

FIT3152 Data analytics – Lecture 12

Revision and Exam Preparation

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
28/2/22	1	Intro to Data Science, review of basic statistics using R	...		
7/3/22	2	Exploring data using graphics in R	T1		
14/3/22	3	Data manipulation in R	T2	Released	
21/3/22	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
28/3/22	5	Network analysis	T4		
4/4/22	6	Regression modelling	T5		
11/4/22	7	Classification using decision trees	T6		
		Mid-semester Break		Submitted	
25/4/22	8	Naïve Bayes, evaluating classifiers	T7		Released
2/5/22	9	Ensemble methods, artificial neural networks	T8		
9/5/22	10	Clustering	T9		
16/5/22	11	Text analysis	T10		Submitted
23/5/22	12	Review of course, Exam preparation	T11		

SETU

Student Evaluation of Teaching and Units (SETU) has opened for Semester 1.

- All students are encouraged to participate. Your feedback is very important.
- You will see a block in Moodle linking you to the survey.

End of semester exam

The end of semester exam:

- Will be online. The university is yet to make a formal announcement, but I expect on campus students will sit their exam at Monash. The university will advise you of the arrangements for sitting the exam.
- The exam is closed book. You are allowed two sheets of blank A4 paper for working.
- You may use a calculator: graphing, scientific, or CAS.
- A mock (practice) exam has been setup. It is available from a link under the Assessments tile in Moodle. It is a good indicator of length and complexity.
- Solutions will be released at the beginning of SWOT VAC.

The eExam

- 11 Question groups in the final exam.
- Format of the eExam makes it seem like there are more questions than a paper exam (but there are not).
- You will write answers in spaces provided.
- Attempt all questions in any order you like.
- You can commence writing at the start of the exam but read through the whole exam first.
- You do not have to finish the paper to pass.
- *Do the easy questions first and get them right!*
- *Do not email/post that exam was too long. It wasn't!*

eExam sample: formulas & navigation

FIT3152 - Mock Exam 2

Dashboard / My courses / FIT3152 - Mock Exam 2 / Preview

Exam navigation Questions attempted: 0/52

R Coding
1 2 2a 2b 2c
2d 2e 2f

Regression
i 3 3a 3b 3c
4 4a 4b 4c

Networks
6a 6b 6c 7

Naïve Bayes'
8 9

Visualisation
10

Decision Trees
11 11a 11b 11c 11d
12a 12b 12c 12d

ROC and Lift
13 13a 13b 14 15
15a 15b

Clustering
16 16a 16b 17 18

A Tour Through the Visualization Zoo – Summary of Graphic Types

Time-Series Data

- Index Charts
- Stacked Graphs
- Small Multiples
- Horizon Graphs

Statistical Distributions

- Stem-and-Leaf Plots
- Q-Q Plots
- SPLOM
- Parallel Coordinates

Maps

- Flow Maps
- Choropleth Maps
- Graduated Symbol Maps
- Cartograms

Hierarchies

- Node-Link diagrams
- Adjacency Diagrams
- Enclosure Diagrams

Networks

- Force-Directed Layouts
- Arc Diagrams
- Matrix Views

Entropy

If S is an arbitrary collection of examples with a binary class attribute, then:

$$\text{Entropy}(S) = -P_{C1} \log_2(P_{C1}) - P_{C2} \log_2(P_{C2})$$
$$= -\frac{N_{C1}}{N} \log_2\left(\frac{N_{C1}}{N}\right) - \frac{N_{C2}}{N} \log_2\left(\frac{N_{C2}}{N}\right)$$

where C_1 and C_2 are the two classes. P_{C1} and P_{C2} are the probability of being in Class 1 or Class 2. Specifically, N_{C1} and N_{C2} are the number of examples in each class. N is the total number of examples.

Note: $\log_2 x = \frac{\log_{10} x}{\log_{10} 2} = \frac{\log_{10} x}{0.301}$

Information gain

The $\text{Gain}(S, A)$ of an attribute A relative to a collection of examples, S, with v groups having $|S_v|$ elements is:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} * \text{Entropy}(S_v)$$

Formulas are information pages

Side panel shows groups and individual question sections

eExam sample: question

FIT3152 - Mock Exam 2

Dashboard / My courses / FIT3152 - Mock Exam 2 / Preview

Exam navigation

Questions attempted: 0/52

R Coding

i i
1 2 2a 2b 2c
2d 2e 2f

Regression

i 3 3a 3b 3c
4 4a 4b 4c

Networks

5 5a 5b 5c 6
6a 6b 6c 7

Naïve Bayes'

8 9

Visualisation

10

Decision Trees

11 11a 11b 11c 12
12a 12b 12c 12d

8 of 27

Edit question

Use data below and Naïve Bayes classification to predict whether the following test instance will be happy or not.

ID	AgeRange	Occupation	Gender	Happy
1	Young	Tutor	F	Yes
2	Middle-aged	Professor	F	No
3	Old	Tutor	M	Yes
4	Middle-aged	professor	M	Yes
5	Old	Tutor	F	Yes
6	Young	Lecturer	M	No
7	Middle-aged	lecturer	F	No
8	Old	Tutor	F	No

Test instance: (AgeRange = young, Occupation = professor, Gender = F, Happy = ?)

A \neg B I U x_2 x^2 \equiv \equiv \equiv \equiv \equiv ABC 123 C C

Space to type your answers

eExam sample: summary page

Filter by

52 All

0 Attempted

0 Incomplete

52 Not Attempted

0 Unsure

FIT3152 - Mock Exam 2

Dashboard / My courses / FIT3152 - Mock Exam 2 / Preview / Summary of attempt

Questions attempted: 0/52

No	Question	Status	Marks
1	The DunHumby (DH) data frame records the DATE a ...	Not Attempted	4
2a	DHY = DH[AS.DATE(DH\$VISIT_DATE,"%D-%M-%Y") < ...	Not Attempted	1
2b	CUSTSPEND = AS.TABLE(BY(DHY\$VISIT_SPEND, DH...	Not Attempted	1
2c	CUSTSPEND = SORT(CUSTSPEND, DECREASING = T... CUSTSPEND = RANK(CUSTSPEND, DECREASING = T...)	Not Attempted	1
2d	CUSTSPEND = RANK(CUSTSPEND, DECREASING = T... CUSTSPEND = RANK(CUSTSPEND, DECREASING = T...)	Not Attempted	1
2e	RANK(CUSTSPEND, DECREASING = T...) IN CUSTPEN...	Not Attempted	1
2f	... + FACET_WRAP(~ CUSTOMER_ID, NROW = 3)	Not Attempted	1
3a	Write down the regression equation predicting log(pri...	Not Attempted	1
3b	Explain the different data types present in the variab...	Not Attempted	2
3c	What is the predicted price for a diamond of 1 carat ...	Not Attempted	1
4a	Which colour diamonds can be reliably assumed to h...	Not Attempted	2
4b	Which colour diamonds have the lowest value? How ...	Not Attempted	2

Use this page to see what you've not attempted yet.

Consultations

Current consultation schedule will be continued.

We will put multiple tutors on certain consultations during Swot Vac and the exam period depending on demand.

Lecture 1 – Introduction to Data Science

Recent examples:

- crime, food webs, global warming

Common themes:

- large data sets with complex interactions within the data. Importance of graphics for analysis and to display results

Review of basic statistics using R

Criminal investigation

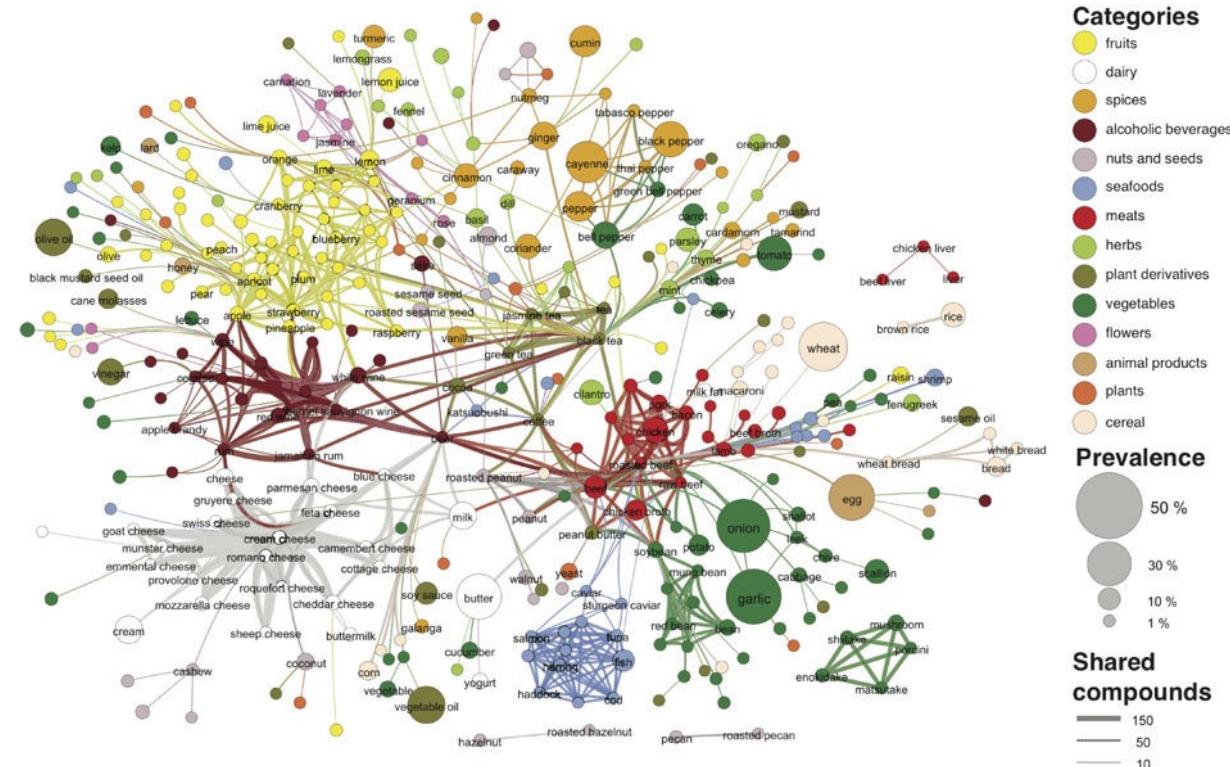
'Needle in a haystack' "We had to sift through a lot of information. The statements of the 100 people who were at the campsite, CCTV footage, data from phones..." he said.

Detective Supt Wilde said they were able to "build a picture of who was supposed to be there and who was not supposed to be there" by piecing together all the information they received and by "placing people in certain locations at certain times".

The Australian newspaper reported the breakthrough came when police traced a mobile phone number to a phone tower near the campsite around the time of Cleo's abduction. This allegedly focused their attention on Mr Kelly.

