BSc - Data Mining1

Topic 12 : Wrap Up

Part 01 : Overview

Dr Bernard Butler
Department of Computing and Mathematics, WIT.
(bbutler@wit.ie)

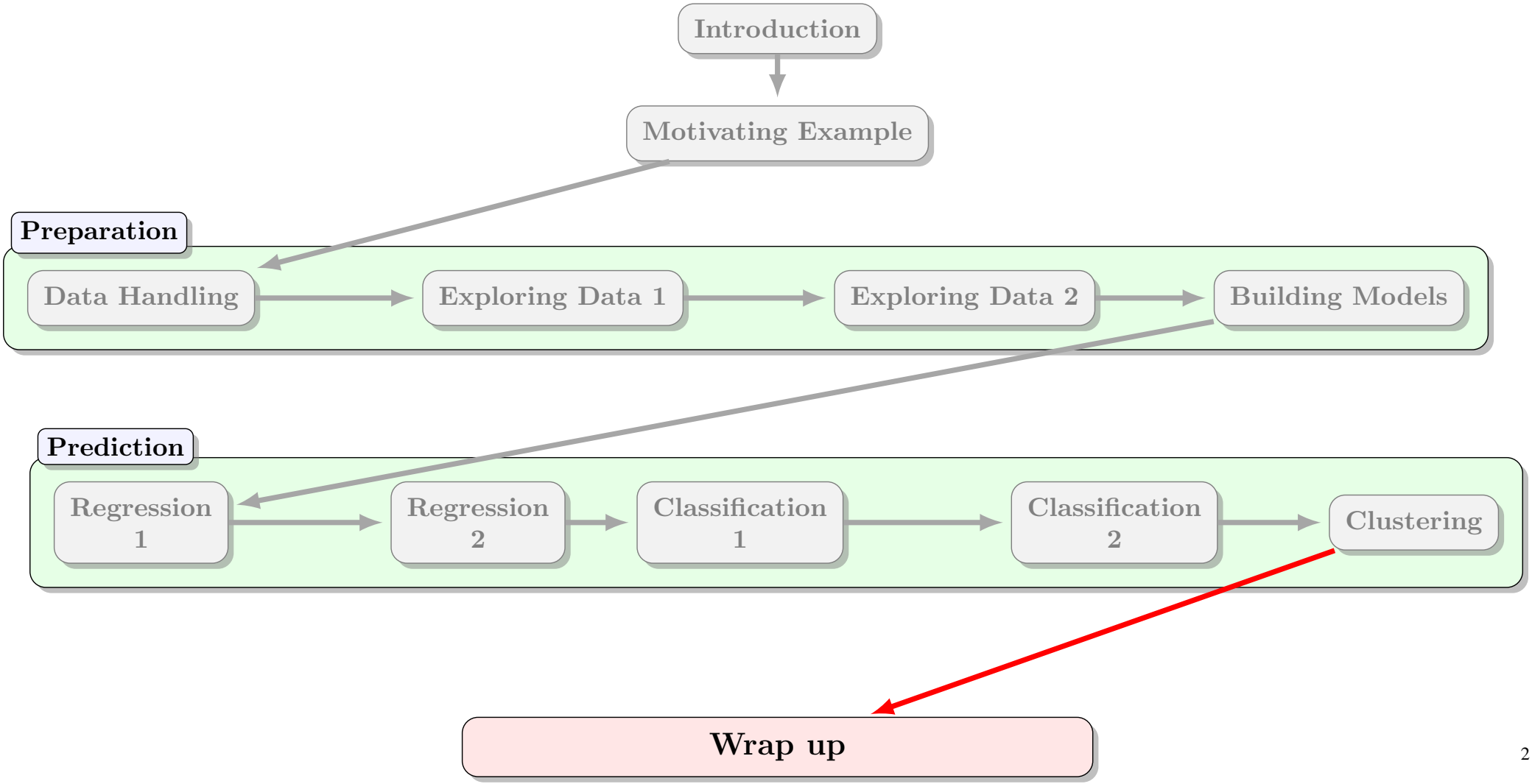
Autumn Semester, 2021

Outline

- Hierarchical clustering - continuing from last week
- Review of module
- Sample exam

