

**CS9102**

**THE UNIVERSITY OF WARWICK**

**MEng Examinations: Summer 2017**

**CS910: Foundations of Data Analytics**

---

---

**Time allowed: 2 hours.**

Answer **SIX** questions only: **ALL THREE** from Section A and **THREE** from Section B.

Read carefully the instructions on the answer book and make sure that the particulars required are entered on **each** answer book.

Only calculators that are approved by the Department of Computer Science are allowed to be used during the examination.

---

---

---

**Section A**      Answer **ALL** questions
 

---

1. The following questions relate to the log-log plot.

[10]

- (a) Give a function  $f(x)$  which does not appear as linear on a log-log plot. [3]
- (b) Give a function  $f(x)$  which appears as linear on a log-linear plot (the  $x$ -axis uses a logarithmic scale and the  $y$ -axis uses a linear scale). [3]
- (c) Consider the function  $f(x)$  shown on the log-log plot from Figure 1. Give the expression of  $f(x)$ . [4]

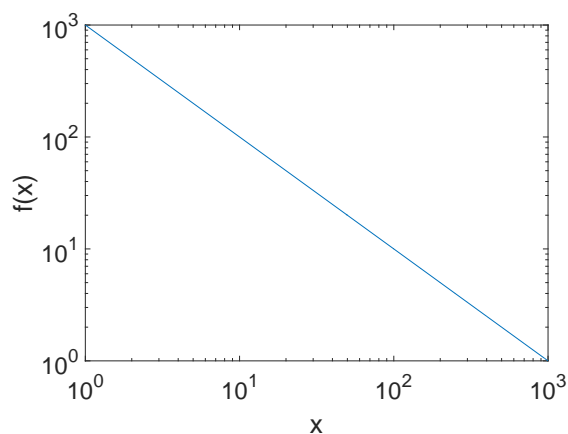


Figure 1: The log-log plot of  $f(x)$

**Solution:** *Comprehension – requires student to show understanding of concepts*

(a)

$$f(x) = e^{-x}$$

(b)

$$f(x) = -\log x + 3$$

(c) The function  $f(x)$  must satisfy

$$\log f(x) = -\log x + 3,$$

which yields

$$f(x) = 10^3 x^{-1}.$$

---

2. Consider the strings  $s_1$ ="saturday" and  $s_2$ ="sunday".

[10]

- (a) Compute the Hamming distance between  $s_1$  and  $s_2$ . [2]

- (b) Compute the (text) edit distance between  $s_1$  and  $s_2$ . [2]  
 (c) Could the cosine similarity distance between  $s_1$  and  $s_2$  be 0? Justify! [3]  
 (d) Could the cosine similarity distance between  $s_1$  and  $s_2$  be different than 0? Justify! [3]

**Solution:** *Application – student needs to apply techniques they have learned*

- (a) Undefined because  $s_1$  and  $s_2$  have different lengths.  
 (b) 3  
 (c) Yes. We could consider two dimensions corresponding to the two words. The encodings will then be  $[1\ 0]$  and  $[0\ 1]$  and hence the cosine similarity would be 0.  
 (d) Yes. We could consider multiple dimensions, each corresponding to an alphabetical letter. Given the overlap between  $s_1$  and  $s_2$  in terms of letters, the cosine similarity would be non zero.

3. Consider  $n$  paired observations  $(x_i, y_i)$  of some random variables  $X$  and  $Y$ . [20]  
 (a) Provide a full derivation of a linear regression model

$$y = ax + b$$

using the principle of least squares. The answer should include the expressions of the parameters  $a$  and  $b$  in terms of  $X$  and  $Y$ . [10]

- (b) Fully simplify the sum of residuals [3]

$$\sum_{i=1}^n (y_i - ax_i - b)$$

for the values of  $a$  and  $b$  obtained in (a).

- (c) Assume that your data satisfies  $x_i \geq 0$  and  $y_i = x_i^2 \forall i = 1 \dots n$ . Which of the following two regression models would best fit the data?

Model 1:  $y = ax + a^2x$

Model 2:  $y = ax$

Justify! (Note that in both models  $a$  is a regression parameter.) [7]

**Solution:** *Bookwork – primarily requires recollection of taught concepts*

- (a)

$$a = \frac{\text{cov}(X, Y)}{\text{Var}[X]}, \quad b = E[Y] - aE[X].$$

The derivations were shown in the slides.

