FIT3152 Mock eExam with brief Answers/Marking Guide

R Coding: 10 Marks

```
> 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 R 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.

Calculate the correlation of sepal length and width [1 Mark]
Calculate the correlation of petal length and width [1 Mark]
Rename and merge data frames [1 Mark]
Round the values [1 Mark]
Save as a csv file [1 Mark]

Describe the action of each function or purpose of each variable in the space provided.

as.data.frame	Coerce the previous output into a data frame [1 Mark]
merge	Merge data frames using a common column as an index [1 Mark]
by	Apply a function to a data frame split by factors [1 Mark]
df	Temporary data frame passed to the function [1 Mark]
round	Round the data to a given number of decimal places (or digits) [1 Mark]

(10 marks)

Regression: 10 Marks

A subset of the 'diamonds' data set from the R package 'ggplot2' was created. The data set reports price, size(carat) and quality (cut, color and clarity) information as well as specific measurements (x, y and z). The first 6 rows are printed below.

```
> head(dsmall)
      carat       cut color clarity depth table price    x    y    z
46434  0.59 Very Good     H    VVS2  61.1    57  1771 5.39 5.48 3.32
35613  0.30      Good     I     VS1  63.3    59   473 4.20 4.23 2.67
43173  0.42   Premium     F      IF  62.2    56  1389 4.85 4.80 3.00
11200  0.95     Ideal     H     SI1  61.9    56  4958 6.31 6.35 3.92
37189  0.32   Premium     D    VVS1  62.0    60   973 4.40 4.37 2.72
45569  0.52   Premium     E     VS2  60.7    58  1689 5.17 5.21 3.15
```

The least squares regression of log(price) on log(size) and color is given below. Note that 'log' in this context means '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

> contrasts(color)
  2 3 4 5 6 7
D 0 0 0 0 0 0
E 1 0 0 0 0 0
F 0 1 0 0 0 0
G 0 0 1 0 0 0
H 0 0 0 1 0 0
I 0 0 0 0 1 0
J 0 0 0 0 0 1
```

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

```
log(price) = 1.74 * log(carat) + 8.61 + color(i),
where i = indicates color(D,E,F,G,H,I,J) [1 Mark]
```

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

```
carat is a numerical variable (treated as a number) [1 Mark]
color is a factor – it is included in the regression equation
as a contrast whereby each level is estimated individually.
[1 Mark]
```

(c) What is the predicted price for a diamond of 1 carat of color H?

```
log(price) = 1.74 * log(carat) + 8.61 + color(i),
log(price) = 1.74 * log(1) + 8.61 -0.21,
log(price) = 1.74 * 0 + 8.61 -0.21,
log(price) = 8.61 -0.21 = 8.40
price = e ^ 8.40 = $ 4447.06 [1 Mark]
```

(d) Which colour diamonds can be reliably assumed to have the highest value? Explain your reasoning. How sure can you be?

```
Color D diamonds have the highest value since the coefficient
for this factor is 0 and all the others are negative. [1 Mark]
For surety, use the significance of the regression equation
overall (***) so better than 0.0001 [1 Mark]
```

(e) Which colour diamonds have the lowest value? How reliable is the evidence? Explain your reasoning.

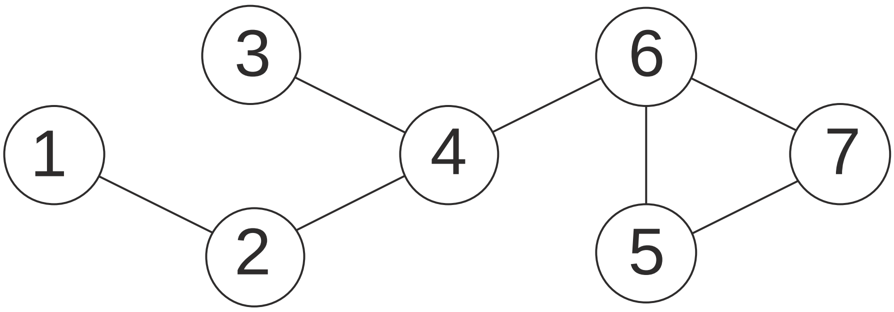
```
Color J diamonds have lowest value (coeff = -0.51) [1 Mark]
Significance better than 0.0001 [1 Mark]
```