Wrap up

Data Mining (Week 12)



Overview — Summary

This Week's Aim

This week's aims are to

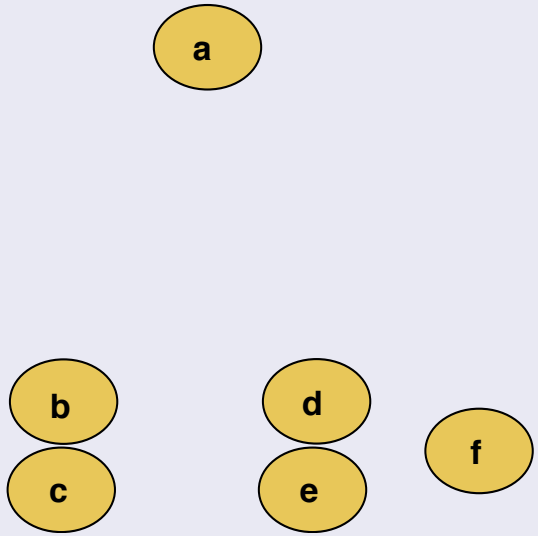
- Complete the coverage of clustering - consider hierarchical methods
- Review the material covered in the module, especially the algorithms
- Outline the structure and marking of the sample exam

Introduction

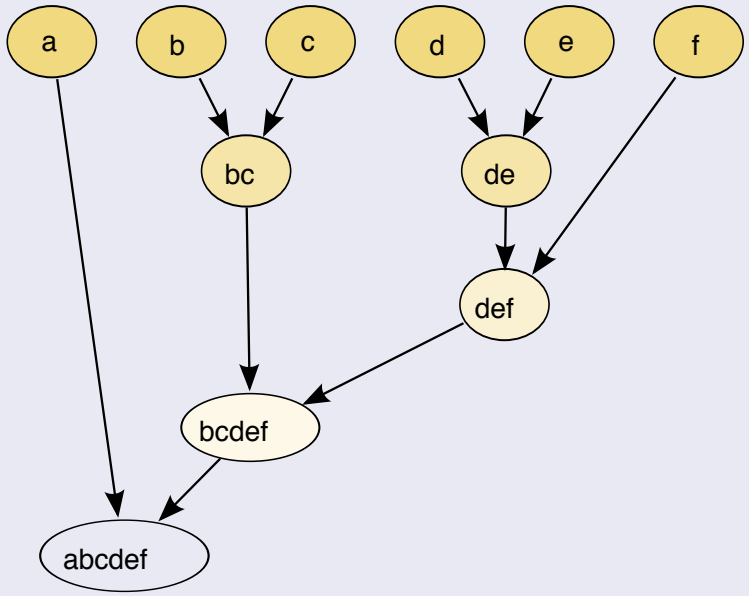
- Last week we saw clustering as a partitioning problem
  - Given data: assign each instance of that data to a single cluster, based on its *similarity* to other instances in that cluster
- Need to choose
  - Distance function and number of clusters for *centre-based* clustering (k-means etc.)
  - Density threshold - distances and/or minpts for *density-based* clustering (DBSCAN etc.)
- But what if we can delay making such hyperparameter choices until after we have visualised the structure in the data?
- Enter Hierarchical Clustering. . .

Simple example of data and its dendrogram

Data



Dendrogram



This illustrates how *agglomerative* clustering works. *Divisive* clustering works in the opposite direction, but the (flipped) dendrogram is the same in this case.
Sometimes, the vertical separation between Level i and $i + 1$ of the tree indicates the distance between clusters that are merged at Level $i + 1$.
Diagram Source: wikipedia

Dendrograms

Definition 1 (dendrogram)

The dendrogram shows clusters and their subclusters, in the form of a tree. The root cluster contains all elements; each leaf contains a single element. Clusters are merged in order of their similarity.

