitting?

Choose at least one correct answer

- (A) You can consolidate categorical inputs — *Correct answer*
- (B) To reduce the degrees of freedom used by the model, several levels of the categorical variable can be assigned to a single dummy variable — *Correct answer*
- (C) To reduce the degrees of freedom used by the model, several levels of the categorical variable can be assigned to multiple dummy variables
- (D) You can calibrate the model performance with an independent validation sample — *Correct answer*

Week4 Quiz

Question 3

10 Points

Which of the following statements best describes the optimal approach to model selection?

Choose at least one correct answer

- (A) The simplest model with the best performance on the training data.
- (B) The simplest model with the best performance on the validation data. *Correct answer*
- (C) The most complex model with the best performance on the training data.
- (D) The most complex model with the best performance on the validation data.

Week4 Quiz

Question 4

10 Points

Which of the following statements is true in relation to odds?

Choose at least one correct answer

- (A) The odds ratio shows the strength of the association between the predictor variable and the response variable — *Correct answer*
- (B) If the odds ratio is 1, then there is no association between the predictor variable and the response variable — *Correct answer*
- (C) If the odds ratio is greater than 1, then the group in the denominator has higher odds of having the event
- (D) The odds ratio represents the multiplicative effect of each input variable — *Correct answer*

Week4 Quiz

Question 5

10 Points

In the following logistic regression model

Logit(p) = - 0.7567 + 0.4373 * gender

What is the estimated odds ratio (females to males)?

Choose at least one correct answer

- (A) 1.55 — *Correct answer*
- (B) 3.42
- (C) 1.12
- (D) 0.39

Calculation odds ratio = $(e^{-.7567+.4373})/(e^{-.7567}) = 1.55$

Basic concept of Tree:

Main Points:

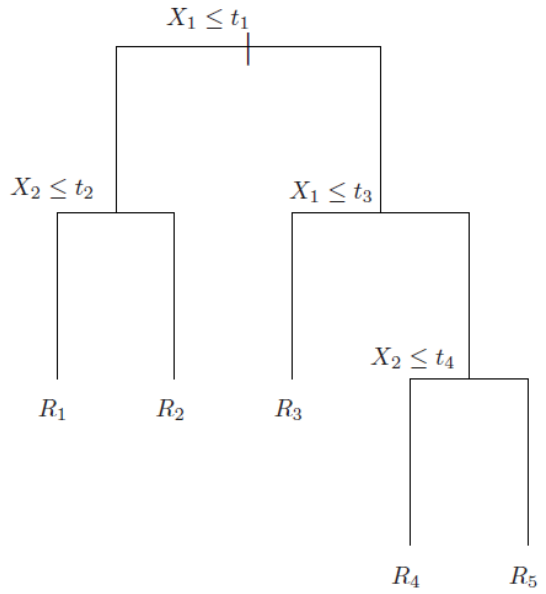
- Tree-based methods for regression and classification partition the predictor space (all possible x values) into a set of regions (rectangles), and fit a very simple model in each region.
- For example, in the regression case we can predict Y in each region as the average y value of the training points falling in the region.
- In classification, we can predict Y in each region as the most frequent class of the training points falling in the region.
- The partitions are based on successive splits of the predictor space (x values) into two non - overlapping subsets, enabling an interpretable visualisation of the prediction rule as a decision tree.
- We begin at the top of the tree with all x belonging to a single region. Each split is done by separating the x with smaller values of one of the predictors (say $X1$) from the x with larger values of the same predictor. The split is indicated via two new branches further down the tree. Best split is The highest logworth.

· The next few slides illustrate a tree constructed for two predictors, $X1$ and $X2$.