Food networks

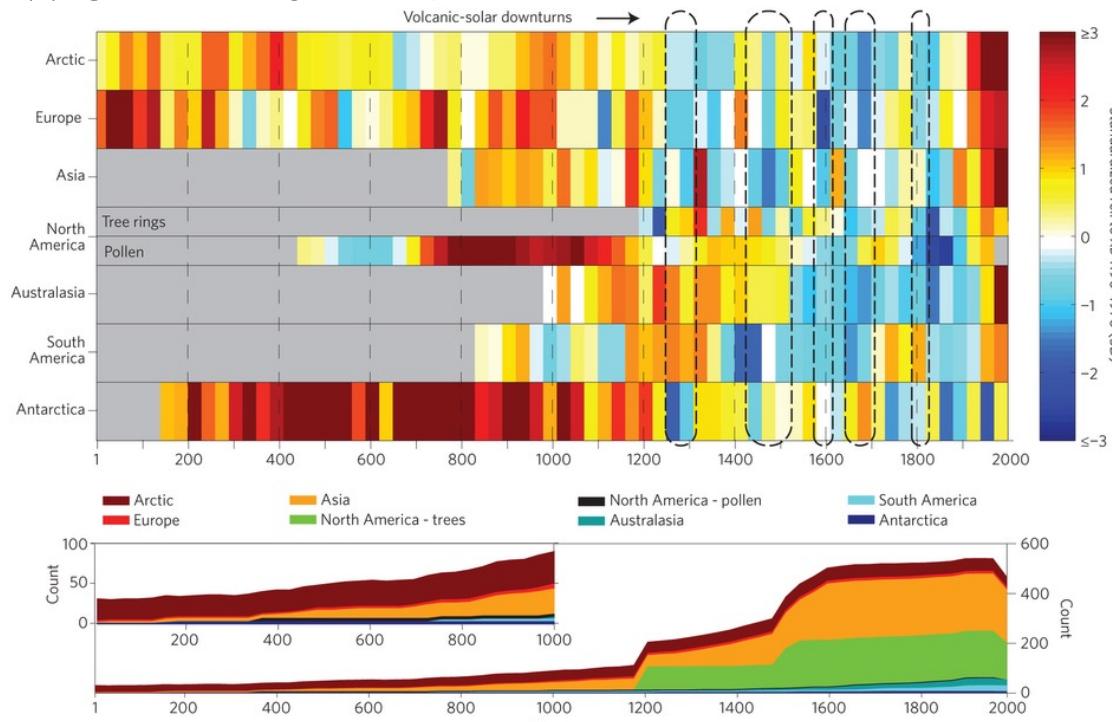
Flavor network and the principles of food pairing



<http://www.nature.com/srep/2011/111215/srep00196/full/srep00196.html>

Climate change

Continental-scale temperature variability during the past two millennia



<http://www.nature.com/ngeo/journal/v6/n5/full/ngeo1797.html>

Beautiful graphic design



<http://xiaoji-chen.com/blog/2011/sky-color-of-10-chinese-cities/>

Deep learning



Go, a complex game popular in Asia, has frustrated the efforts of artificial-intelligence researchers for decades.

ARTIFICIAL INTELLIGENCE

Google masters Go

Deep-learning software excels at complex ancient board game.

<https://www.nature.com/news/google-ai-algorithm-masters-ancient-game-of-go-1.19234>

Data science: some common themes

Previous examples illustrate:

- Complex problems, of societal concern.
- Large data sets, multiple data sets (mashups), messy, incomplete, heterogeneous, non-traditional, open data.
- Often using data repositories created for another purpose (food network): One description of Data Science is making a product out of data...
- Data collection and analysis on a scale that would have been unthinkable 15 years ago.
- Use of high-quality graphics for communicating results!

Data science: high-level skills

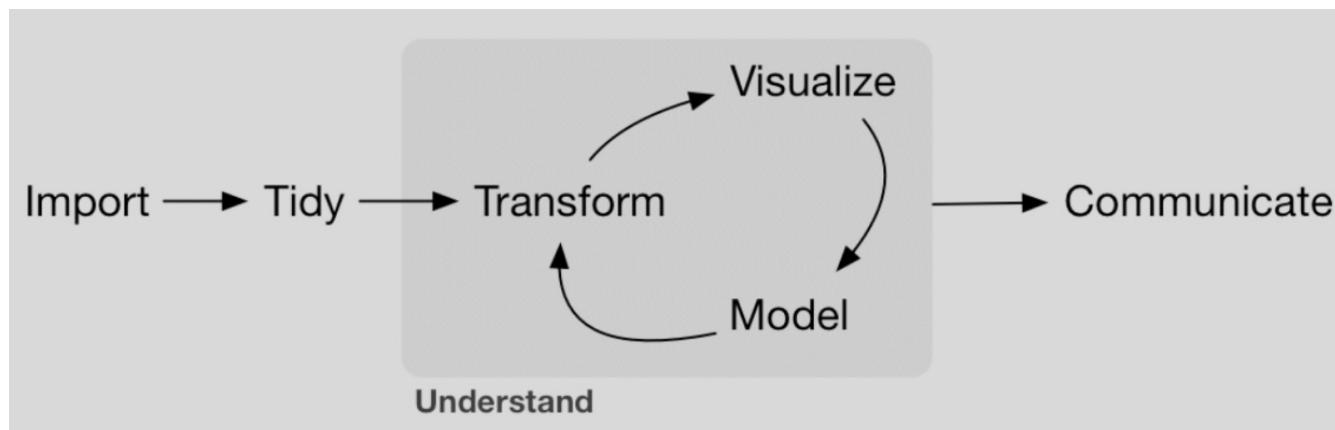
Some necessary skills for a data scientist:

- Understand a problem from client's perspective,
- Collect, cleanse, manage and combine data – which may come from disparate sources,
- Understand the data, most likely using visualization tools as a starting point,
- Analyze and model the data using statistical and (AI) machine learning techniques,
- Communicate the results simply and effectively.

The data science process

Generic methodologies for data analysis:

- For example, the data analysis process from Wickham and Gromelund:



<https://r4ds.had.co.nz/>

R

Rationale for using R, some background.

Obtaining and installing R and R studio,

Using R: syntax, data types, functions,

Data structures: data frames, arrays etc.,

Packages,

Help and References in R,

Review of basic statistics using R.

Lecture 2 – Visualization of data

Recent examples, common themes:

- Graphics present many dimensions using: colour, size, position, adjacency, connection, shape...

Visualization using R

- First steps: looking at the data (Iris data)
- Visualization for analysis – base graphics
- Presentation quality graphics – ggplot2, lattice

A tour through the visualization zoo

Summary of Graphic Types

- Time-Series Data
 - > Index Charts, Stacked Graphs, Small Multiples, Horizon Graphs
- Statistical Distributions
 - > Stem-and-Leaf Plots, Q-Q Plots, SPLOM, Parallel Coordinates
- Maps
 - > Flow Maps, Choropleth Maps, Graduated Symbol Maps, Cartograms
- Hierarchies
 - > Node-Link diagrams, Adjacency Diagrams, Enclosure Diagrams
- Networks
 - > Force-Directed Layouts, Arc Diagrams, Matrix Views

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

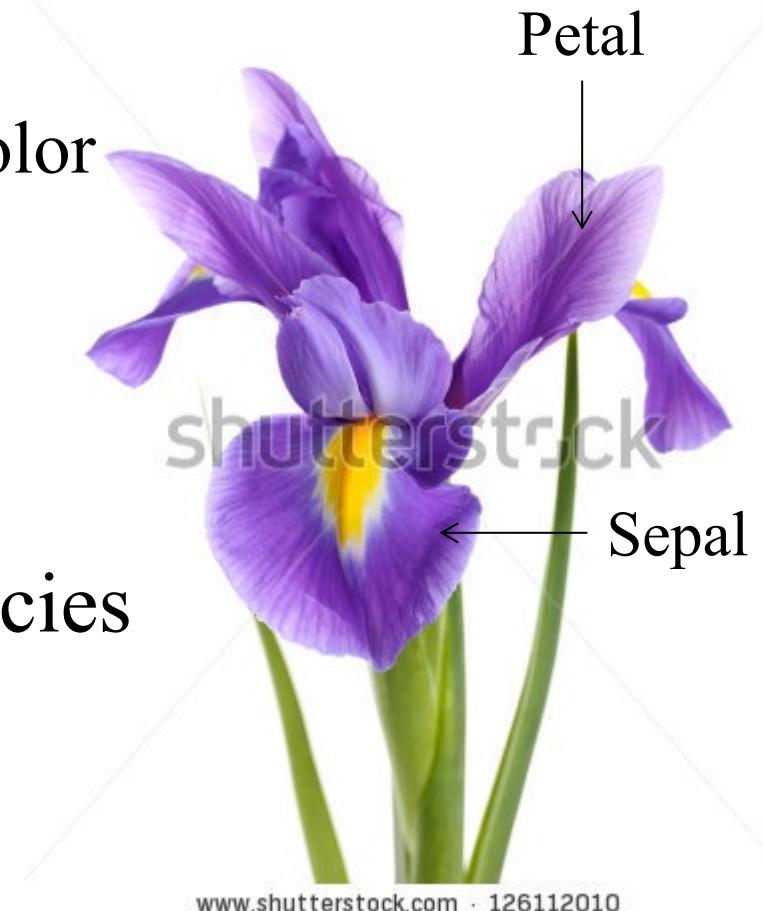
Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species
using physical measurements?

- Data is packaged with R: “iris”

http://en.wikipedia.org/wiki/Iris_flower_data_set



Print

```
> iris # = print(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
...					

Sample exam question:



6

Marks

A World Health study is examining how life expectancy varies between men and women in different countries and at different times in history. The table below shows a sample of the data that has been recorded. There are approximately 15,000 records in all.

Country	Year of Birth	Gender	Age at Death
Australia	1818	M	9
Afghanistan	1944	F	40
USA	1846	F	12
India	1926	F	6
China	1860	F	32
India	1868	M	54
Australia	1900	F	37
China	1875	F	75
England	1807	M	15
France	1933	M	52
Egypt	1836	M	19
USA	1906	M	58

Using one of the graphic types from the Visualization Zoo (see formulae and references for a list of types) suggest a suitable graphic to help the researcher display as many variables as clearly as possible.

Explain your decision. Which graph elements correspond to the variables you want to display?

Lecture 3 – Data Manipulation in R

Making tables and summaries

Working with factors

- E.g., to apply functions by groups (species in iris data)

Transforming data: analysis + reformatting output

- aggregate, cor, by, as.table, as.data.frame, colnames, merge, cbind, rbind, max, which.max, do.call

High level view

Data analysis is easier if you have a high level view of the data:

- 4 columns + 1 factor (Species)
- Two pairs of related columns: sepals & petals

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Setosa
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Virginica
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Versicolor

?by: applying the cor function

Looking more closely at the way correlation is calculated:

Data frame Column of factors

```
> by(iris, iris[5], function(df) cor(df$Sepal.Length,  
df$Sepal.Width))
```

Declaring a new anonymous function on the fly. Parameter is temporary data frame created for each factor

Values in temp data frame passed to cor function

Sample exam question:



5

Marks

An excerpt of the Iris data set is shown below.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2   setosa
2          4.9         3.0          1.4         0.2   setosa
3          4.7         3.2          1.3         0.2   setosa
4          4.6         3.1          1.5         0.2   setosa
5          5.0         3.6          1.4         0.2   setosa
6          5.4         3.9          1.7         0.4   setosa
```

The following code is run:

```
Petal.cor <- as.data.frame(as.table(by(iris, iris[5], function(df)
  cor(df[3], df[4]))))
colnames(Petal.cor) <- c("Species", "Petal.cor")
Sepal.cor <- as.data.frame(as.table(by(iris, iris[5], function(df)
  cor(df[1], df[2]))))
colnames(Sepal.cor) <- c("Species", "Sepal.cor")
iris.cor <- merge(Sepal.cor, Petal.cor, by = "Species")
iris.cor[,2] = round(iris.cor[,2], digits = 3)
iris.cor[,3] = round(iris.cor[,3], digits = 3)
write.csv(iris.cor, file = "Iris.cor.csv", row.names=FALSE)
```

Describe the action and outputs of the R code.