(f) Comment on the reliability of the model as a whole giving reasons.

```
Reliability of model is high overall:
Multiple R-squared = 0.94,
p-value very small,
median residual close to 0. [1 Mark each up to 2 Marks]
```

Networks: 10 Marks

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



(a) Calculate the **betweenness centrality** for nodes 4 and 6.

Node(4) betweenness = 11 [1 Mark]
(It is in the following geodesics: 1-5, 1-6, 1-7, 2-5, 2-6, 2-7, 3-5, 3-6, 3-7, 1-3, 2-3.)
Node(6) betweenness = 8 [1 Mark]
(It is in the following: 1-5, 1-7, 2-5, 2-7, 3-5, 3-7, 4-5, 4-7.)

(b) Calculate the **closeness centrality** for nodes 4 and 6.

Node(4) closeness = 1/9 [1 Mark] (Since sum of shortest paths to others = 2 + 1 + 1 + 1 + 2 + 2.)
Node(6) closeness = 1/10 [1 Mark] (3 + 2 + 2 + 1 + 1 + 1)

(c) Calculate the **degree** of nodes 4 and 6.

|4| = 3 [1 Mark]
|6| = 3 [1 Mark]

(d) Giving reasons based on your results in Parts a – c, which node is most **central** in the network?

Node(4) is most central [1 Mark] It has the greatest betweenness centrality and closeness centrality [1 Mark]

(e) Write down the adjacency matrix for the network.

$$\begin{array}{c|ccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 7 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

Correct form [1 Mark], Correct values [1 Mark]

Naïve Bayes: 4 Marks

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

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

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

Test instance: (Age Range = young, Occupation = professor, Gender = F, Happy = ?)
p(Happy = yes) 0.5
p(Happy = no) 0.5 [1 Mark] 用左加，應該用乘

YES		P(young/yes)	P(professor/yes)	P(F/yes)	$P(C_j) \times P(A_1 \mid C_j) \times P(A_2 \mid C_j) \times \ldots \times P(A_n \mid C_j)$
p(yes)	0.5	0.250	0.250	0.500	0.016
NO		P(young/no)	P(professor/no)	P(F/no)	$P(C_j) \times P(A_1 \mid C_j) \times P(A_2 \mid C_j) \times \ldots \times P(A_n \mid C_j)$
p(no)	0.5	0.250	0.250	0.750	0.023

Correct calculations [1 Mark]

So classify as Happy = No [1 Mark or H]

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

NUM		P(young/yes)	P(professor/yes)	P(F/yes)	$P(C_j) \times P(A_1 \mid C_j) \times P(A_2 \mid C_j) \times \ldots \times P(A_n \mid C_j)$
p(yes)	0.5	0.250	0.250	0.500	0.016
DENOM		P(young)	P(professor)	P(F)	P(A1)×P(A2)× … ×P(An)
		0.250	0.250	0.625	0.039

So p(yes|attributes) = 0.016/0.039 = 0.41 [1 Mark or H]

Visualisation: 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?

Appropriate main graphic [1 Mark] with explanation. [1 Mark]

Mapping of variables to attributes in the graphic and/or data reduction (summary) as appropriate with explanation. [1 Mark each up to 4 Marks]