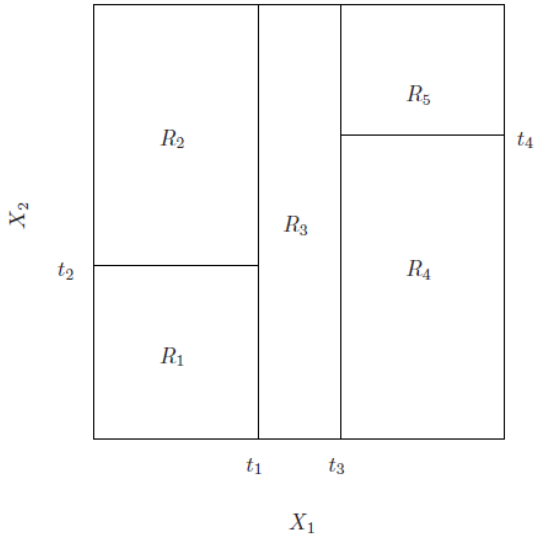
Basic concept of Tree:

- regions $R1, R2, \ldots, R5$ are known as **terminal nodes** or *leaves* of the tree

Example



Left: a tree



Right: corresponding partition

Basic concept of Tree:

Tree size

How large should we grow the tree?

- A large tree will clearly overfit the data, with the small number of observations in each region leading to high variance estimators of the constants *cm*.
- A small tree may fail to capture important structure in the data.
- The tree size is therefore a tuning parameter governing the complexity of the model, and we should select it by using the data. Which mean growing tree interactively.

Tree pruning

We adopt the following strategy to select the tree size:

1. Grow a large tree, stopping the splitting process only when it reaches the minimum node size (i.e. when splitting the nodes further would result in too few observations in the new regions).
2. Use **cost-complexity pruning** to reduce the tree back to an optimal size.
3. pruning adjust the complexity to avoid the overfitting

Basic concept of Tree:

Classification tree

- A **classification tree** is very similar to a regression tree, except that we predict a categorical response instead of a numerical one.
- We split the predictor space in non-overlapping regions $R1, R2, \ldots, RM$, and estimate the class probabilities within each region. We then classify the observations according to the estimated probabilities.

Growing a classification tree

- We need an appropriate criterion for growing a classification tree by binary splitting (the role of the RSS in a regression tree).
- We use **node impurity measures** such as misclassification error, Gini index, and cross-entropy.

Basic concept of Tree:

Node impurity measures

- When evaluating a split we weight the node impurity measures for the two child nodes by the number of observations in each node. E.g. for misclassification error, we look at the total number of misclassifications.

- Gini index and cross entropy are better at encouraging pure nodes than the misclassification error, and, thus, are generally the preferred measures when growing the tree.

- Gini index and the cross entropy also have the advantage of being differentiable functions of the proportions, which makes them more amenable to numerical optimisation.

- We can use any of the measures (including the misclassification error) to prune the tree.

Basic concept of Tree:

Categorical predictors

- Trees can easily handle categorical predictors without creating dummy variables: a split on such a variable comes down to assigning some of the categories to one branch, and assigning the remaining categories to the other branch.

- Moreover, the case of ordinal categorial variables (i.e. those whose categories can be ordered) is simpler, because we can perform the usual binary splitting by separating the larger and the smaller categories.

- Thus, trees handle ordinal variables in a more natural way than OLS (where a numerical score is assigned to each category).

Basic concept of Tree:

Advantages of trees

- Trees are simple and are easy to explain.
- Trees lead to highly interpretable prediction rules.
- Trees can easily handle categorical predictors without the need to create dummy variables.
- Trees can approximate complex nonlinearities, including interactions.
- Trees are the basic component of powerful prediction methods such as random forests and boosting.

Disadvantages of trees

- Instability: due to their hierarchical nature, trees are inherently unstable and have high variance. Small changes in the data may lead to a very different sequence of splits, compromising the interpretation of the model.
- Lack of smoothness: trees lead to non-smooth prediction functions and decision boundaries. Especially in the regression setting, this can degrade the performance.
- Therefore, decision trees may have low predictive accuracy.

