CS9101

THE UNIVERSITY OF WARWICK

MSc Examinations: Summer 2018

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. Consider the following data points $\{3, 5, 3, 3, 7, 5\}$. [10]

   (a) Sketch the frequency plot. [2]

   (b) Sketch the frequency/rank plot. [2]

   (c) (independent of (a) and (b)). Assume some data set which is likely to be heavy-tailed. To fit a heavy-tailed distribution you could plot on a log-log scale either the frequency plot or the frequency/rank plot. What would be your choice and why? [6]

2. The following questions concern the q-q plot. [10]

   (a) Consider the data sets $X_1 = \{3, 5, 3, 3, 7, 5\}$ and $X_2 = \{50, 70, 30, 30, 30, 50\}$. Sketch the q-q plot of X_1 and X_2. [3]

   (b) Consider the data sets $Y = \{y_1, y_2, \ldots, y_m\}$ and $Z = \{z_1, z_2, \ldots, z_n\}$ for some $m, n \geq 1$. Prove that the q-q plot of Y and Z exhibits a non-decreasing behavior. [7]

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 $y_i = ae^{bx_i}\ \forall i = 1 \ldots n$, where a and b are unknown parameters. Can linear regression be used to fit a and b? What if both parameters a and b were known? [7]

Section B Choose **THREE** questions.

4. Consider a random sample $Y_1, Y_2, \ldots, Y_n$ of a random variable Y with expectation $\mu := E[Y]$ and variance $\sigma^2 = \mathrm{Var}[Y]$. [20]

   (a) Prove that
   $$Z_\theta := \overline{Y} := \frac{Y_1 + \cdots + Y_n}{n}$$
   is an unbiased estimator for μ. [4]

   (b) Prove that
   $$Z_\theta := \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$
   is an unbiased estimator for σ^2, where $\overline{Y}$ was defined in (a). [8]

   (c) If Y is nonnegative prove that
   $$\mathbb{P}(Y \geq y) \leq \frac{E[Y]}{y}$$
   for all $y > 0$. [8]

5. Consider the following data set with three attributes (X_1, X_2, and C, the last one being the target attribute (class)): [20]

X_1	X_2	C
1	1	1
0	0	1
0	1	0
1	0	1

   (a) Does the data satisfy the Naïve Bayes independence assumption? [5]

   (b) Partition the data set into Training and Test data sets such that the accuracy of the Naïve Bayes classifier (on the Test set) is 0. [7]

   (c) Provide a data set with 4 distinct records, and 4 binary attributes (X_1, X_2, X_3, and C, the last one being the target attribute), such that data satisfies the Naïve Bayes independence assumption. [8]

   Note: all answers must be briefly justified!

6. Consider a set of points in the Euclidean space $X_1, X_2, \ldots, X_n$. Recall that the objective of the k-means clustering algorithm is to find k points $C_1, C_2, \ldots, C_k$ minimizing [20]
   $$\sum_{i=1}^{N} \min_{j\in\{1,2,\ldots,k\}} \|X_i - C_j\|_2 \,,$$
   where $\|\cdot\|_2$ denotes the standard Euclidean distance metric.

Continued

(a) Is it a good idea to redefine the k-means clustering by minimizing after k as well? In other words, the new objective would be to minimize

$$\min_k \sum_{i=1}^{N} \min_{j \in \{1,2,\ldots,k\}} \|X_i - C_j\|_2 \, .$$

[5]

(b) Assume the input points $\{1, 3, 10, 14\}$ and $k = 3$. Does the Lloyd's k-means clustering algorithm *always* result in an optimal clustering assignment on such input? [7]

(c) Assume $m + n$ distinct points in the 1-dimensional Euclidean space, and the optimal $2 - means$ clustering $\{X_1, X_2, \ldots, X_m\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ where $X_1 < X_2 < \cdots < X_m$ and $Y_1 < Y_2 < \cdots < Y_n$. Provide a necessary condition for such a clustering. [8]

Note: all answers must be briefly justified!

7. Consider the directed graph $(\{1, 2, 3, 4\}, \{(1, 2), (2, 4), (1, 3), (3, 4), (4, 1)\})$ representing links between four web-pages (e.g., $(1, 2)$ means that there is a link from page 1 to page 2). [20]

(a) Write the transition matrix A and iterate the derivation of the importance (column) vector r_t, for $t = 2, 3, 4$, in a simplified version of PageRank whereby $r_t = A * r_{t-1}$ for all $t \geq 2$; assume that the initial importance vector is $r_1 = (1/4,\ 1/4,\ 1/4,\ 1/4)^T$. [6]

(b) What is the key shortcoming of the simplified version of PageRank from (a)? How can you fix it? [9]

(c) Describe in one sentence the main objective of PageRank. Further describe in one sentence how this objective is achieved. [5]

End