Lecture 4 – Data science industry

The data science industry

Data science workflow models

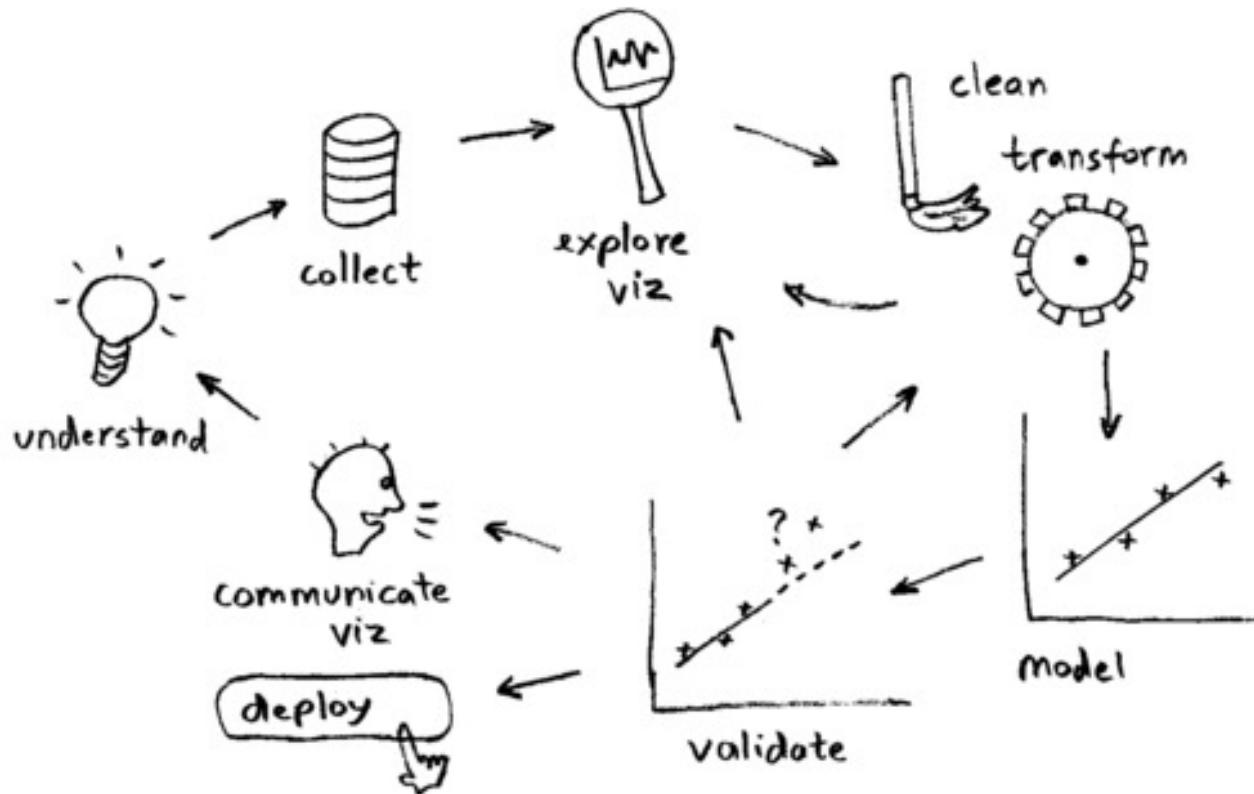
Dirty, and Tidy data

Transforming data: recoding, extracting subsets,
working with dates.

The data science industry

- What is a Data Scientist?
 - > What skills are required?
- The data mining process and need for a consistent analytics methodology:
 - > KDD, SEMMA, CRISP-DM
 - > Business understanding, Data understanding, Data preparation, Modelling, Evaluation, Deployment
- Data preparation and pre-processing, visualization
 - > Sampling, impute missing variables, transformations...

The data science workflow



<http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/>

Dirty Data

Data in the real world is dirty:

- Incorrect data
- Inaccurate data
- Business rule violations
- Inconsistent data
- Incomplete data
- Non-integrated data

See: A taxonomy of dirty data

Sample exam question:



5

Marks

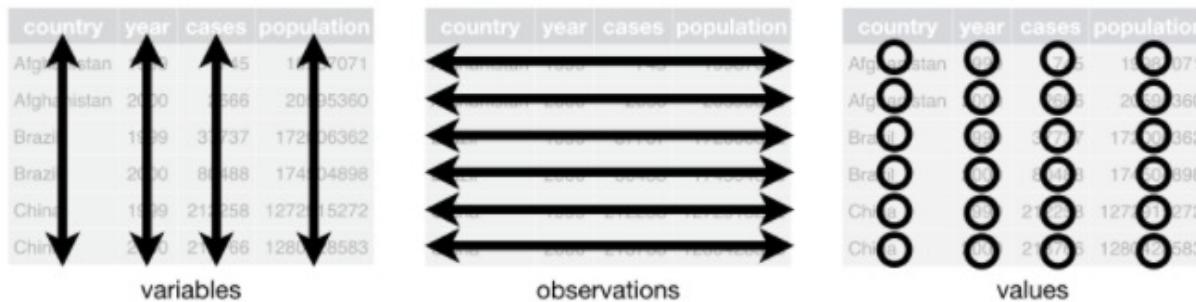
The table below is an extract from the list of books in the British Library. Identify the instances of dirty data present, stating the way in which the data is dirty. One mark will be given for each correct instance up to a maximum of 5 marks.

Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors
206		London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A.	A. A.	FORBES, Walter.
216		London; Virtue & Yo	1868	Virtue & Co.	All for Greed. [A novel. The dedication s	A., A. A.	BLAZE DE BURY, Ma
218		London	1869	Bradbury, Evans &	Love the Avenger. By the author of &Eos	A., A. A.	BLAZE DE BURY, Ma
472		London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to	A., E. S.	Appleyard, Ernest Si
480	A new edition, revis	London	1857	Wertheim & Macint	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
519		London	1872	The Author	Lagonells. By the author of Darmayne	A., F. E.	ASHLEY, Florence En
667		pp. 40. G. Bryan & Co: Oxford, 1898			The Coming of Spring, and other poems	A. J. A. J.	ANDREWS, J. - Writ
874		London]	1676		A Warning to the inhabitants of England	Rema��z.	ADAMS, Mary.
1143		London	1679		A Satyr against Vertue. (A poem: suppos	A., T.	OLDHAM, John.
1280		Coventry	1802	Printed by J. Turner	An Account of the many and great Loan		CARTE, Samuel. JAC
1808		Christiania	1859		Erindringer som Bidrag til Norges Histor	AALL, Jacob.	AALL, J. C. LANGE, C
1905		Firenze	1888		Gli Studi storici in terra d'Otranto ... Fra	AAR, Ermanno - pse	S., L. G. D. SIMONE,
1929		Amsterdam	1839, 38-54		De Aardbol. Magazijn van hedendaagsc		WITKAMP, Pieter Hz
2836		Savona	1897		Cronache Savonesi dal 1500 al 1570 ... A	ABATE, Giovanni Ag	ASERETO, Giovanni
2854		London	1865	E. Moxon & Co.	See-Saw; a novel ... Edited [or rather, w	ABATI, Francesco.	READE, William Win
2956		Paris	1860-63		Gelstodeltsie d'une partie de la Haute E	ABBADIE, Antoine T	RADAU, Rodolphe.
2957		Paris	1873		[With eleven maps.]	ABBADIE, Antoine T	RADAU, Rodolphe.
3017	Nueva edicion, anot	Puerto-Rico	1866		[Historia geograf��rica, civil y politica de	ABBAD Y LASIERRA, ACOSTA Y CALBO, Jo	
3131		New York	1899	W. Abbott	The Crisis of the Revolution, being the s	ABBATT, William.	ANDREI, John - Ma
4598		Hull	1814	The Author	Peace: a lyric poem. [With prefatory ad	ABBOTT, Thomas Es	WRANGHAM, Franc
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chr		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdullahi Histori��] A��gypti compen		WHITE, Joseph - Car
5382		London	1847, 48 [1846-48]	Punch Office	The Comic History of England ... With ... A'	BECKETT, Gilbert	LEECH, John - Artist
5385	[Another edition.] II	London	[1897?]	Bradbury, Agnew &	[The comic history of England ... With tw	BECKETT, Gilbert	LEECH, John - Artist
5389	[Another edition.]	London	[1897?]	Bradbury, Agnew &	[The Comic History of Rome ... Illustrate	BECKETT, Gilbert	LEECH, John - Artist
5432		Milano	1893		Signa: opera in tre atti [founded on the	A' BECKETT, Gilbert	MAZZUCATO, Giova
6036		London	1805	C. & R. Baldwin	The Venetian Outlaw, a drama in three		ELLISTON, Robert W
6821		Aberdeen	1837	J. Davidson & Co.	Description of the Coast between Aberd		DUNCAN, William - i

Tidy data

Tidy data seeks a consistent format that has:

- Each variable in its own column.
- Each observation in its own row.
- Each value in its own cell.



- 2 benefits: consistency and exploits R's vector nature

<http://www.jstatsoft.org/v59/i10/paper>

Lecture 5 – Networks

Structure

- Nodes (vertices) and edges (arcs), directed – undirected, weighted – unweighted.
- Walk, Path, Cycle, Geodesic, Length, Connected,
- Loop, Complete, Subgraph, Clique, Simple.

Network Statistics

- Diameter, Average path length, Degree distribution, Density and Clustering coefficient.

Vertex Characteristics

- Degree, Centrality: Betweenness, Closeness, Eigenvector.

Analysis

- Igraph package in R
- Graph data object. Creating graphs in R. Data entry using different methods, edge list, csv, etc.
- Plots
- Detecting the most important actors in a network using centrality measures.
- Theoretical graph types (E-R (random), B-A (scale-free), W-S (small world)).

Vertex summary

Degree, betweenness and closeness centrality

```
> degree(g)  
J A S D R M  
3 4 2 1 3 1
```

```
> betweenness(g)  
J A S D R M  
4 5 0 0 1 0
```

```
> format(closeness(g), digits = 2)  
J A S D R M  
"0.143" "0.167" "0.111" "0.091" "0.143" "0.100"
```

Vertex summary cont...

Eigenvector centrality (first create the eigenvector and then format the output).

```
> e = evcent(g)
> format(e$vector, digits = 2)
```

J	A	S	D	R	M
"0.80"	"1.00"	"0.69"	"0.29"	"0.90"	"0.36"

What does this all mean?

Summary: who is the most important person in the network?

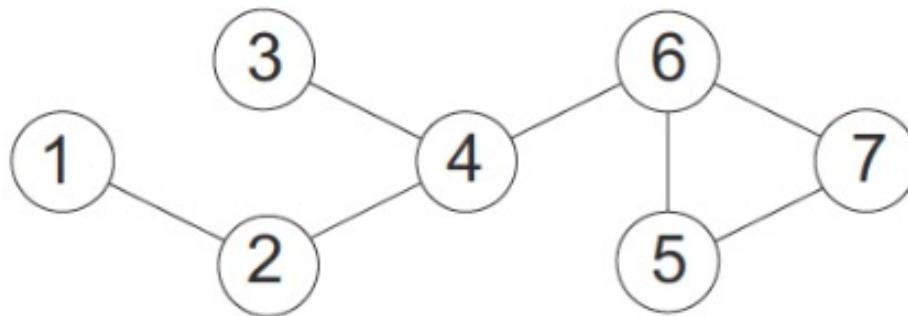
	J	A	S	D	R	M
Degree	3.00	4.00	2.00	1.00	3.00	1.00
Betweenness	4.00	5.00	0.00	0.00	1.00	0.00
Closeness	0.14	0.17	0.11	0.09	0.14	0.10
Eigenvector	0.80	1.00	0.69	0.29	0.90	0.36

Sample exam question



3
Marks

The social network of a group of friends (numbered from 1 – 7) is drawn below.



6a)

1
Marks

Calculate the density of the graph.