The dendrogram makes the internal similarity structure of the data more visible.
Sometimes, the vertical separation between Level i and $i + 1$ of the tree is proportional to the distance between clusters that are merged at Level $i + 1$.
An alternative representation is to display the data as a point cloud and to overlay nested clusters over the data.

Overview of Hierarchical Clustering Algorithm

Method (Agglomerative hierarchical clustering (AGNES))

Initialise the Cluster set $C = \{\boldsymbol{x}_i\}, i = 1, \ldots, n$;
$q \leftarrow |\{c_i\}| = n$;
Compute the $n \times n$ proximity matrix D where $D_{ij} = d(c_i, c_j)$;
repeat
  Find i, j associated with $\min_{i,j} D$, where i, j are indices of clusters that are nearest each other;
  Create the merged cluster c_i' containing the elements of cluster c_i and c_j;
  Record the merge operation so the dendrogram data structure can be built;
  Drop the old c_j cluster since it is not needed any more;
  Delete row $D(j, :)$ and column $D(:, j)$ from D
  $q \leftarrow q - 1$;
  Update row $D(i, :)$ and column $D(:, i)$ to compute distance between new cluster c_i' and remaining $q - 2$ clusters;
until $q = 1$ and hence only one cluster remains;

As can be seen, this is a deterministic search algorithm.
However, there is scope for different definitions of the distance function $d(c_i, c_j)$ betwen clusters c_i and c_j.

Distance between clusters: linkage

Earlier, we looked at different ways of computing the *distance between two points*.
For hierarchical clustering, we need to compute the *distance between two clusters*.

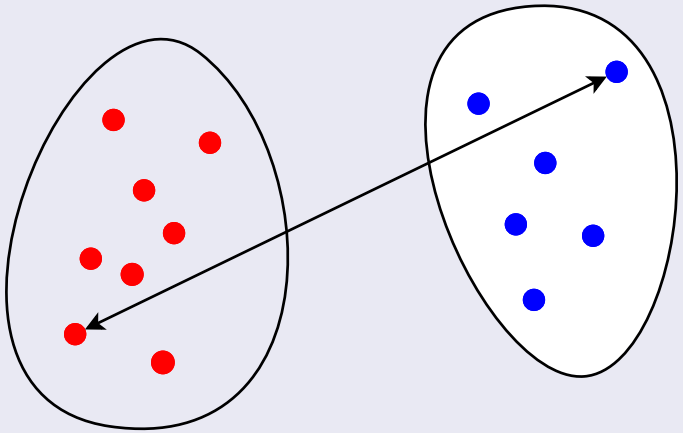
Definition 2 (Linkage function)

For Complete Linkage: $D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$.
For Single Linkage: $D(X, Y) = \min_{x \in X, y \in Y} d(x, y)$.
For Average Linkage: $D(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y)$. This is also known as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) linkage.
For Ward linkage: the initial (single-point) cluster distances are simply the Euclidean distances between the points. The clusters are merged based on a minimum variance criterion. The distance between any point and a merged cluster is calculated using a recursive formula of Lance-Williams type.

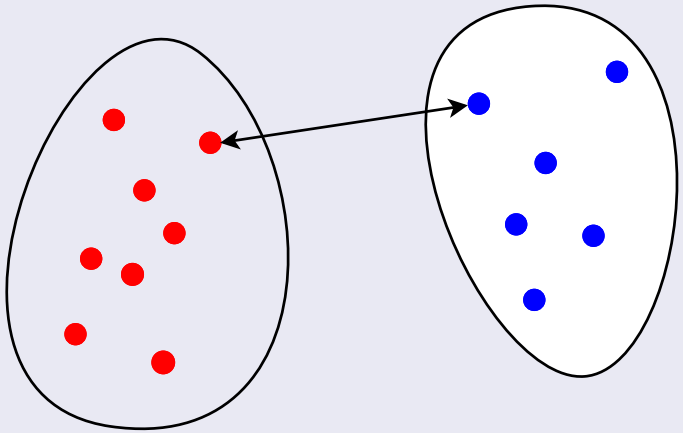
Generally, Complete Linkage and Ward's minimum variance linkage give the most balanced and useful clusters.