Bagging:

- Since single tree based method is week and having high variance, to get a higher accuracy from the tree based method ensembled method been introduced, where the idea is construct more than one tree from a single data set and combined the result.

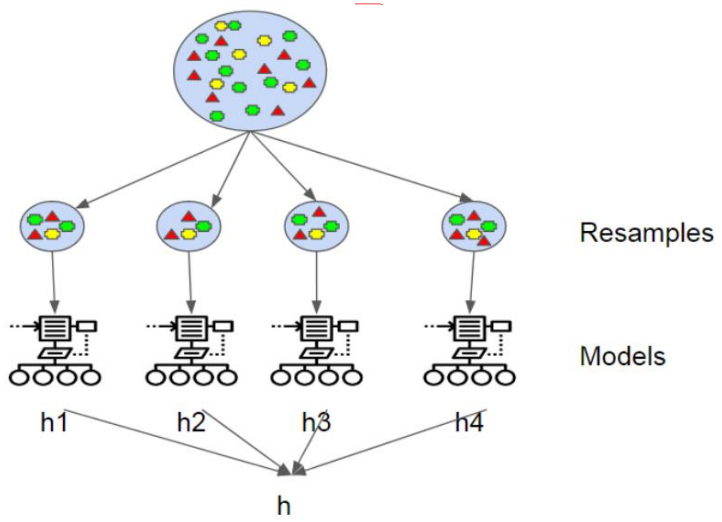
- How many trees:
  - Regression: Square of the error
  - Classification: Gini index, deviance

- Example:
  - ✓ Bagging
  - ✓ Boosting
  - ✓ Random forests

Bagging:

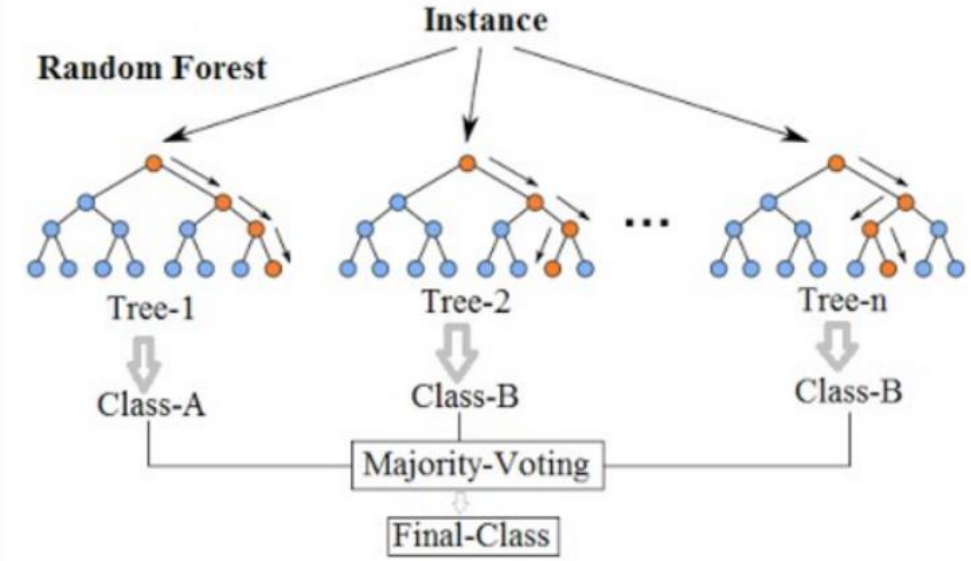
Bagging:

- this is also called as bootstrap aggregation.
- Step1: bootstrap random sample (n<=N) with replacement.
  say if there are 100 data points, you may take 60 from there with replacement.
- Step2: train the learning model separately say from the picture there should be 4 model, separately from h1,h2,h3,h4
- Step3: final prediction is the mean ()Regression or majority votes (classification).