Lecture 6 – Linear regression

Fitting the regression

- Assumptions, LOBF, linear model object: coefficients & residuals, prediction

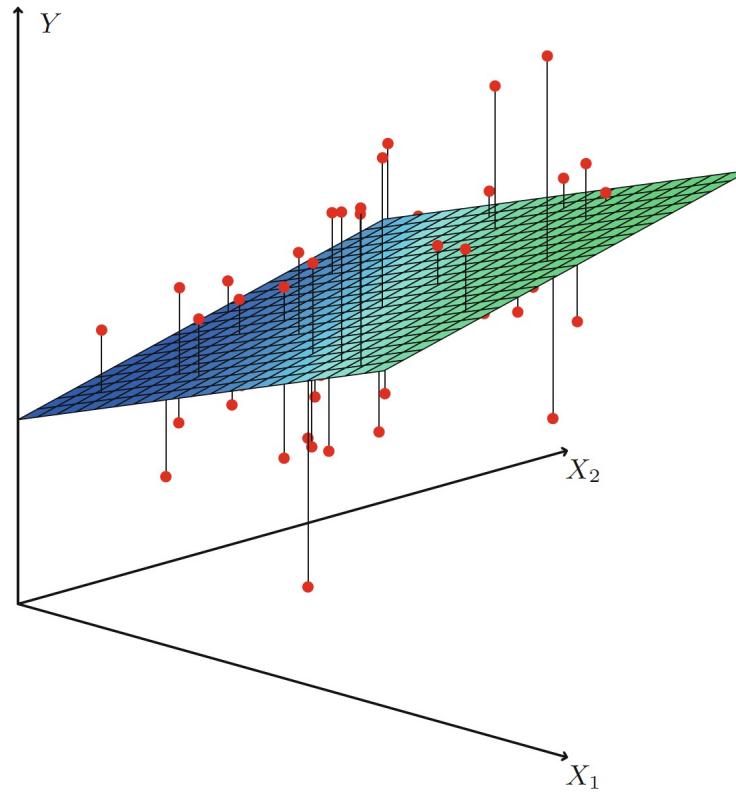
Regression diagnostics.

Multiple linear regression

Regression with qualitative variables

- Using non-numerical factors

Multiple linear regression



From: G. James et al., An Introduction to Statistical Learning: with Applications in R (2013).

Diagnostics – summary

```
> summary(fit)
```

Call:

```
lm(formula = Function ~ Price)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.3839	-6.8347	0.0382	8.1903	13.4312

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	44.020	4.565	9.642	3.09e-10 ***
Price	6.942	1.502	4.621	8.43e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '.'

Residual standard error: 9.185 on 27 degrees of freedom

Multiple R-squared: 0.4416, Adjusted R-squared: 0.421

F-statistic: 21.36 on 1 and 27 DF, p-value 8.428e-05

Median close to 0

Coefficients: α, β

Hypothesis test that $\alpha, \beta = 0$ vs $\alpha, \beta \neq 0$

Coefficient of Determination: r^2

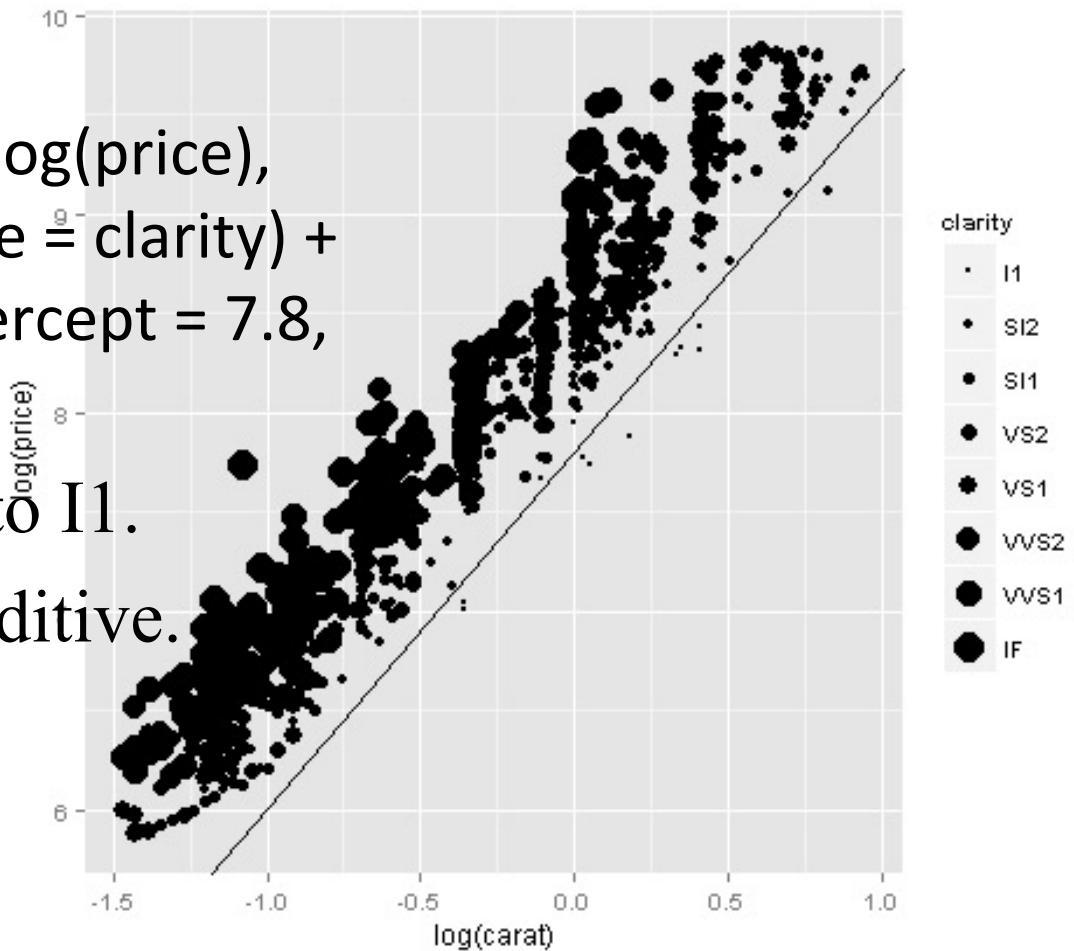
Overall significance of regression: that at least one coefficient $\neq 0$

Fitted model

$\ln(\text{price})$ v $\ln(\text{carat})$

```
> qplot(log(carat), log(price),  
       data = dsmall, size = clarity) +  
       geom_abline(intercept = 7.8,  
                   slope = 1.8)
```

- Basic model fitted to I1.
- Quality increase additive.



Sample exam question:



4

Marks

The least squares regression of $\log(\text{price})$ on $\log(\text{size})$ and color is given below. Note that 'log' in this context means ' $\text{Log}_e(X)$ '. Based on this output, answer the following questions.

```
> library(ggplot2)
> set.seed(9999) # Random seed
> dsmall <- diamonds[sample(nrow(diamonds), 1000), ] # sample of 1000 rows
> attach(dsmall)
> contrasts(color) = contr.treatment(7)

> d.fit <- lm(log(price) ~ log(carat) + color)
> d.fit

> summary(d.fit)

