CS9101

THE UNIVERSITY OF WARWICK

MSc 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]

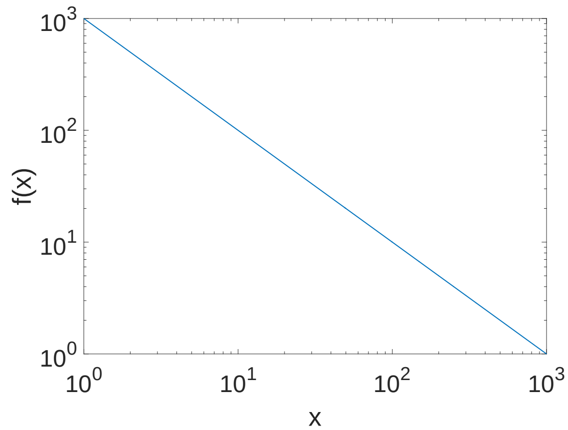


Figure 1: The log-log plot of $f(x)$

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]

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

$$y = ax$$

   using the principle of least squares. The answer should include the expression of the parameter a in terms of X and Y. [10]

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

$$\sum_{i=1}^{n}(y_i - ax_i)^2$$

for the value of a obtained in (a).

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

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

Justify! (Note that a and b are regression parameters.) [7]

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

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

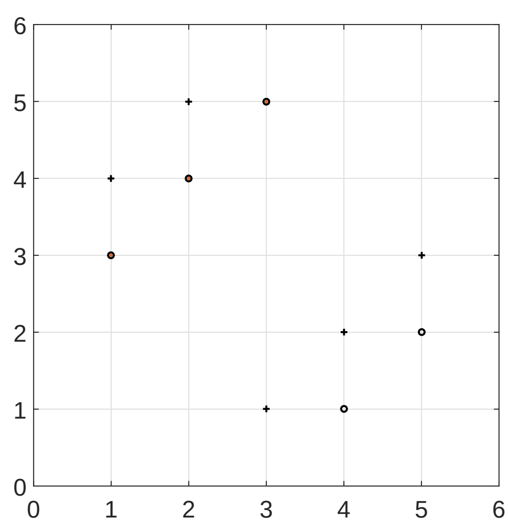


Figure 2: Points belonging to two classes

   (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]

---

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]

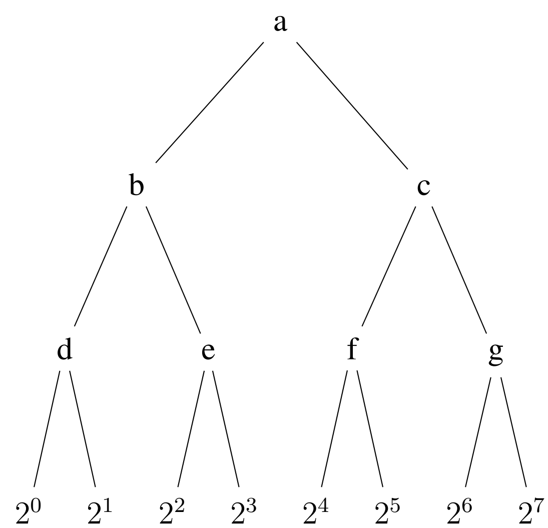
---

7. Consider the points $2^0, 2^1, 2^2, \ldots, 2^{2^n-1}$ 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 $d(X, Y) := \min d(x \in X, y \in Y)$, 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).

[8]



End