(b)

$$\begin{aligned}\sum_{i=1}^n (y_i - ax_i - b) &= nE[Y] - \frac{\text{cov}(X, Y)}{\text{Var}[X]} nE[X] - n \left( E[Y] - \frac{\text{cov}(X, Y)}{\text{Var}[X]} E[X] \right) \\ &= 0.\end{aligned}$$

(c) Both models fit the data equally well. The reason is that the solutions for the linear regressions would be identical. Concretely, Model 2 yields a positive value for  $a$ , which could be further expressed in terms of  $b + b^2$ .

---

**Section B**      Choose **THREE** questions.
 

---

4. The following questions relate to random variables.

[20]

- (a) Give a random variable whose median is strictly smaller than its mean. Justify! [2]  
 (b) Give a random variable whose median is strictly smaller than its mode. Justify! [3]  
 (c) Prove that for any random variable  $X$  the following holds [3]

$$\text{Var}[X] = E[X^2] - (E[X])^2.$$

- (d) Give a random variable  $X$  with at least two values, each having positive probability, such that [5]

$$E\left[\frac{1}{X}\right] = \frac{1}{E[X]}.$$

- (e) Prove that for any random variables  $X$  and  $Y$  such that  $E[X] = E[Y] = 0$  the following holds [7]

$$(E[XY])^2 \leq E[X^2]E[Y^2].$$

**Solution:** *Comprehension – requires student to show understanding of concepts*

- (a) The exponential distribution with density  $f(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$  and  $\lambda > 0$ . The median is  $\frac{\ln 2}{\lambda}$  whereas the mean is  $\frac{1}{\lambda}$ .  
 (b) Take the random variable

$$X = \begin{pmatrix} 1 & 2 & 3 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}.$$

The median is 2 whereas the mode is 3.

- (c) One can write

$$\begin{aligned} \text{Var}[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2. \end{aligned}$$

- (d) Take the random variable

$$X = \begin{pmatrix} -1 & \frac{1}{2} & 2 \\ \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix},$$

for which

$$E[X] = -\frac{1}{9} + \frac{2}{9} + \frac{8}{9} = 1 = -\frac{1}{9} + \frac{8}{9} + \frac{2}{9} = E\left[\frac{1}{X}\right].$$

(e) For some real number  $a$  one has

$$0 \leq E[(X - aY)^2] = E[X^2] - 2aE[XY] + a^2E[Y^2].$$

Choose  $a = \frac{E[XY]}{E[Y^2]}$  and the claim follows.

5. Consider running a k-NN classifier using Euclidean distance on the data set from Figure 2, whereby each point belongs to one of two classes: + and o. [20]

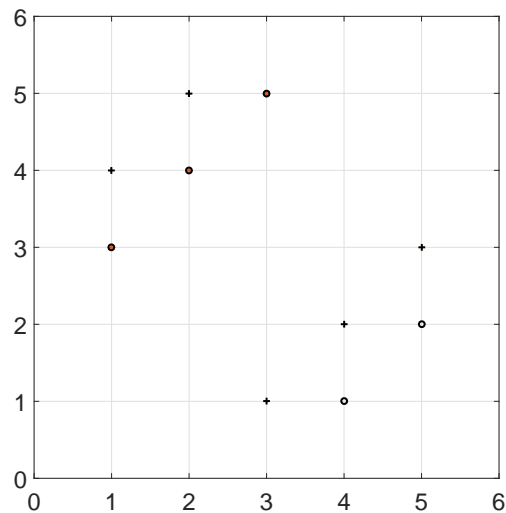


Figure 2: Points belonging to two classes

- (a) What is the 10-fold cross validation error when  $k = 1$ ? [5]  
 (b) Which of the values  $k \in \{3, 4, 5, 9\}$  yields the minimum number of 10-fold cross validation errors? [7]  
 (c) Give a distance metric, instead of the Euclidean distance, such that the 10-fold cross validation error of 1-NN is  $\frac{4}{10}$ . [8]

**Solution:** *Application – student needs to apply techniques they have learned*

- (a) Each point is misclassified and hence the error is 1.  
 (b) For each  $k \in \{3, 5, 9\}$ , each point is misclassified. When  $k = 4$  some points may be correctly classified, depending on how 4-NN handles ties.  
 (c) The inverse of the Euclidean distance.