Random Forests:

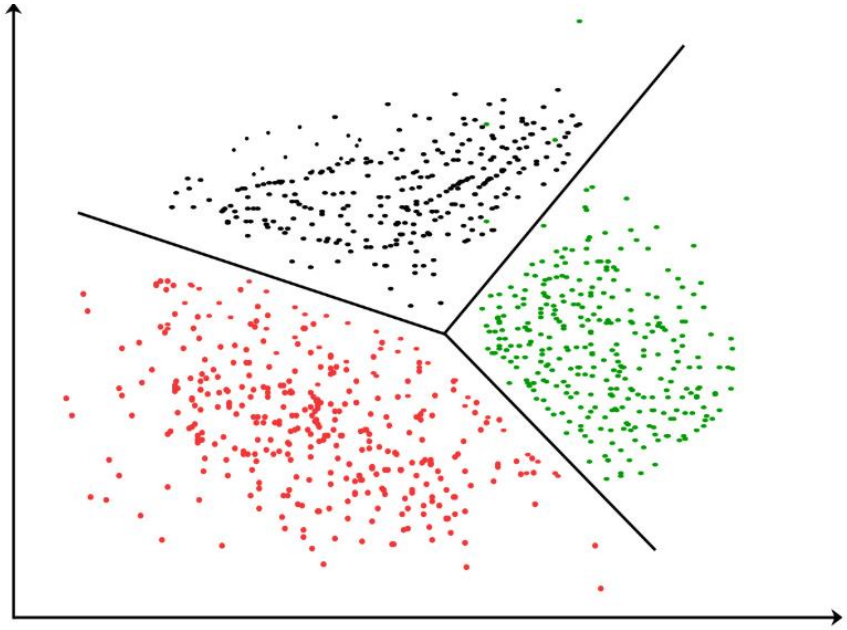
- Bagging uses full set of predictors, to determine the split.
- Random forest uses random number of predictors (subset) to determine the split. Which predictor also be the random.
- It use the optimal number of predictors.
- Sqrt(k) for the classification and k/3 for regression. K= total number of predictor. But it can be vary for getting better result.
- Ntree is 500 by default



Clustering:

What is clustering:

- Methods of grouping samples (*x*) that are ***similar*** – according to some pre-defined criteria.
- A form of ***unsupervised learning*** – no label information (*y*) is used to tell the algorithm which observations should be grouped together.
- It is often used for ***exploratory data analysis*** – a way of looking for patterns or structure in the data that are of interest.



Clustering:

Basic Principle of clustering:

Aim: to group observations that are "similar" based on predefined criteria.

Issues:
- Data types - counts, ratio, ordinal, categorical and continuous.
- Missing data
- Scaling
- (Dis)similarity metric (a critical step in clustering):
  - Euclidean, Manhattan, Pearson correlation, Spearman correlation etc.

Algorithm:
- Hierarchical clustering
- k-means clustering
- Advanced:
  - Fuzzy c-means clustering, Semi-supervised clustering, bi-clustering

Clustering:

Commonly used (dis)similarity measures:

- A metric is a measure of the similarity or dissimilarity between two data objects and it's used to form data points into clusters

- **Two main classes of distance:**

– Correlation coefficients (compares shape of expression curves)

– Distance metrics

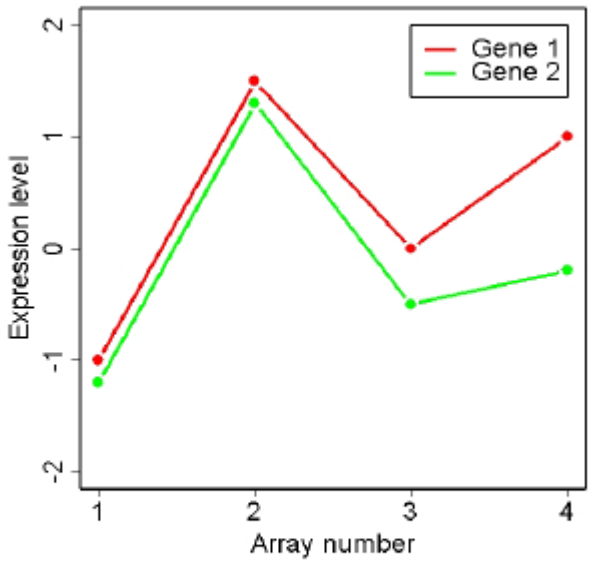
- City Block (Manhattan) distance: $d(X,Y) = \sum_i |x_i - y_i|$