Call:
lm(formula = log(price) ~ log(carat) + color)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.97535 -0.16001  0.01106  0.15500  0.99937 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.61356   0.02289 376.259 < 2e-16 ***
log(carat)  1.74075   0.01365 127.529 < 2e-16 ***
color2     -0.06717   0.02833 -2.371  0.0179 *  
color3     -0.05469   0.02783 -1.965  0.0496 *  
color4     -0.07139   0.02770 -2.578  0.0101 *  
color5     -0.21255   0.02973 -7.148  1.7e-12 ***
color6     -0.32995   0.03175 -10.393 < 2e-16 ***
color7     -0.50842   0.04563 -11.143 < 2e-16 *** 
---
Residual standard error: 0.2393 on 992 degrees of freedom
Multiple R-squared:  0.9446,    Adjusted R-squared:  0.9443 
F-statistic: 2418 on 7 and 992 DF,  p-value: < 2.2e-16
```

3a)

1

Marks

Write down the regression equation predicting log(price) as a function of size and color.



Report question issue

3b)

2

Marks

Explain the different data types present in the variables: carat and color. What is the effect of this difference on the regression equation?

Lecture 7 – Decision Trees

Overview of data mining and machine learning

Introduction to classification and decision trees:

- Leaf nodes, Branches, Tree nodes...

A specific decision tree algorithm: ID3

Entropy and information gain

- Decision attributes, Attribute splitting

Model accuracy; training and performance evaluation

Implementing a decision tree using R

Calculating entropy

For a two-class problem: c_1 and c_2 :

- P indicates the probability of belonging to each class, the number in each class is $N_{c1} + N_{c2} = N$.

$$\begin{aligned}\text{Entropy}(S) &= -P_{c1} \log_2(P_{c1}) - P_{c2} \log_2(P_{c2}) \\ &= -\frac{N_{c1}}{N} \log_2\left(\frac{N_{c1}}{N}\right) - \frac{N_{c2}}{N} \log_2\left(\frac{N_{c2}}{N}\right)\end{aligned}$$

For a multi-class problem

$$\begin{aligned}\text{Entropy}(S) &= -\sum_{i=1}^C P_i \log_2(P_i) \\ &= -\sum_{i=1}^C \frac{N_i}{N} \log_2\left(\frac{N_i}{N}\right)\end{aligned}$$

Information gain

Information gain is the expected reduction in entropy caused by partitioning the examples according to an attribute A.

- Gain(S, A) of an attribute A, relative to a collection of examples S (with v groups having $| |$ elements) is:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

Entropy before split

Expected entropy after split

Playing tennis: initial entropy

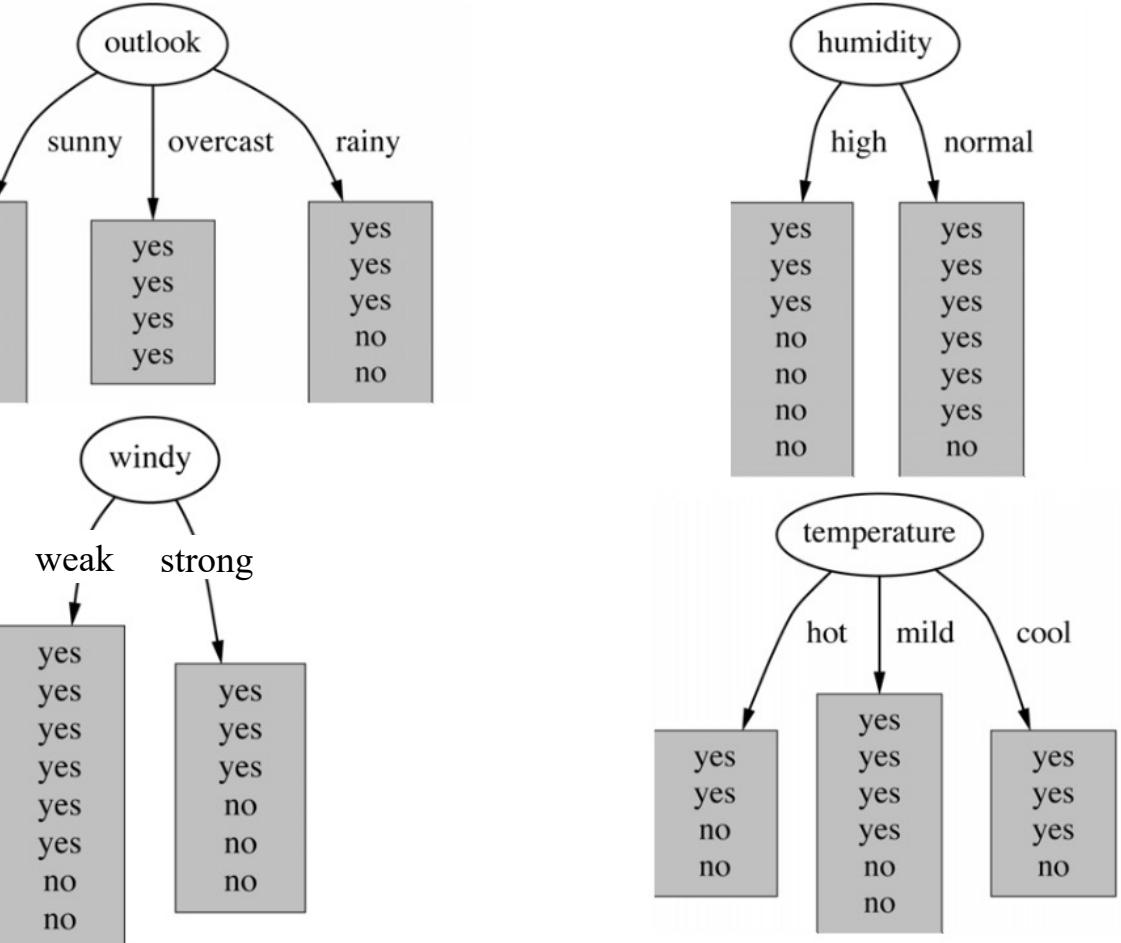
Training set (S): Initial entropy before splitting based on 9 Yes/5 No:

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Yes	No	P(Yes)	Log2(Yes)	P(No)	Log2(No)	Entropy
9	5	0.6429	-0.6374	0.3571	-1.4854	0.9403

Which attribute to select?

Remember - ID3 chooses the attribute which gives the greatest information gain (reduction in Entropy), or the ‘purest’ result.



We next calculate the information gain for each attribute in turn.

Metrics for Performance Evaluation

		PREDICTED CLASS	
		Class=Yes	Class>No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class>No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

Also:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

Sample exam question



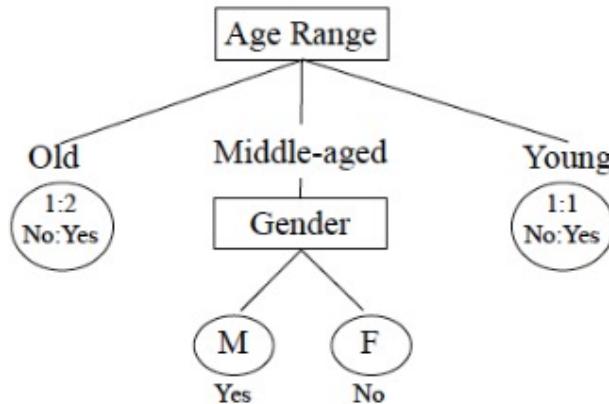
4

Marks

Eight university staff completed a questionnaire on happiness. The results are given below.

ID	Age Range	Occupation	Gender	Happy
1	Young	Tutor	F	Yes
2	Middle-aged	Professor	F	No
3	Old	Tutor	M	Yes
4	Middle-aged	Professor	M	Yes
5	Old	Tutor	F	Yes
6	Young	Lecturer	M	No
7	Middle-aged	Lecturer	F	No
8	Old	Tutor	F	No

A decision tree was generated from the data.



Sample exam question

12b)

1
Marks

What is the entropy of "Happy"?



12c)

2
Marks

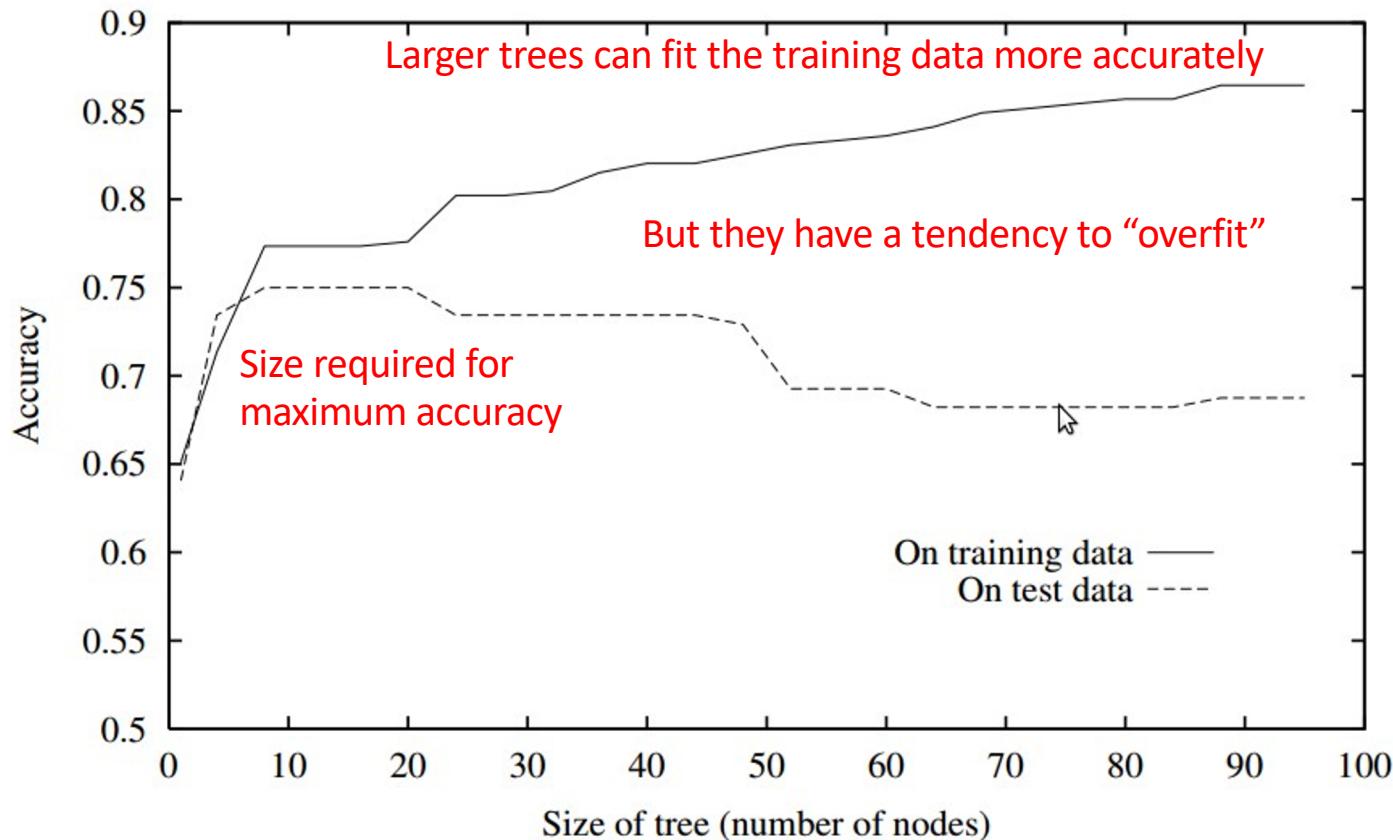
What is information gain after the first node of the decision tree (Age Range) has been introduced?

Lecture 8 – Classification continued

Classification Continued:

- Improving the basic decision tree: pruning and cross-validation,
- Naïve Bayes classification,