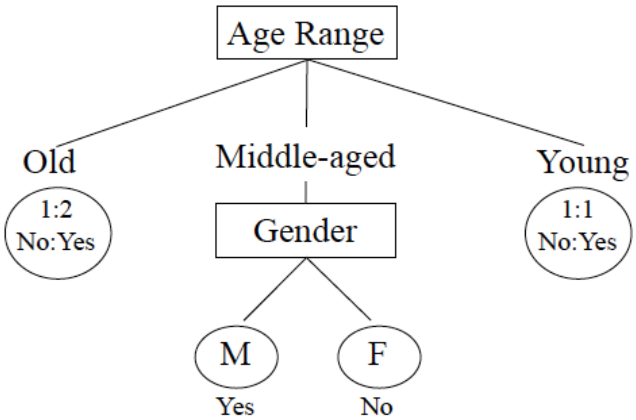
For example, one approach would be a heat map with time intervals on the x axis (perhaps every 10 or 50 years depending on range) and continents or countries on the y axis (depending on how many countries there are). Each cell could then be coloured for average age of death. You could either have two heat maps (male/female) or interleave cells so that m/f for each time period were adjacent.

Decision Trees: 10 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.



(a) Using the decision tree generated from the data provided, assuming a required confidence level greater than 60% to classify as 'Happy', what is the predicted classification for the following instances:

Instance 1: (Age Range = Young, Occupation = Professor, Gender = F, Happy = ?)

Instance 2: (Age Range = Old, Occupation = Professor, Gender = F, Happy = ?)

Instance 1: Happy = No, because confidence for Happy = Yes is 50%, which is less than required confidence level. [1 Mark]

Instance 2:Happy = Yes, because confidence for Happy = Yes is 66.67%, which is greater than required confidence level. [1 Mark]

(b) Is it possible to generate a 100% accurate decision tree using this data? Explain your answer.

Instances 5 and 8 have identical decision attributes, but belong to different classes, so No (Old, Tutor, F = Yes; Old, tutor, F = No). Therefore a 100% accurate decision tree could not be generated from this data. (Or equivalent) [1 Mark]

(c) Explain how the concept of entropy is used in some decision tree algorithms.

Information gain is used in the ID3 algorithm to determine which attribute to split on. Information gain calculates the reduction in entropy when splitting on a specific attribute and chooses the attribute which gives the greatest reduction in entropy or greatest information gain. (Or something similar) [1 Mark]

(d) Do you think entropy was used to generate the decision tree above? Explain your answer.

The Occupation attribute appears more homogeneous in terms of the class attribute Happy than the Age attribute. (Or Similar) [1 Mark] Therefore, no, entropy was not used. (or similar) [1 Mark]

(e) What is the entropy of "Happy"?

50:50 Yes:No = 1 by inspection. [1 Mark]

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

$E(2:1) = -\frac{2}{3} \cdot log_2\left(\frac{2}{3}\right) - \frac{1}{3} \cdot log_2\left(\frac{1}{3}\right) = 0.9184$ [1 Mark]

$Gain(S, AgeRange) = E(S) - \left(\frac{3}{8}0.9184 + \frac{3}{8}0.9184 + \frac{2}{8}1.0\right)$

$Gain(S, AgeRange) = 1 - (0.9388) = 0.0612$ [1 Mark]

(f) Explain why some decision tree algorithms are referred to as greedy algorithms.

Decision tree algorithms always choose the best option to branch on at each step without taking into account future choices. Is never able to back track in order to improve the final solution. [1 Mark]

ROC and Lift: 10 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	20%+	80%+
c1	0.9	1	1	1
c2	0.8	1	1	1
c3	0.7	0	1	0
c4	0.7	1	1	0
c5	0.6	1	1	0
c6	0.5	1	1	0
c7	0.4	0	1	0
c8	0.4	1	1	0
c9	0.2	0	1	0
c10	0.1	0	0	0

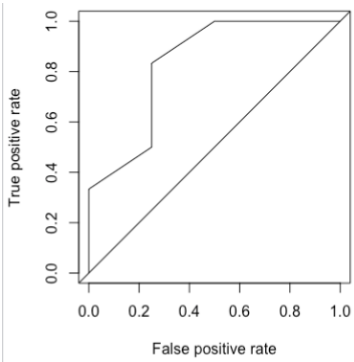
(a) Calculate the **True Positive Rate** and the **False Positive Rate** when a confidence level of 20% is required for a positive classification.