- Euclidean distance: $d_{euc}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

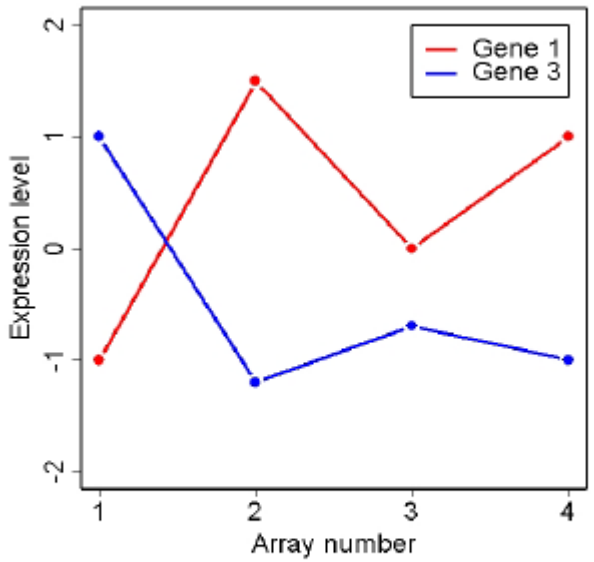
Clustering:

Correlation (a measure between -1 and 1)

- You can use absolute correlation to capture both positive and negative correlation



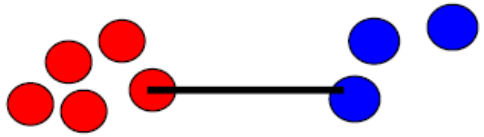
Positive correlation



Negative correlation

Clustering:

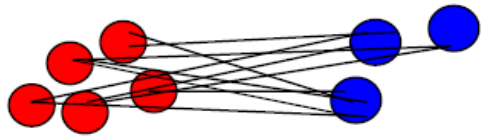
- Distance between clusters (between-cluster dissimilarity measures):



Single (minimum)

Complete (maximum)