- Classifier model evaluation: ROC evaluation,
- Lift.

Example: size of tree vs accuracy



Avoiding over-fitting

Two approaches:

1. Pre-pruning: stop growing the tree earlier, before it reaches the point where it over fits on the training data.

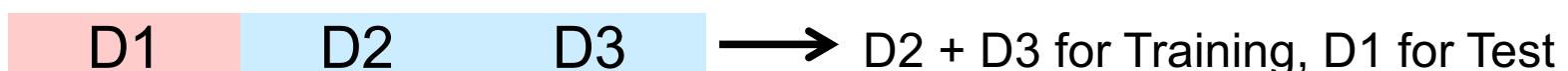
2. Post-pruning: allow the tree to overfit the data, and then post-prune the tree. (more effective in practice).

Example: 3 fold cross validation

Partitioning:

- 2/3 training (making the model)
- 1/3 testing (measuring performance of model)

Repeat the procedure three times so that every case has been used exactly once for testing.



Average performance across D1, D2 and D3.

Bayesian classifiers

Approach:

- Compute the posterior probability for all values of C_j using Bayes theorem

$$P(C_j | A_1 \cap A_2 \cap A_3 \dots \cap A_n) = \frac{P(C_j) \cdot P(A_1 \cap A_2 \cap A_3 \dots \cap A_n | C_j)}{P(A_1 \cap A_2 \cap A_3 \dots \cap A_n)}$$

- Choose the value of C_j that maximises this posterior probability $P(C_j | A_1 \cap A_2 \cap A_3 \dots \cap A_n)$.
- Is equivalent to maximising $P(C_j) \cdot P(A_1 \cap A_2 \cap A_3 \dots \cap A_n | C_j)$, since $P(A_1 \cap A_2 \cap A_3 \dots \cap A_n)$ same for all C_j .

Naïve Bayes classifier

- Assume independence among attributes A_i when class is given, thus

$$P(A_1 \cap A_2 \cap A_3 \dots \cap A_n | C_j) = P(A_1 | C_j) \times P(A_2 | C_j) \times \dots \times P(A_n | C_j)$$

- Estimate $P(A_i | C_j)$ for all A_i
- A new point is classified to the C_j which maximises

$$P(C_j) \times P(A_1 | C_j) \times P(A_2 | C_j) \times \dots \times P(A_n | C_j)$$

- Classification confidence (probability) is given by

$$P(C_j | A_1 \cap A_2 \cap A_3 \dots \cap A_n) = \frac{P(C_j) \cdot P(A_1 \cap A_2 \cap A_3 \dots \cap A_n | C_j)}{P(A_1 \cap A_2 \cap A_3 \dots \cap A_n)}$$

Sample exam question



3

Marks

Use data below and Naive Bayes classification to predict whether the following test instance will be happy or not.

ID	AgeRange	Occupation	Gender	Happy
1	Young	Tutor	F	Yes
2	Middle-aged	Professor	F	No
3	Old	Tutor	M	Yes
4	Middle-aged	professor	M	Yes
5	Old	Tutor	F	Yes
6	Young	Lecturer	M	No
7	Middle-aged	lecturer	F	No
8	Old	Tutor	F	No

Test instance: (AgeRange = young, Occupation = professor, Gender = F, Happy = ?)



Sample exam question



1

Mark

Use the complete Naïve Bayes formula to evaluate the confidence of predicting Happy = yes, based on the same attributes as the previous question: (AgeRange = young, Occupation = professor, Gender = F).



Hint: adapt $P(C_j | A_1 \cap A_2 \cap A_3 \dots \cap A_n) = \frac{P(C_j) \cdot P(A_1 \cap A_2 \cap A_3 \dots \cap A_n | C_j)}{P(A_1 \cap A_2 \cap A_3 \dots \cap A_n)}$ for independent events.

Receiver Operating Characteristic

Developed in the 1950s for signal detection - to analyse noisy signal transmission.

- Characterises the tradeoff between positive hits and false alarms.
- ROC plots True Positive Rate, TPR, (on y axis) against False Positive Rate, FPR, (on x axis).
- Performance of a single classifier presented as a single point of ROC curve.
- Changing the threshold of algorithm, sample distribution or cost matrix etc. changes that point: this lets a profile of classifier to be developed.

Receiver Operating Characteristic

Calculating TPR and FPR

- True Positive Rate, TPR, also called *sensitivity* indicates how good a test is for correctly predicting “yes” when it should predict “yes”. (Think of confidence)

$$\text{True Positive Rate: } TPR = \frac{TP}{TP + FN}$$

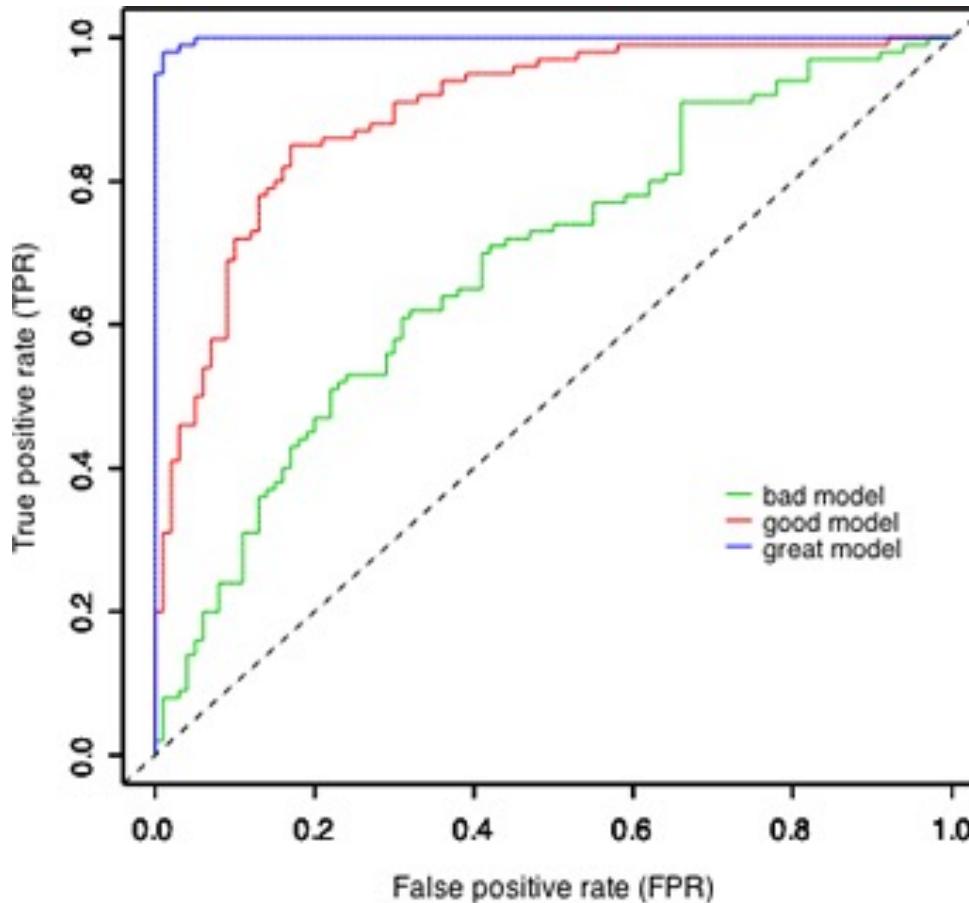
- False Positive Rate, FPR, also known as a ‘*false alarm*’ is:

$$\text{False Positive Rate: } TPR = \frac{FP}{FP + TN}$$

- ROC plots TPR against FPR

Using ROC for Model Comparison

ROC curves for three different classifiers



Sample exam question



4
Marks