TP = 6, FP = 3, TN = 1, FN = 0. All correct 1 Mark
TPR = 6/(6+0) = 1, FPR = 3/(3+1) = 0.75. All correct 1 Mark

(b) Calculate the True Positive Rate and the False Positive Rate when a confidence level of 80% is required for a positive classification.

TP = 2, FP = 0, TN = 4, FN = 4. All correct 1 Mark
TPR = 2/(2+4) = 0.33, FPR = 0/(0+4) = 0. All correct 1 Mark

(c) The ROC chart for the previous question is shown below. Comment on the quality of the model overall. Give a single measure of classifier performance.



Exact = 0.83 accept 0.7 – 0.9. [1 Mark]
Classifier is good. [1 Mark]

(d) What is the lift value if you target the top 40% of customers that the classifier is most confident of?

P(true) = 6/10, for top 40% P(true) = 3/4 [1 Mark]
Lift = (3/4) / (6/10) = 1.25 [1 Mark or H]

(e) Explain what the value of lift means in the previous question.

Lift is the increase in the response rate over randomly selection [1 Mark] by choosing those you are most confident of. [1 Mark]

Clustering: 10 Marks

(a) What does the 'k' refer to in k-means clustering. Who/what determines the value of k?

K is the number of clusters. [1 Mark] This is pre-defined by the user before running the algorithm. [1 Mark]

(b) Describe the steps involved with k-means clustering.

1. Define the number of clusters required, k. [1 Mark]
2. Declare k centroids. [1 Mark]
3. Assign each data point to the closest centroid; 4. Re-calculate the centroids. [1 Mark]
5. Repeat 3 and 4 until the cluster centroids do not change. [1 Mark]

(c) Are clustering algorithms supervised or unsupervised learning algorithms? Explain.

They are unsupervised algorithms designed to find groupings of similar instances. [1 Mark]
Unlike classification, there is no 'class' attribute that can be used to help determine the clusters. [1 Mark]

(b) Is the k-means clustering algorithm a partitional or hierarchical clustering algorithm? Explain your answer.

k-Means is partitional. [1 Mark] There is no hierarchy from which clusters can be chosen. The number of clusters cannot be changed, once set. [1 Mark]

Text Analytics: 10 Marks

9 (a) Explain what is meant by the 'bag of words' approach to text mining.

Each document in the collection is assumed to be just a set of words and it is the entire collection of words that is used in the analysis. [1 Mark]
The semantics or meaning of the text in the documents is not considered in the 'bag of words' approach. [1 Mark]

(b) What is the main problem associated with the bag of words approach? Provide an example.

The main problem is that semantics are not considered and two documents that mean quite different things, but contain the same words, will be considered to be similar. [1 Mark]
Example:
- while licking their ice creams, the children chased the dog
- the dog chased the children and licked their ice creams

[1 Mark]

(c) Describe an application where text mining could be used, giving an example of how it would be applied.

Grouping articles by similar content (or similar). [1 Mark]
For example, job applications, tweets, emails etc. [1 Mark]

(d) 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 choir sang loudly. }
Doc2 = { The boys were singing in church. }
Doc3 = { The boy asked to sing a song. }

(choir, sing-, loud-)
(boy, sing-, church) [Tokenise and stop words 1 Mark]
(boy, ask, sing- song). [Stemming and overall format 1 Mark]

(e) Construct the term document frequency matrix for the processed text documents above. [2 Marks].

	ask	boy	choir	church	loud-	sing-	song
Doc1	0	0	1	0	1	1	0
Doc2		1	0	1	0	1	0
Doc3	1	1	0	0	0	1	1

Matrix correct format: words = cols, docs = rows [1 Mark]
Indicators are correct [1 Mark or H]

Ensemble Methods 4 Marks

(a) Describe the main similarities of the three ensemble classifiers (bagging, boosting and random forests) studied.

Create multiple data sets by resampling or cloning [1 Mark] Build multiple classifiers [1 Mark] Combine classifiers (ave or vote) [1 Mark up to a total of 2]

(b) How do boosting and random forests differ from bagging?

