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]

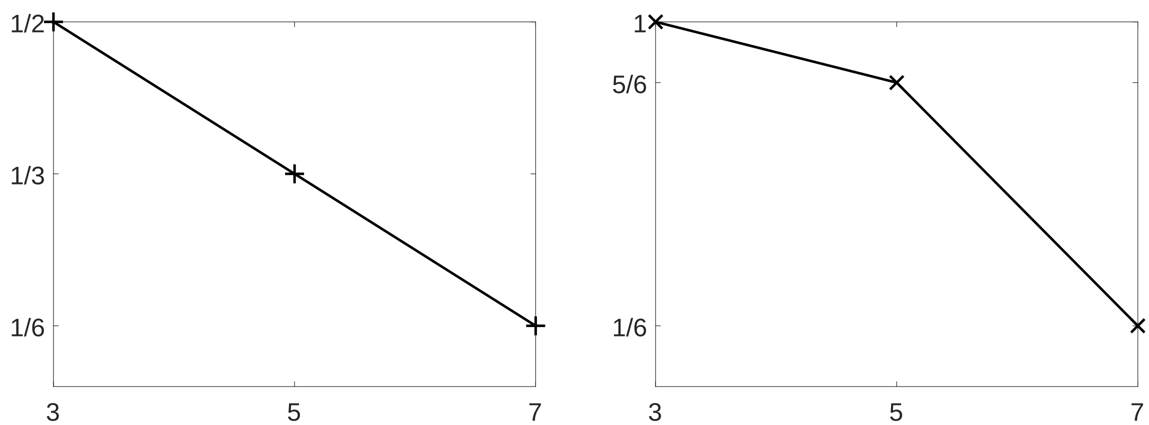
   (c) (independent of (a) and (b)). Assume some data set which is likely to be heavy-tailed. To fit a heavy-tailed distribution you could plot on a log-log scale either the frequency plot or the frequency/rank plot. What would be your choice and why? [6]

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

(a) See Figure 1.(a).

(b) See Figure 1.(b).

(c) The frequency/rank plot because the CCDF is a monotonous function. Moreover, unlike the PDF function, the CCDF displays a significantly lower variability in the tail.



(a) frequency plot

(b) frequency/rank plot

Figure 1: Frequency (PDF) and Frequency/Rank (CCDF) plots for Problem 1

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]

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

(a) See Figure 2.

(b) Take two points (a_1, b_1) and (a_2, b_2) from the q-q plot, such that $a_1 \le a_2$. According to the CDF's monotonicity:

$$a_1 \le a_2 \Rightarrow \mathbb{P}(X \le a_1) \le \mathbb{P}(X \le a_2)$$

which implies from the definition of the Q-Q plot that

$$\mathbb{P}(Y \le b_1) \le \mathbb{P}(Y \le b_2) \Rightarrow b_1 \le b_2 \ .$$

Therefore the q-q plot/curve is non-decreasing.

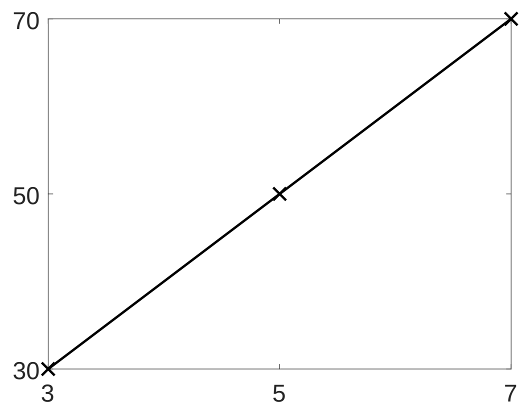


Figure 2: The q-q plot

---

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]

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

(a)

$$a = \frac{E[XY]}{E[X^2]} \ .$$

The derivations were shown in the slides.

(b)

$$n\left(E[Y^2] - \frac{E[XY]^2}{E[X^2]}\right) \ .$$

The derivations were shown in the slides.

(c) Yes. Denoting $t_i = \ln y_i$ we have the linear regression model

$$t_i = \ln a + bx_i \ .$$

If a and b were known then there is no need for a regression model since the data is fully described.