The following table shows the outcome of a classification model for customer data. The table lists customers by code and provides the following information: The model confidence of a customer buying/not buying a new product (confidence-buy); whether in fact the customer did or did not buy the product (buy = 1 if the customer purchased the model, buy = 0 if the customer did not buy the model).

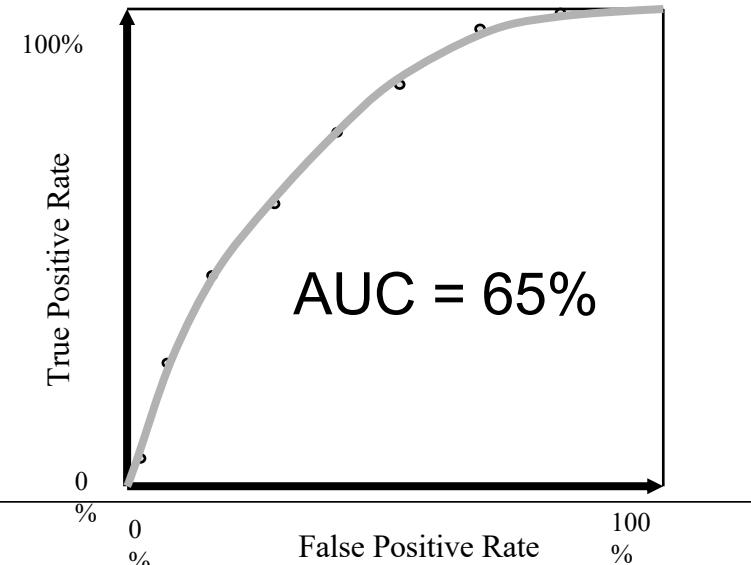
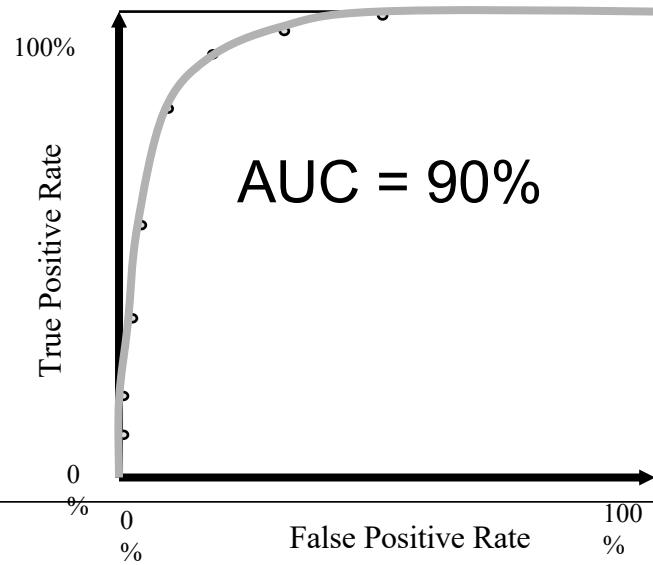
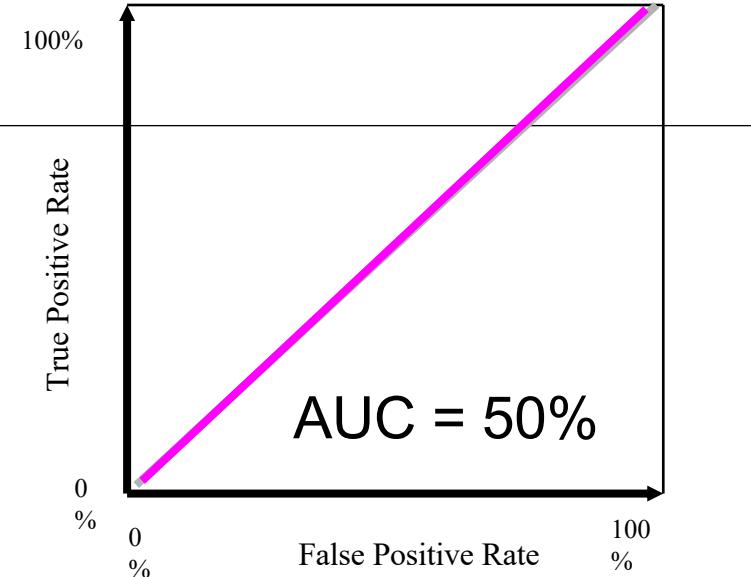
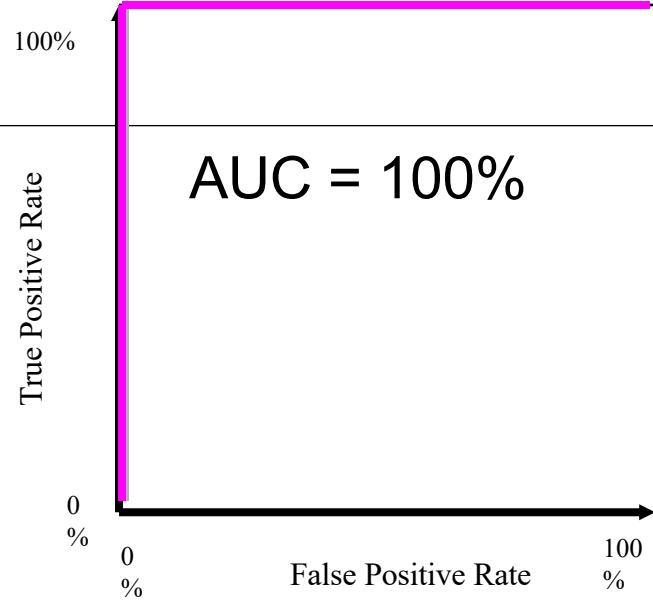
customer	confidence-buy	buy-not-buy
c1	0.9	1
c2	0.8	1
c3	0.7	0
c4	0.7	1
c5	0.6	1
c6	0.5	1
c7	0.4	0
c8	0.4	1
c9	0.2	0
c10	0.1	0

13a)

2
Marks

Calculate the True Positive Rate and the False Positive Rate when a confidence level of 20% is required for a positive classification.

AUC for ROC curves

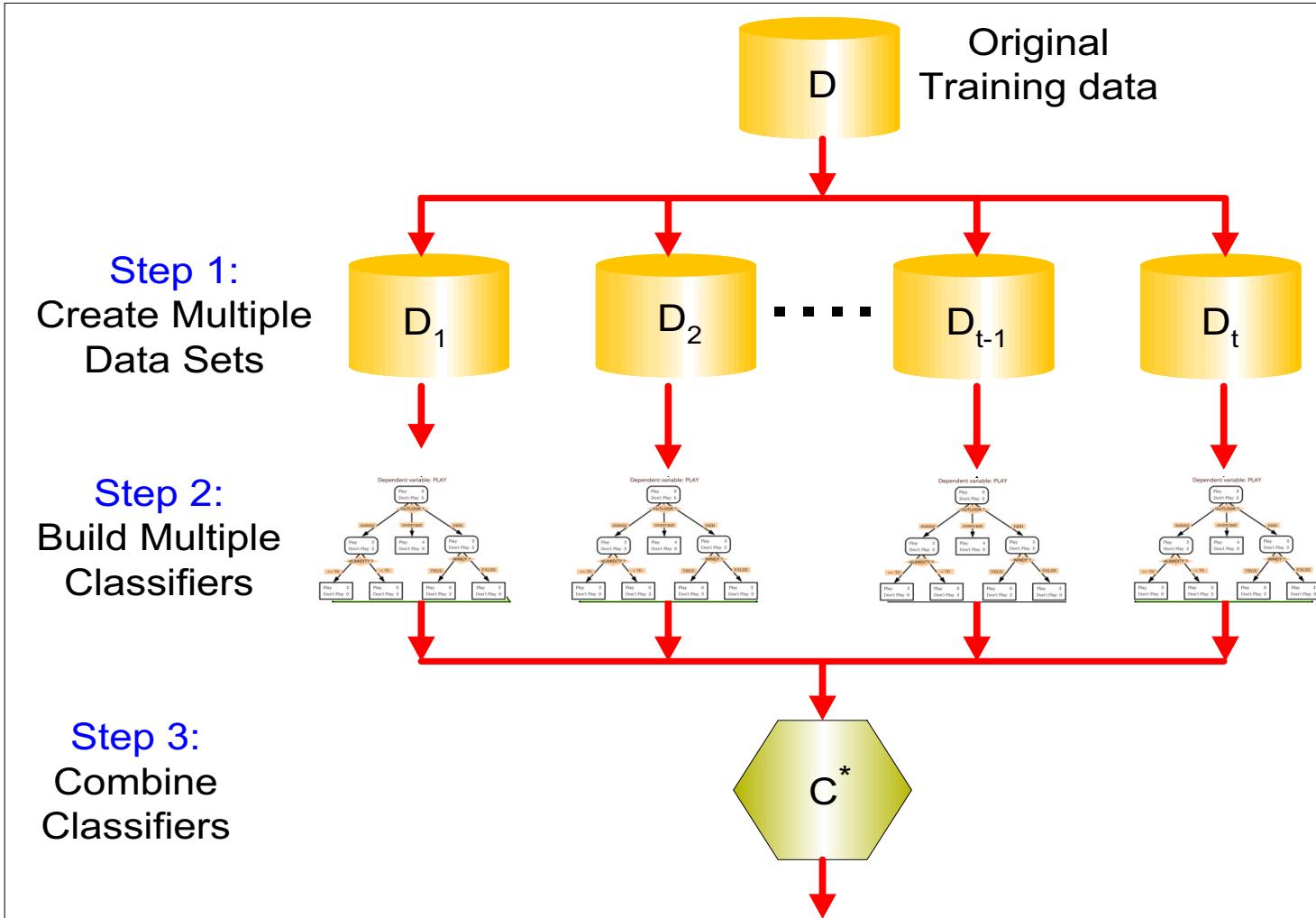


Lift

For binary classification and prediction models:

- Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- Lift charts are visual aids for measuring model performance
- Lift factor = success rate with model / success rate without model

Lecture 9 Ensemble methods + ANN



Bagging (Bootstrap aggregation)

- Bootstrapping: resampling the original data set to produce multiple synthetic data sets.
- Sampling is uniform, each bootstrap replicate may have multiple instances of the original data points and contains approximately 63% of the original data set.
- Build a classifier on each of the replicates and combine results by voting. Multiple classifiers reduce the (high) variance of individual decision trees, (recall: sample variance is $\frac{\sigma}{\sqrt{n}}$).

Boosting

Multiple trees are grown slowly, using incremental improvement.

Data set is not bootstrap sampled, but training examples are weighted, with a higher weight given to hard to classify examples at each step.

Classification is by a weighted sum from each classifier, with more accurate classifiers having a greater weight.

Random Forest

A refinement of bagged decision trees.

Specifically designed for decision trees.

Random Forest Algorithm

- Create multiple data sets from the original training set using subsets of data points and subsets of attributes.
- Build a decision tree classifier for each new data.
- Combine the classifiers by taking a majority vote to produce the final decision.

Artificial Neural Networks, ANNs

- Are computer models of neural behaviour in the (human) brain.
- Are applicable to wide range of problems
- Have the ability to ‘learn’ by weighting the contribution of each neuron to a decision (output).
- They are accurate, can handle redundant attributes and noisy data.
- Large ANNs give rise to ‘deep learning’.

Setting up ANNs

Pre-processing

- One input neuron for each input variable,
- One output neuron for each output class,
- Inputs can only be numerical (R will accept binary TRUE FALSE),
- Data should be normalised,
- Categorical, data needs to be converted to binary columns as indicator variables,
- No missing values.

Lecture 10 – Clustering

k-Means clustering

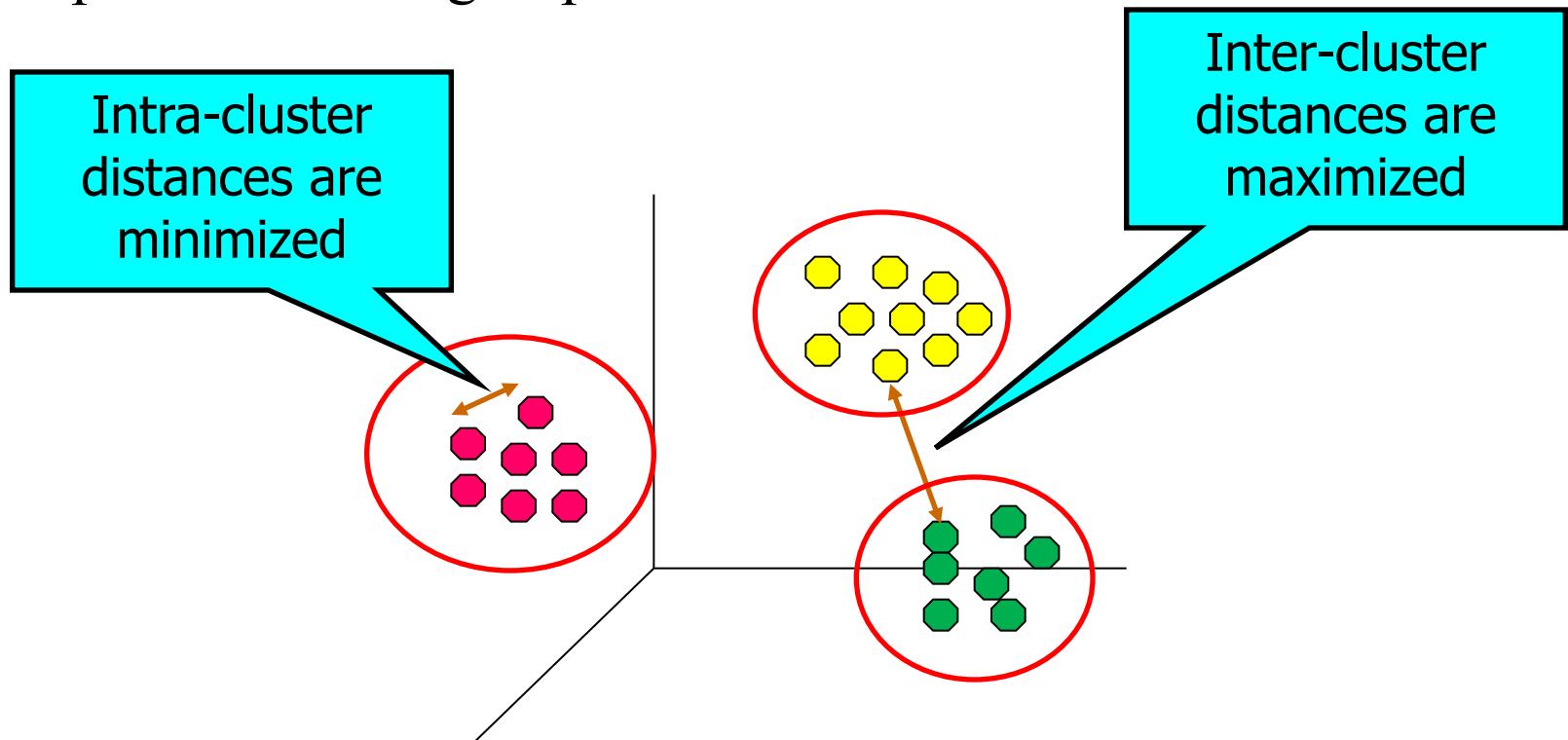
Hierarchical clustering

Clustering using R

Recall: clustering is unsupervised learning!

What is Cluster Analysis?

Finding groups of points such that the points in a group will be similar (or related) to one another and different from (or unrelated to) the points in other groups



k-Means Clustering

Partitional clustering approach

Each cluster is associated with a **centroid** (center point)

Each point is assigned to the cluster with the closest centroid

Number of clusters, k , must be specified

The basic algorithm is very simple:

1. Select k points (at random) as the initial centroids
2. **Repeat**
 3. Form k clusters by assigning all points to the closest centroid
 4. Re-compute the centroid of each cluster
5. **Until** the centroids don't change

Sample exam question



4

Marks

A k-Means clustering algorithm is fitted to the iris data, as shown below.