Boosting re-weights attributes to favour hard to classify cases. [1 Mark] Random Forests varies the attributes used in samples as well. [1 Mark]

Dirty Data 6 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.

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 & C	Love the Avenger. By the author of â€œ	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 Henr
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Henr
519		London	1872	The Author	Lagonells. By the author of Darmayne (F	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. - Write
874		London]	1676		A Warning to the inhabitants of England	RemaÊ¿.	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 hedendaagscl		WITKAMP, Pieter Ha
2836		Savona	1897		Cronache Savonesi dal 1500 al 1570 ... A	ABATE, Giovanni Ag	ASSERETO, Giovanni
2854		London	1865	E. Moxon & Co.	See-Saw; a novel ... Edited [or rather, w	ABATI, Francesco.	READE, William Win
2956		Paris	1860-63		GeÌ▯odeÌ▯sie 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 geograÌ▯fica, civil y politica de	ABBAD Y LASIERRA,	ACOSTA Y CALBO, Jo
3131		New York	1899	W. Abbatt	The Crisis of the Revolution, being the s	ABBATT, William.	ANDREÌ▯, John - Ma
4598		Hull	1814	The Author	Peace: a lyric poem. [With prefatory add	ABBOTT, Thomas Ea	WRANGHAM, Franci
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chri		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdollatiphi HistoriÃ¦ Ã†gypti compen		WHITE, Joseph - Car
5382		London	1847, 48 [1846-48]	Punch Office	The Comic History of England ... With ...	A'BECKETT, Gilbert A	LEECH, John - Artist
5385	[Another edition.] Il	London	[1897?]	Bradbury, Agnew &	[The comic history of England ... With tw	A'BECKETT, Gilbert A	LEECH, John - Artist
5389	[Another edition.]	London	[1897?]	Bradbury, Agnew &	[The Comic History of Rome ... Illustrate	A'BECKETT, Gilbert A	LEECH, John - Artist
5432		Milano	1893		Signa: opera in tre atti [founded on the	A'BECKETT, Gilbert A	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 -

Most of these are instances of incorrect data, although many records are incomplete also. [1 Mark each up to maximum 6]. 1 = incorrect/duplicate (has publisher and place in same cell. 2 = incorrect/duplicate etc, 3 = incorrect/inaccurate using abbreviation for "Oxford", 4 = incorrect/inaccurate etc.

Formulas and references

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:

$$Entropy(S) = -P_{C1}log_2(P_{C1}) - P_{C2}log_2(P_{C2})$$

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

where $C1$ and $C2$ are the two classes. P_{C1} and P_{C2} are the probability of being in Class 1 or Class 2 respectively. 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 $Gain(S, A)$ of an attribute A relative to a collection of examples, S, with v groups having $|S_v|$ elements is:

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

Networking

Closeness Centrality: $C_{CL}(v) = \frac{1}{\sum_{u \in V} dist(u,v)}$

Betweenness Centrality: $C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$,

where (s, t) is the number of shortest paths between s and t.
$(s, t|v)$ is the number of shortest paths between s and t passing through v

Density: $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2}$,

where $|E_g|$ is number of edges, $|V_g|$ is number of vertices

Clustering coefficient: $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau 3(g)}$,

where $3\tau_{\Delta}(g)$ is number of triangles, $\tau 3(g)$ is number of connected triples

Naïve Bayes'

$For\ events\ A_1, A_2, \dots, A_n\ and\ event\ C$, classification probability is

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

For Bayesian classification, a new point is classified to C_j if $P(C_j) * \mathrm{P}(A_1|C_j) * \mathrm{P}(A_1|C_j) * \dots * \mathrm{P}(A_n|C_j)$ is maximised.

Naïve Bayes assumes $P(A \cap B) = P(A) * P(B)$ etc.

Cosine or normalised dot product

For documents i and j with terms w

$$Sim(D_i, D_j) = \frac{\sum_{t=1}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^{N}(w_{it})^2 * \sum_{t=1}^{N}(w_{jt})^2}}$$

ROC

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$