Distance between centroids

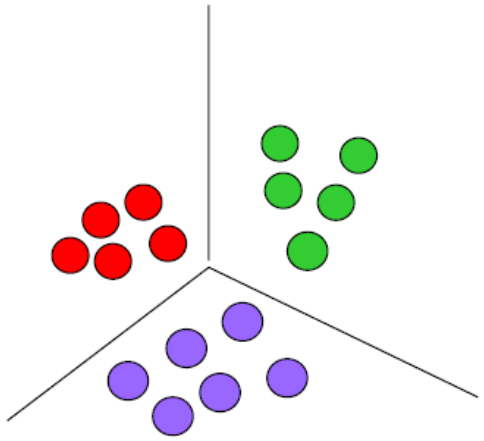


Average (Mean) linkage

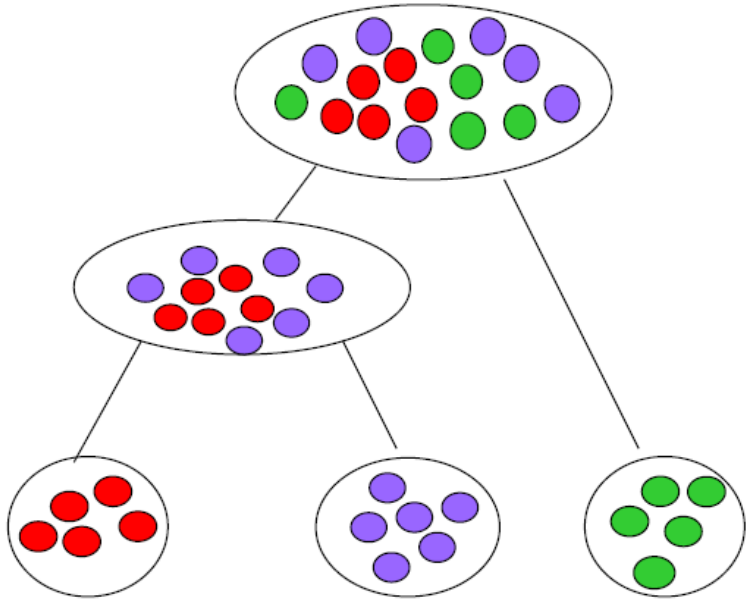
Clustering:

- Clustering algorithms:
  - ✓ Clustering algorithm comes in 2 basic flavors

Partitioning

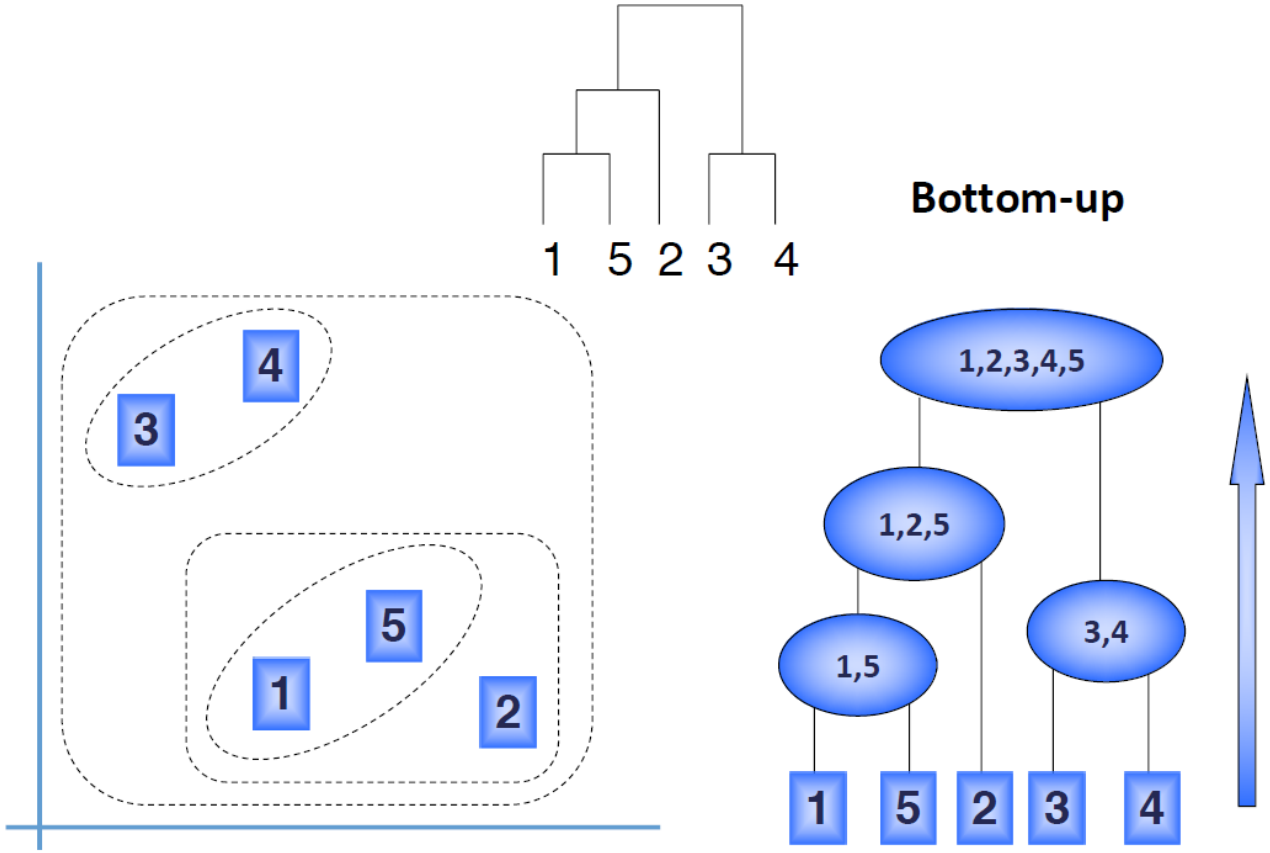


Hierarchical



hierarchical clustering:

- Hierarchical clustering methods produce a tree or dendrogram.
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.



hierarchical **Clustering:**

- **Bottom-up tree building procedure:**

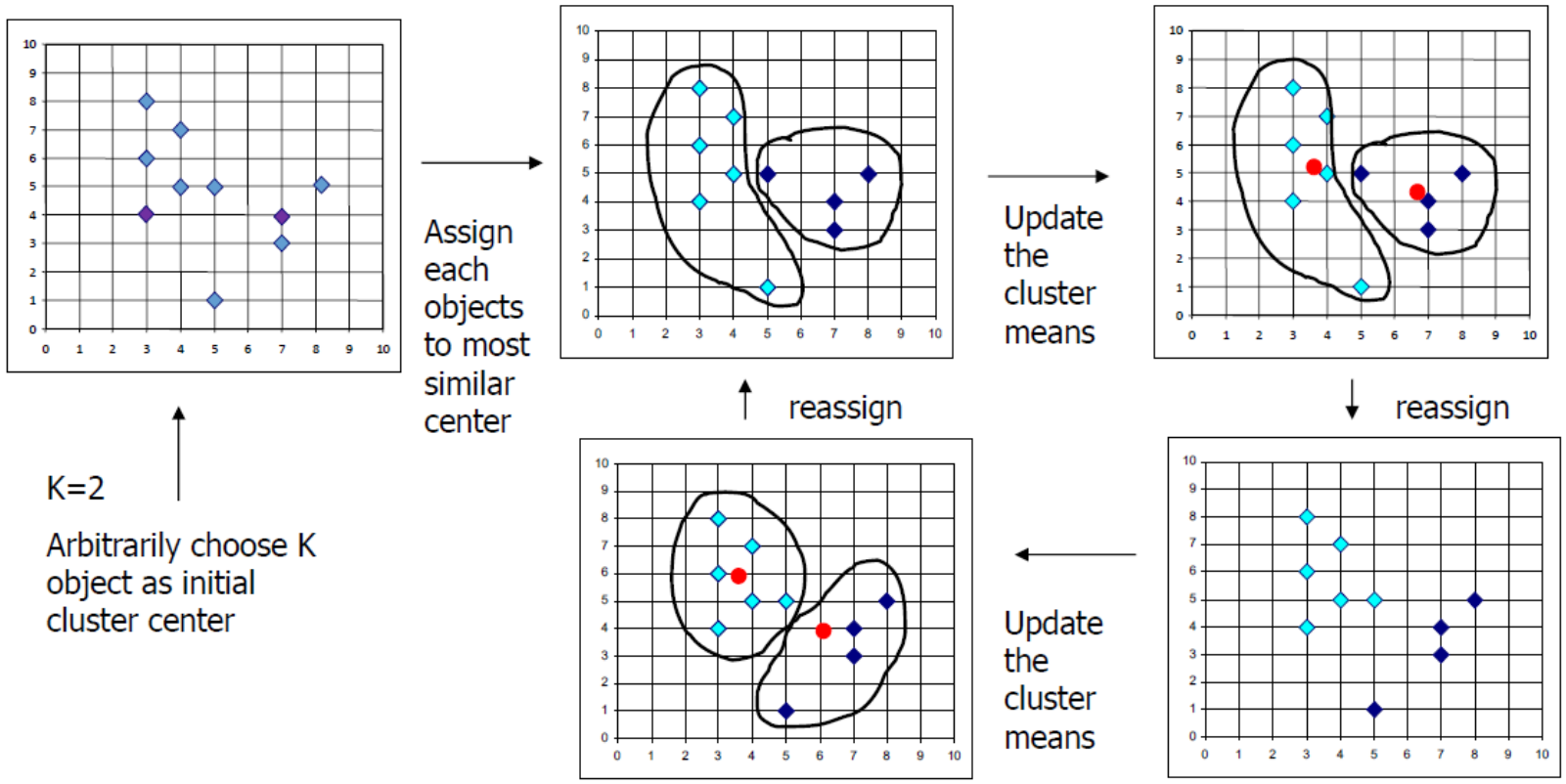
  - Start with n sample (or m feature) clusters
  - At each step, merge the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters
  - The distance between clusters is defined by the method used (e.g., if complete linkage, the distance is defined as the distance between furthest pair of points in the two clusters)