Distance between clusters: linkage visualisation

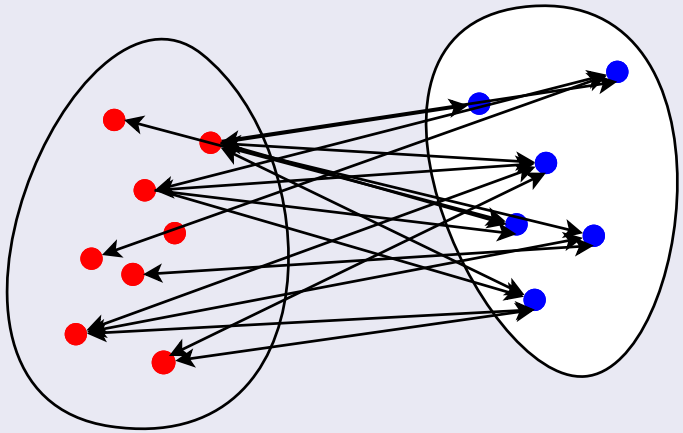
Complete Linkage



Single Linkage



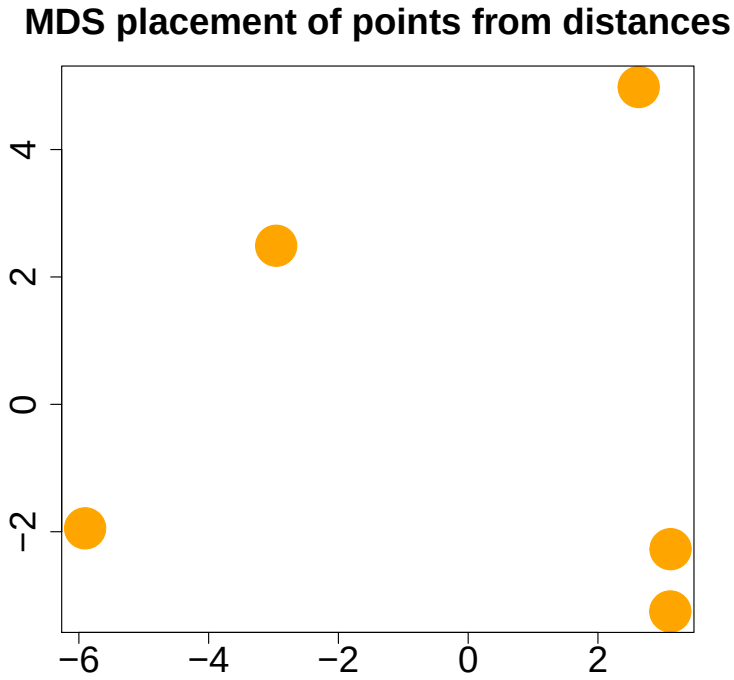
Average Linkage



AGNES worked example: setting the scene

Distance Matrix, Step 0

	A	B	C	D	E
A	0				
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0



Use of distance matrices

Many algorithms in data mining either start from a distance matrix representation, or need to create it themselves. Reversing the process(from distances to locations) is not unique, but MultiDimensional Scaling often gives an attractive placement (as above, centred on origin).

AGNES worked example: Initial iterations

First clustering: CE, A, B, D

The smallest distance is 2 between C-E. We cluster these points and compute the distance of the remaining points from the CE cluster. Because of single linkage, we store the minimum such distance in the revised table beside.