6. Consider the data from the table below in which the attribute A is binary, whereas the values  $a_i \in \{Y, N\}$  are unknown. [20]

Name	Sex	A
Alex	M	$a_1$
Mary	F	$a_2$
Alex	F	$a_3$
Alex	F	$a_4$
John	M	$a_5$
Zoe	F	$a_6$
Nina	F	$a_7$
Dan	M	$a_8$

- (a) Ignoring the attribute A, what would a Naïve Bayes Classifier predict on the input Alex, i.e., M or F? Justify! [4]
- (b) Again, ignoring the attribute A, build a decision tree classifier which would predict M on the input Alex. [5]
- (c) Determine some values  $a_i$  such that a Naïve Bayes Classifier would predict M on the input (Alex,Y) . [6]
- (d) What is the key advantage of the Naïve Bayes Classifier over decision tree classifiers? What is its weakness? [5]

**Solution:** *Application – student needs to apply techniques they have learned*

- (a) We have

$$P(\text{Alex}|\text{M})P(\text{M}) = \frac{1}{3} \frac{3}{8} = \frac{1}{8}$$

and

$$P(\text{Alex}|\text{F})P(\text{F}) = \frac{2}{5} \frac{5}{8} = \frac{2}{8}$$

and hence the prediction is F.

- (b) Choose the attribute Name at the root and the binary split: Alex and the rest of names. Label the leaf with the minority class. Note that the leaf Alex will be labelled with M.

- (c) Consider

Name	Sex	A
Alex	M	Y
Mary	F	N
Alex	F	Y
Alex	F	Y
John	M	Y
Zoe	F	N
Nina	F	N
Dan	M	Y

in which case we would have

$$P(Alex|M)P(Y|M)P(M) = \frac{1}{3} \frac{3}{8} = \frac{1}{8}$$

and

$$P(Alex|F)P(Y|F)P(F) = \frac{2}{5} \frac{5}{8} = 0$$

and hence the prediction is M.

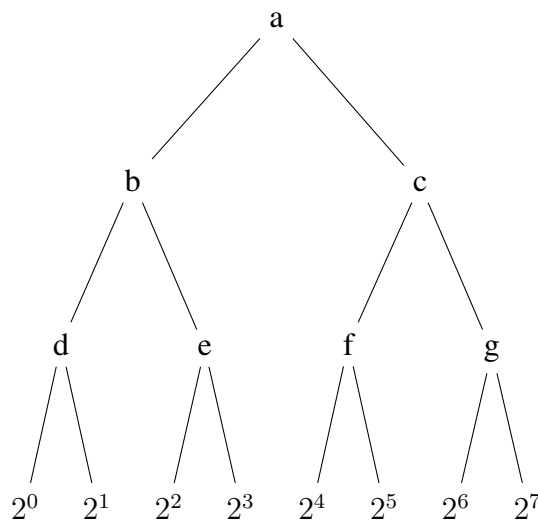
(d) The ability to account for correlation amongst attributes.

The assumption that each attribute is conditionally independent of others.

7. Consider the points  $2^0, 2^1, 2^2, \dots, 2^{2^n-1}$  for some  $n \geq 1$ .

[20]

- (a) Sketch the clustering trees produced by hierarchical clustering with Euclidean distance and the following inter-cluster distances: single-link, complete-link, and average-link. (Recall that for single-link  $d(X, Y) := \min_{x \in X, y \in Y} d(x, y)$ , whereas for complete and average-link the ‘min’ is replaced by ‘max’ and ‘avg’, respectively.) [12]
- (b) Replace the Euclidean distance metric from (a) by another distance *function* (which does not necessarily have to obey the *metric* rules) such that hierarchical clustering with single-link would produce a full binary tree (each node, except for the leaves, has exactly two children). For instance, if  $n = 3$ , the tree below would be produced. (Note: for two points  $x$  and  $y$  you need to construct a function  $d(x, y)$  obeying the requirements).



**Solution:** Application – student needs to apply techniques they have learned

(a) All three inter-cluster distances yield the figure below.

(b)  $d(a, b) = \frac{\max(a, b)}{\min(a, b)}$