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 $\sigma^2 = \text{Var}[Y]$. [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]

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

(a)
$$E\left[Z_\theta\right] = E\left[\frac{Y_1 + \cdots + Y_n}{n}\right] = \frac{nE\left[Y_1\right]}{n} = E[Y]$$

(b)
$$\begin{aligned}
E\left[Z_\theta\right] &= \frac{1}{n-1} E\left[\sum_i Y_i^2 - 2\sum_i \overline{Y} Y_i + n\overline{Y}^2\right] = \frac{1}{n-1} E\left[\sum_i Y_i^2 - n\overline{Y}^2\right] \\
&= \frac{1}{n-1}\left(nE\left[Y^2\right] - nE\left[\overline{Y}^2\right]\right) = \frac{1}{n-1}\left(nE\left[Y^2\right] - n\left(\text{Var}\left[\overline{Y}\right] + E\left[\overline{Y}\right]^2\right)\right) \\
&= \frac{1}{n-1}\left(n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \sigma^2 \,,
\end{aligned}$$

(c) Denoting by $f(x)$ the density of Y we can write
$$\begin{aligned}
E[Y] &= \int_0^\infty x f(x) dx \\
&\geq \int_y^\infty x f(x) dx \\
&\geq \int_y^\infty y f(x) dx \\
&= y\mathbb{P}(Y \geq y)\,.
\end{aligned}$$

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!

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

(a) No, because

$$1 = \mathbb{P}(X_1 = 1, X_2 = 0|C = 1) \neq \mathbb{P}(X_1 = 1|C = 1)\mathbb{P}(X_2 = 0|C = 1) = \frac{2}{3}\frac{2}{3}$$

(b) Take the Test set as $(0, 1, 0)$ and the Training set as the rest, in which case

$$\mathbb{P}(C = 1)\mathbb{P}(X_1 = 0|C = 1)\mathbb{P}(X_2 = 1|C = 1) = 1\frac{1}{3}\frac{1}{3} > 0 = \mathbb{P}(C = 0) \cdot \ldots$$

and hence $(0, 1)$ is incorrectly classified as '1'.

(c)

X_1	X_2	X_3	C
1	0	1	1
0	1	1	1
1	1	1	1
0	0	1	1

Note that $\mathbb{P}(x_1, x_2, x_3|c = 1) = \frac{1}{4}$ for the four tuples in the set.

---

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.

(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!

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

(a) No, because the overall minimum would be zero, attained by letting $C_i = X_i$ for $i = 1, 2, \ldots, n$.

(b) No. Assume the initial choice is $\{1, 3, 10\}$ in which case the clusters would be $\{1\}$, $\{3\}$, and $\{10, 14\}$. The corresponding centroids would be $\{1, 3, 12\}$ and the clusters would not change; moreover, the objective would be $0 + 0 + 4 + 4 = 8$. However, the optimal (center) points would be $\{2, 10, 14\}$ which yield the objective $1 + 1 + 0 + 0 = 2$.

(c) $[X_1, X_m]$ and $[Y_1, Y_n]$ must be non-overlapping intervals. Assume without loss of generality that $X_1 < Y_1$, and assume by contradiction that $Y_1 < X_m$ ($Y_1 = X_m - \epsilon$ with $\epsilon > 0$). If C_1 and C_2 are the centers of the two optimal clusters, then a better clustering would be obtained by assigning Y_1 to the first cluster because

$$X_m - C_1 \le C_2 - X_m \Rightarrow (X_m - \epsilon) - C_1 < C_2 - (X_m - \epsilon) \Rightarrow Y_1 - C_1 \le C_2 - Y_1 \ .$$

This is a contradiction.

---

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 $t \ge 2$; 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]

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

(a) We have

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

The updated importance vectors are

$$r_2 = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{2} \end{pmatrix}, \; r_3 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{4} \end{pmatrix}, \; r_4 = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

(b) The problem is the periodic behavior, meaning that PageRank cannot identify the most important web-page. One solution is to create a self-loop (e.g., $(1, 1)$), in which case the iterative procedure from (a) would be guaranteed to converge.

(c) PageRank computes the importance of web-pages. This is achieved by leveraging the transition matrix.

Solution:

1) Comparing attributes of different cardinality and 2) testing whether some data has some theoretical distribution.