Distance Matrix, Step 1

	CE	A	B	D
CE	0			
A	3	0		
B	7	9	0	
D	8	6	5	0

Second clustering: ACE, B, D

The smallest distance is 3 between A and CE. We cluster these points and compute the distance of the remaining points from the ACE cluster. For example d(B,A) = 9, d(B,C) = 7, d(B,E) = 10, so by single linkage d(B,ACE) = 7 as in the revised table beside.

Distance Matrix, Step 2

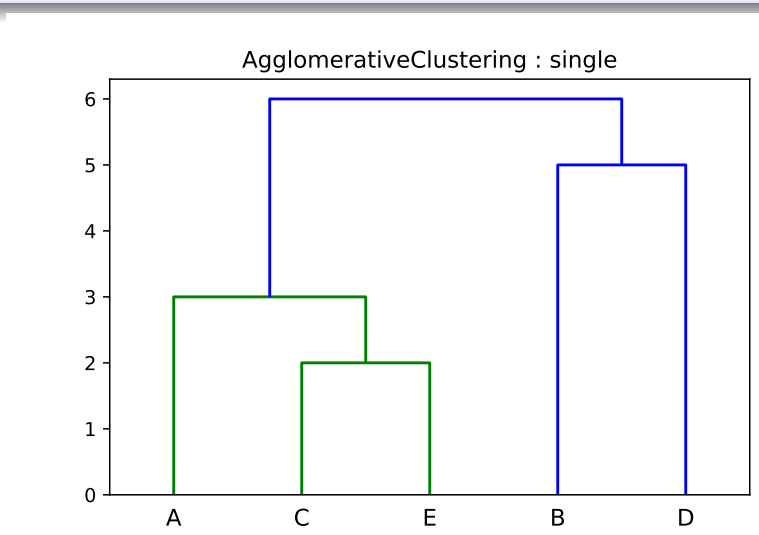
	ACE	B	D
ACE	0		
B	7	0	
D	6	5	0

Minimum distances so far: 2,3

AGNES worked example: Final iterations

Third clustering: CE, A, B, D

The smallest distance is 5, between B and D, so we create a BD cluster. The new distance matrix is shown alongside. The next step after this would be to merge ACE with BD, creating a single ABCDE cluster. The algorithm ends...



Distance Matrix, Step 3

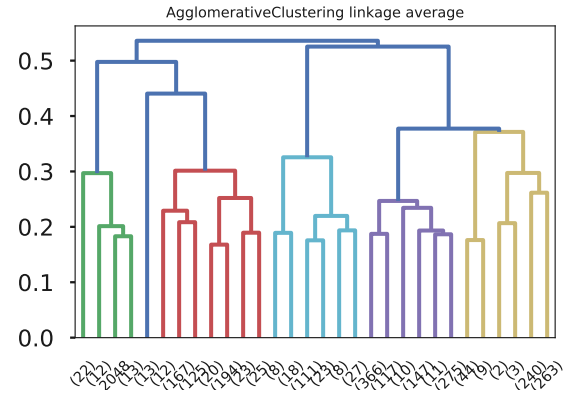
	ACE	BD
ACE	0	
BD	6	0

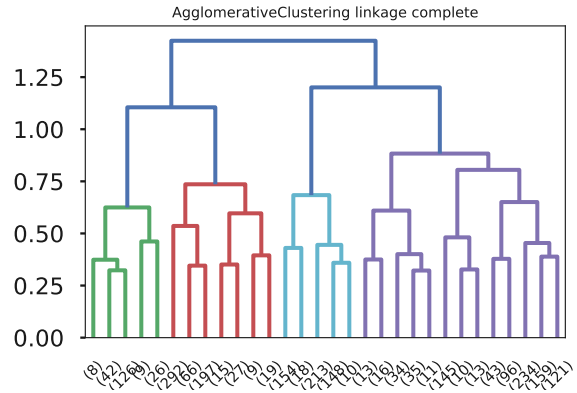
Minimum distances so far: 2,3,5,(6)