```
rm(list = ls())
data("iris")
ikfit = kmeans(iris[,1:2], 4, nstart = 10)
ikfit
table(actual = niris$Species, fitted = ikfit$cluster)
```

Based on the R code and output below, answer the following questions.

```
> ikfit
K-means clustering with 4 clusters of sizes 24, 53, 41, 32

Cluster means:
  Sepal.Length Sepal.Width
1     4.766667    2.891667
2     5.924528    2.750943
3     6.880488    3.097561
4     5.187500    3.637500

Within cluster sum of squares by cluster:
[1] 4.451667 8.250566 10.634146 4.630000
  (between_SS / total_SS =  78.6 %)

> table(actual = niris$Species, fitted = ikfit$cluster)
            fitted
actual      1   2   3   4
  setosa     18   0   0  32
  versicolor  5  34  11   0
  virginica   1  19  30   0
```

Sample exam question

16a)

2

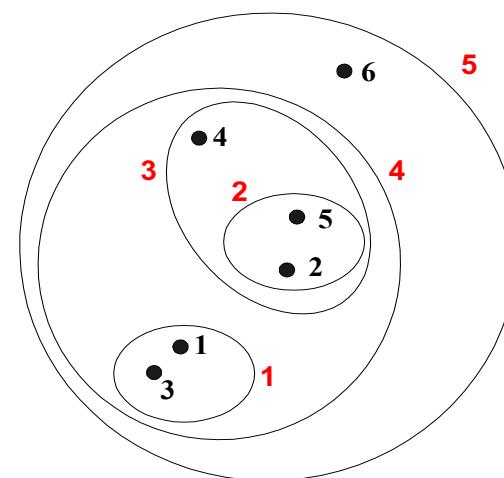
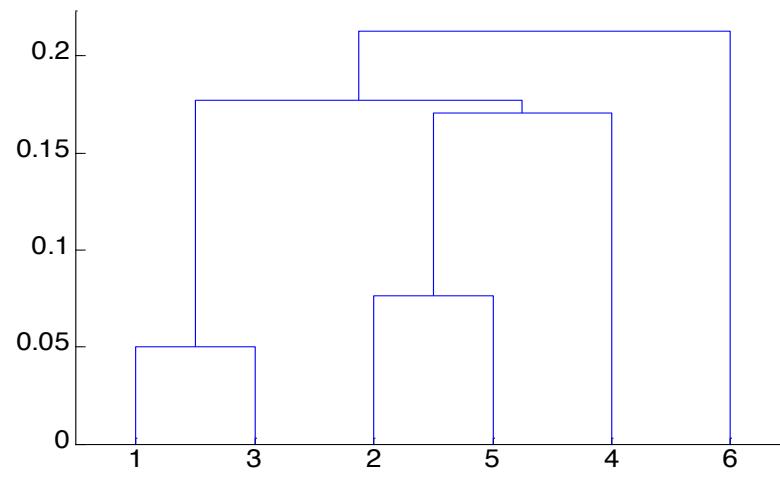
Marks

Comment on the quality of the clustering giving at least one quantitative measure.

Hierarchical Clustering

Produces a set of nested clusters organized as a **hierarchical tree**

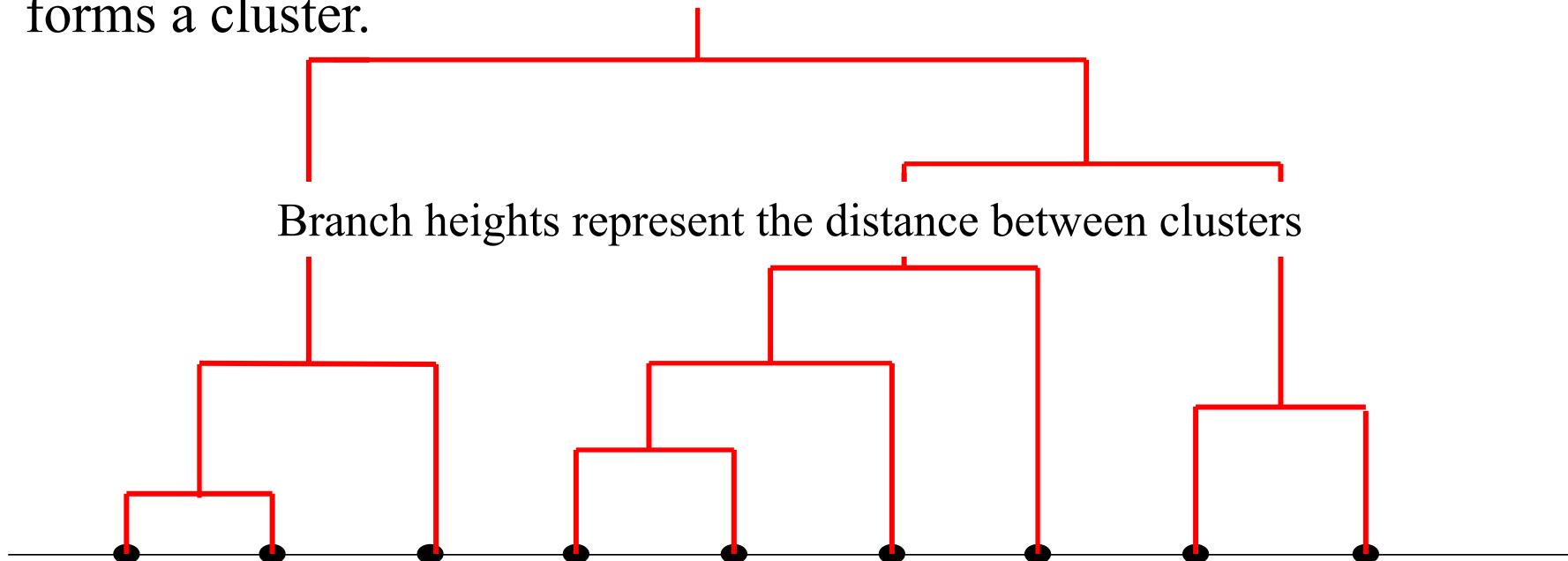
- Records the sequences of merges or splits.
- Can be visualized as a dendrogram or enclosure diagram.



Dendrogram and hierarchies

Decompose data objects into a several levels of nested partitioning (**tree of clusters**), called a **dendrogram**.

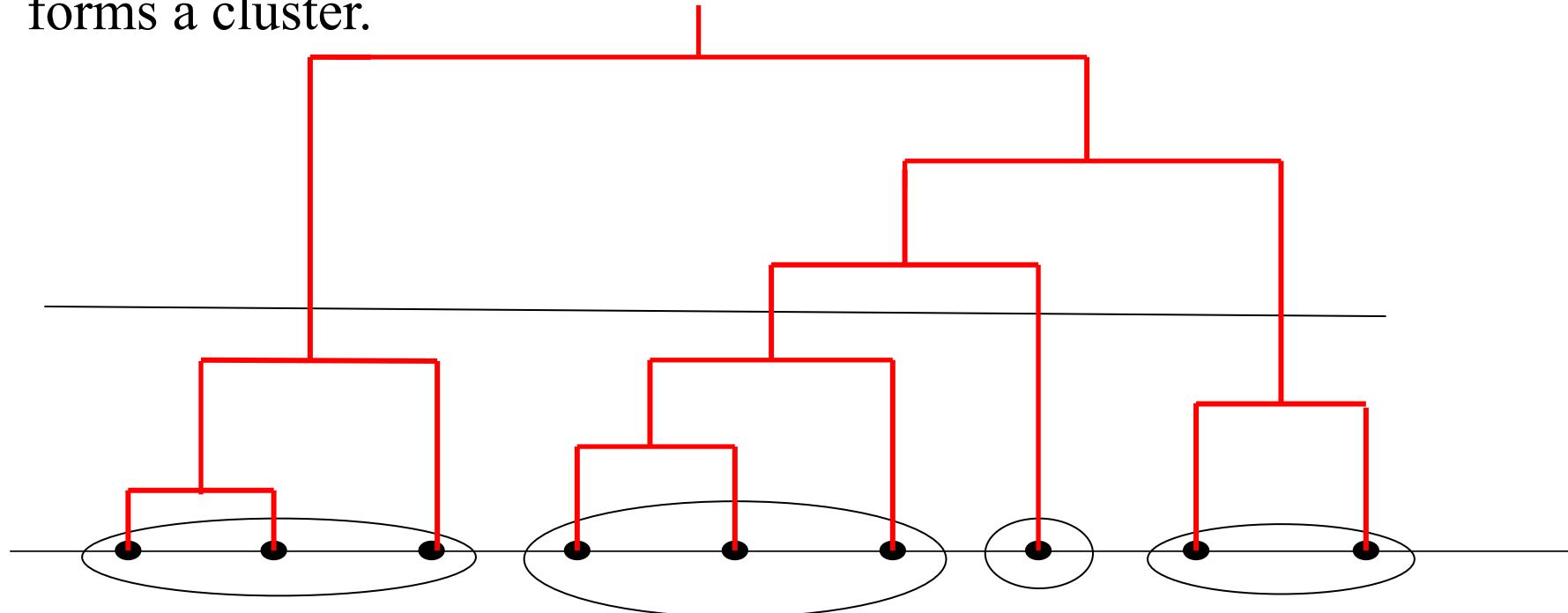
A **clustering** of the data objects is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.



Dendrogram and hierarchies

Decompose data objects into a several levels of nested partitioning (**tree of clusters**), called a **dendrogram**.

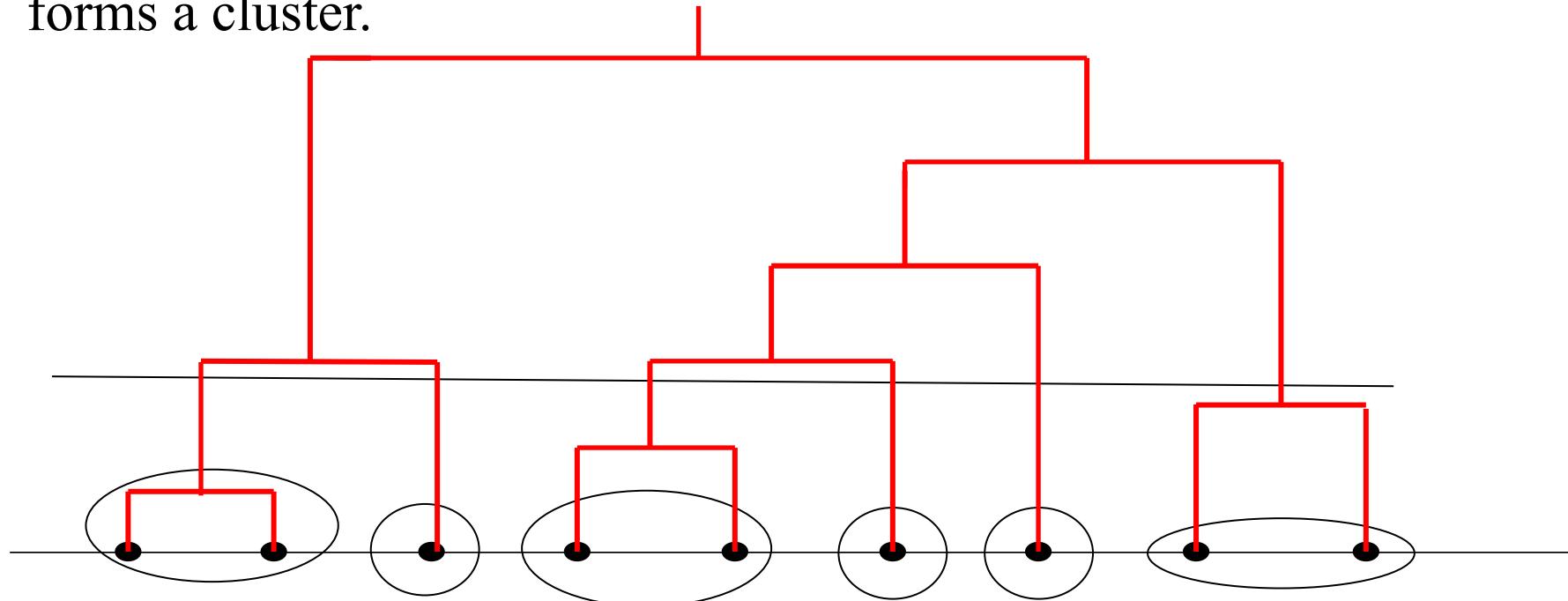
A **clustering** of the data objects is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.



Dendrogram and hierarchies

Decompose data objects into a several levels of nested partitioning (**tree of clusters**), called a **dendrogram**.

A **clustering** of the data objects is obtained by **cutting** the dendrogram at the desired level, then each **connected component** forms a cluster.



Text analytics:

Covers many different activities:

- Information retrieval
- Text mining (*traditional name and still popular*)
- Web mining
- NLP: Natural Language Processing (*words & meaning*)
- Document classification
- Document clustering
- Sentiment analysis
- Topic analysis

Bag of words

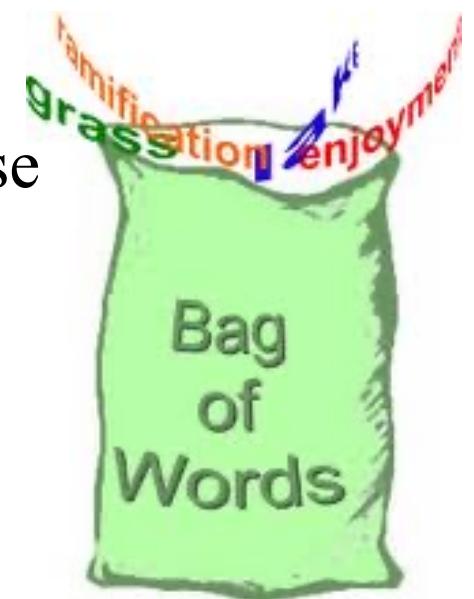
The vector space model uses a ‘bag of words’ approach.

- Each document assumed to be just a collection of words
- Makes implicit assumptions that the order of the words in a document does not matter
- Syntactically similar documents are semantically similar – which is often the case

These assumptions not always valid e.g.

- ‘James and the giant ate a peach.’
- ‘The giant ate James and a peach.’

But works well in practice...



Extracting structure from text

Several steps:

- Tokenise
- Convert case
- Remove stop words
- Stem
- Lemmatize
- Create n-grams

TDM for processed documents

The Corpus:

- Doc1: {dog | dog_eat | eat | eat_homework | homework}
- Doc2: {cat | cat_eat | eat | eat_sandwich | sandwich}
- Doc3: {dolphin | dolphin_eat | eat | eat_homework | homework}

Term-Document (Frequency) Matrix

Document	eat	eat_homework	eat_sandwich	cat	cat_eat	dog	dog_eat	dolphin	dolphin_eat	homework	sandwich
Doc 1	1	1	0	0	0	1	1	0	0	1	0
Doc 2	1	0	1	1	1	0	0	0	0	0	1
Doc 3	1	1	0	0	0	0	0	1	1	1	0

Sample exam question



2

Marks

Apply the five main steps required to pre-process text documents for analysis to the corpus below. Write your processed documents in the space provided.

Doc1 = { The church choir sang loudly. }

Doc2 = { The boys were singing in the church choir. }

Doc3 = { The boy asked to sing a song. }

2

Marks

Construct the term document frequency matrix for the processed text documents above. Use as many cells as you require below.

FIT3152 Data analytics

Summary

Unit objectives (from Lecture 1)

What the course is trying to achieve:

- We are concentrating on fundamental, generic, skills for a data scientist that are independent of software platform or problem domain.
- Problem solving skills, independence and ingenuity. Good communication skills.

What it is not trying to achieve:

- Introduction to the vast range of software, techniques and computing platforms available to data scientists.

Technical skills

Using R as a platform we have covered:

- Basic statistics
- Exploring data using graphics
- Data manipulation
- Linear regression
- Network analysis
- Decision trees, ensemble methods, ANNs (supervised)
- Clustering (unsupervised)
- Text analysis

High-level skills (Lecture 1)

Some necessary skills for a data scientist:

- Understand a problem from client's perspective
- Collect, cleanse, manage and combine data – which may come from disparate sources
- Understand the data, most likely using visualization tools as a starting point
- Analyze and model the data using statistical and (AI) machine learning techniques
- Communicate the results simply and effectively.

Where to next?

If you want to take data science further:

- Kaggle competitions
www.kaggle.com
- MeetUp groups
<https://www.meetup.com/Data-Science-Melbourne/>
- Employment: business, finance, government, all areas of industry, science, consulting ...
- Further study (FIT has Master of Data Science)

Good Luck!