A typical k-means clustering algorithm:

- Arbitrarily choose k objects as the initial cluster centers

- Until no change, do
  - (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the Cluster

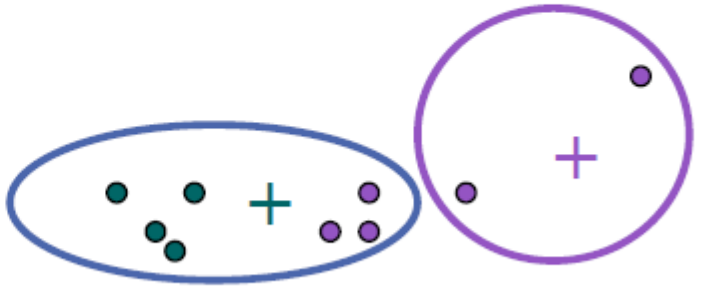
  - Update the cluster means, i.e., calculate the mean value of the objects for each cluster



Problems with kmean Clustering:

- Sensitive to outliers

objects with extremely large values May substantially distort the distribution of the data



Advantage of kmean Clustering:

- Easy to implement
- With huge number of variables k mean clustering is faster
- An instance can change the cluster while centroid is changed
- Train quickly

Disadvantage of kmean Clustering:

- Difficult to predict k value
- Order of data has an impact of final result
- Euclidian distance is not applicable in many cases
- High varience

Work book Exercise:

Question 1

How many rows of data does BIGORGANICS contain? **111,115**

Question 2

How many missing values are there for Organics Purchase Indicator? **0**

Question 3

How many customers purchased organic products? **27,525**

Question 4

Which gender buys more organic products? **Female (21,025)**

Work book Exercise:

Question 5

Which of the following statements most correctly describes the conclusion you can make about those people who buy organic products?

A. People who are younger and more affluent tend to buy more organic products. ✓
B. People who are older and more affluent tend to buy more organic products.
C. Less affluent people buy more organic products.
D. We do not have enough information to make any conclusions.

Question 6

How many observations were used to build the model? **82,040**

Question 7

Which variables are not considered important to the model?

- **Total Spend**
- **Geographic Region**
- **Neighborhood Cluster-7 Level**

Work book Exercise:

Question 8

What is the R square for the new model? **0.2258**

Question 9

What is the K-S statistic for the new model? **0.4706**

Question 10

Which model is selected when the fit statistic is Misclassification?

Logistic – with IM & VarSel

Question 11

Which model is selected when the prediction cutoff is set to 0.25 and the FDR statistic is chosen?

Logistic

JAMES COOK UNIVERSITY
AUSTRALIA

Thank You