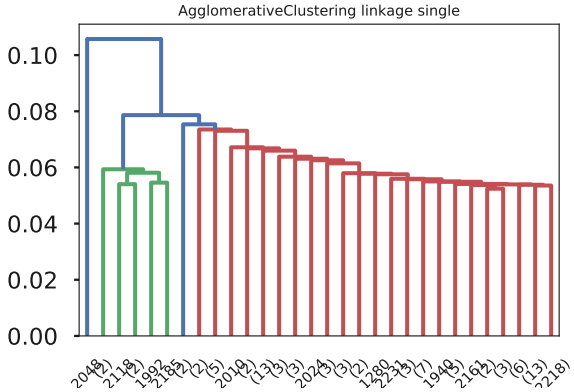
Resulting Dendrogram

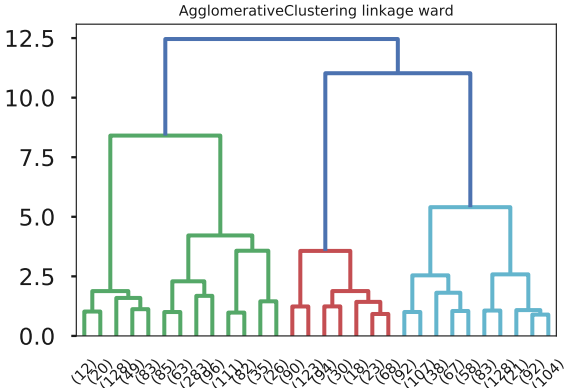
The resulting dendrogram summarises the hierarchical clustering. Note that the joins/splits occur at distances 2, 3, 5 and 6, as noted above.

Comparison of Dendrograms









As can be seen, the choice of linkage function has a dramatic effect on cluster membership.
The underlying data in this instance appeared to have 6 clusters. Can you see this in these dendrograms?

Uses of hierarchical clustering

- Hierarchical clustering can be very helpful for looking at data at a variety of scales, and hence for seeing hierarchical structure in a data set. This can lead to insights that other techniques, which focus on finding a single cluster mapping, cannot offer.
- Hierarchical clustering offers a rich variety of objective functions (primarily relating to linkage), some of which might suit a specific scenario.
- It can be used as a means of estimating parameters for other, perhaps more focused techniques, e.g., to estimate the number of clusters/components in the data.
- Since hierarchical clustering provides more than one candidate clustering, it can be require more system resources (computation and memory) than other techniques. Thus it might not scale very well.
- We have seen agglomerative (bottom-up) clustering. Divisive (top-down) clustering (DIANA) is also available, but has worse scalability.

Introduction

What are the most important take-aways from this module

01-Data Mining History and Process

- Need to cut through hype and commentary by non-experts and those with (commercial) agendas
- Instead, focus on the key concepts and definitions of *big data*, *machine learning*, . . .
- Data mining is the overarching *process*; what are its models and procedures?
- Understand how ICT advances enabled new applications, requiring new machine learning techniques, which enable. . .
- But is this a virtuous cycle? What about societal effects of unethical data mining?
- With the growing maturity of deep learning - how can we trust the ultimate “black box” of deep learning?
- In the lab we considered how unspecialised tools can be used for data analytics

02-Pandas and a simple classification example

- Pandas is the workhorse of data mining tools in python
- Used for data import/export, managing dataframes (naming, adding columns,. . .) and series
- More complex operations (filtering, aggregating, sorting) are also possible
- Used heavily and assessed as such in the CA programming assignments!
- Classification is one of the classic machine learning tasks: predicting a label, given data
- Introduced k nearest neighbours as a simple algorithm, based on voting, to identify the most likely label.
- What are its strengths and weaknesses? When would it be used?

03-Data handling: Storage and Computation

- Storage is used both for persistence and archiving - which is better for data mining?
- There are many data models and management systems - when to use each
- CAP theorem and its implications
- Big data requires a new approach, supporting out-of-memory computation
- Characteristics of Hadoop and Spark - how they work and how they differ
- Computational models and worked examples

04-Exploratory Data Analysis 1 and 2

- EDA requires time and care
- A 3-Phase process was described and students have used this in CA1 and CA2

06-Data Modelling

- This is the central focus of the module!
- Explain what a linear model is and how it can be extended (gneerally with mnore complex features)
- Training vs test data
- Objectives of modelling: explain a domain vs predict a result
- Components of error: bias, variance
- How can we control errors? How do we refine models and when do we stop?

07-Regression1

- What is regression and what types of features, targets are needed?
- What are the assumptions and what happens if they do not hold?
- How do we judge the success of the model?
- Distinguish between and statistical and machine learning metrics

08-Regression2

- What options do we have if vanilla regression models are not sufficient?
- Use of multivariate approaches to improve models
  - Regularisation: ridge regression to down-weight some features; lasso to drop them entirely
  - Dimensionality reduction: find a more economical subset of the features
  - What are the advantages and disadvantages?
- Role of correlation: between features and between a feature and the target

09-Classification1

- How it differs from regression
- conversely, how a variable tranformation can enable logistic regression to be used for classification
- Confusion matrices (true/false positives/negatives) and the derived ratios
- When to use a metric and how to interpret it in practice

10-Classification2

- Use of probability based classification techniques
- Naive Bayes - its derivation and worked examples
- Entropy in data mining and how it leads to the decision tree method
- Algorithm and worked examples
- Choosing classification technique for a given problem

10-Classification2

- How clustering differs from classification
- Partitional vs hierarchical clustering
- Role of distance metrics, their definition and calculation
- EM algorithms - how they work
- Derivation of K-means and how it can be extended if needed
- How GMM relates to k-means
- Motivation for density-based approaches and their pros and cons
- Derivation of the DBSCAN algorithm and how to tune it
- Role of different linkages and interpretation of dendrograms

Overview of Sample Exam

ID	Section	Type	Marks
Q1	A	10 True/False	20
Q2	B	Parts/Essay	10
Q3	B	Parts/Essay	10
Q4	B	Essay/Essay	10
Q5	C	Parts/Worked	25
Q6	C	Parts/Worked	25

- All questions in Sections A, B and C are mandatory.
- There may be choice within a question, e.g., between essay topics

Sample Exam - section A

- True/False questions can come from any part of the module
- They test detailed knowledge: read the statement carefully!
- Be careful of “some” versus “all” statements
- Validate your answer: try to come up with a counterexample, etc
- Note that negative marking applies, so wild guesses are not advised...
- Questions can be drawn from anywhere in the module
- The moodle quizzes are indicative of the style and difficulty, but exam versions are limited to True/False

Sample Exam - section B

- There are three, mostly multipart, questions in this section, each worth 10 marks
- All questions should be attempted (worth 30 marks)
- You may be asked to derive one or more formulas; you will also generally be asked where the formula is used
- you may also be asked to interpret analysis results, e.g., to say how successful it was
- you may also be asked a more essay-style question. For example, it could be a compare-and-contrast question, so you should describe 5 major points of similarity or difference, to obtain the full 10 marks.
- you may also be asked quite specific questions arising from a more general concept
- General comments about reading the question, time management, etc., also apply here!

Sample Exam - section C

- There are two multipart questions in this section, each worth 25 marks
- The early parts of each question are structured like the multi-part questions in Section B
- You may be expected to use one of the algorithms we covered in class to compute some quantity. The data sets will be small to make this possible in exam conditions. You should bring a scientific calculator with you.
- You may be presented with some results and asked to interpret/validate them.
- the questions in Section C are designed to test whether you can “do” data mining, not just remember facts, so they have a more applied feel than those in Sections A and B
- Time management is important here: Section C is worth half the marks, but you could get bogged down. Always write down anything you learned that is relevant.
- Questions may be drawn from anywhere in the module.