CS9102

THE UNIVERSITY OF WARWICK

MEng Examinations: Summer 2019

CS910: Foundations of Data Analytics

Time allowed: 2 hours.

Solve **FIVE** problems only: **BOTH** 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.

Continued

Section A Solve **BOTH** problems

1. The following questions concern the quantile-quantile (q-q) plot. [20]

   (a) State two common uses of the q-q plot (be brief!). [5]

   (b) Assume that you own two ice cream shops. Being away for an extended holiday and wondering about the current prices in the two stores you ask for the q-q plot. Assume that you receive the plot from Fig. 1.(a). Explain in one short sentence whether the plot is valid or not, and what can be deduced from the plot. [5]

   (c) Same question as above but for Fig. 1.(b). [5]

   (d) Same question as above but for Fig. 1.(c) (note that there is a single point in the plot). [5]

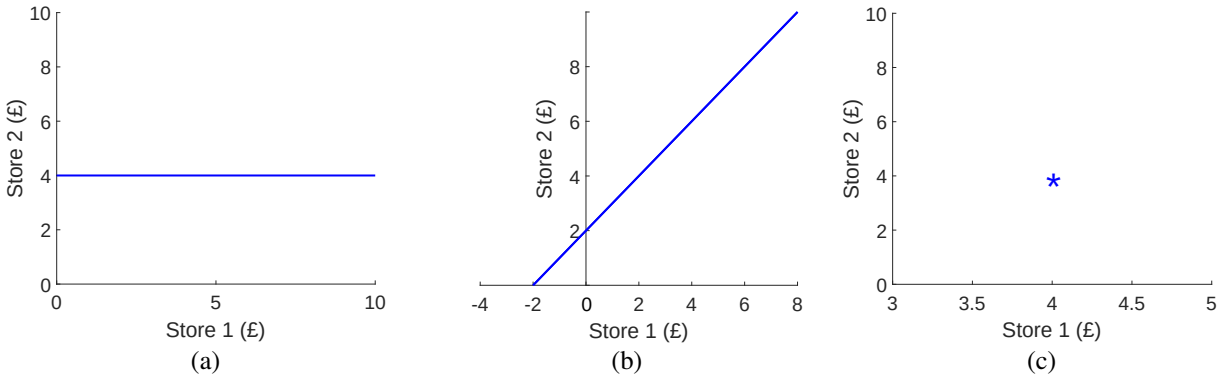

Figure 1: Q-Q plots

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

(a) Comparing two empirical distributions. Comparing theoretical and empirical distributions.

(b) Yes. Store 2 sells everything at the same price of $£4$.

(c) Yes. Store 1 sells some items at negative price.

(d) Yes. The store managers were lazy and computed a single quantile-quantile pair only.

2. 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 $y_i = ae^{2\ln x_i} + be^{\sqrt{x_i}}\ \forall i = 1 \ldots n$, where a and b are unknown parameters. Can linear regression be used to fit a and b? [7]

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

(a)

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

The derivations were shown in the slides.

(b)

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

(c) Yes, by making the transform $t_i = x_i^2$ and $s_i = e^{\sqrt{x_i}}$, in which case y_i linearly depends on t_i and s_i.

Continued

Section B Choose **EXACTLY THREE** problems.

1. The following questions concern probability concepts. [20]

   (a) Prove that
   $$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$$
   for any events A and B. [5]

   (b) You flip three fair coins. Assume that the outcomes of two of them are equal. Because the third coin is fair then the probability that the outcomes of all three coins are equal is 0.5. Do you agree? (Briefly justify your answer!) [5]

   (c) Assume that X and Y are independent and exponentially distributed random variables with the same parameter $\lambda > 0$. Justify whether the product XY is exponentially distributed as well. [10]

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

(a)
$$\begin{aligned} \mathbb{P}(A \cap B) &= 1 - \mathbb{P}(\overline{A} \cup \overline{B}) \\ &= 1 - \big(\mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B}) - \mathbb{P}(\overline{A} \cap \overline{B})\big) \\ &\geq 1 - (1 - \mathbb{P}(A) + 1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - 1 \end{aligned}$$

(b) No. The probability that they are all alike is $\frac{1}{4}$ (because $P(HHH) = P(TTT) = \frac{1}{8}$). Note that the information that two outcomes are the same is irrelevant.

(c) No. The tail is sub-exponential as it could be seen from
$$\mathbb{P}(XY > z) \geq \mathbb{P}(X > \sqrt{z}, Y > \sqrt{z}) = e^{-\lambda\sqrt{z}} e^{-\lambda\sqrt{z}} = e^{-2\lambda\sqrt{z}}$$

2. The following questions concern the ID3 classifier. (*Note: all answers must be briefly justified!*) [20]

   (a) Can you describe a data set with five attributes such that the entropy of one attribute is $\log \frac{1}{2}$ (the logarithm is in base 2). [5]

   (b) What is a key weakness of the ID3 algorithm? [5]

   (c) Provide a training data set with 3 binary attributes (X_1, X_2, C; the last one is the target attribute) and 2 distinct records, such that ID3 incorrectly classifies the (test) record $(0, 0, 1)$. [10]

Solution: *Comprehension – requires student to show understanding of concepts*; *Application – student needs to apply techniques they have learned*

(a) Impossible because the entropy is always positive.

(b) ID3 favors attributes with many values (e.g., "student id") for which reason overfitting is likely.

(c)

X_1	X_2	C
1	1	1
0	1	0

Note that ID3 chooses attribute X_1 and thus $(0, 0, 1)$ is misclassified.

---

3. The following questions concern logistic regression. (*Note: all answers must be briefly justified!*) [20]

   (a) Does logistic regression fall within the class of supervised learning algorithms? [5]

   (b) Assume that you have a data set in which the dependent (response) variable has three possible values. Is it possible to apply logistic regression in this situation? [5]

   (c) Consider the relation between an explanatory and dependent variable from Figure 2; the dependent variable measures some probability (e.g., of rain in a certain day, in which case the explanatory variable may be the time of day). Is it a good idea to use logistic regression in this situation? [5]

   (d) Provide a relation between an explanatory and dependent variable, either as a plot or as a formula, such that logistic regression achieves a root-mean-square-error of 0 (on some data satisfying your plot/formula). [5]

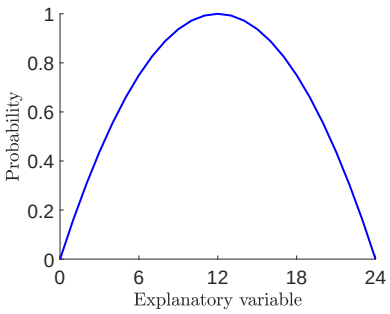

Figure 2: A relation between explanatory and dependent variables.

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

(a) Yes. Because it uses the training data/set to learn the parameters of the model.

(b) Yes. One could use the one-versus-all method.

(c) No, one should rather use a quadratic model, i.e., $Y = aX^2 + bx + C$ where X is the explanatory variable; the logistic regression curves are fundamentally different than quadratic models (see d))

(d) Take $y = \frac{1}{1+e^{-ax+b}}$, which falls in the class of logistic regression curves, and thus the fitting is perfect.

---

4. The following questions concern the K-means clustering algorithm with Euclidean distance. [20]
(*Note: all answers must be briefly justified!*)

   (a) Consider the set of points $X_1 = (1, 3)$, $X_2 = (2, 3)$, $X_3 = (3, 0)$, $X_4 = (5, 0)$, $X_5 = (5, 3)$, in two-dimensional Euclidean space, that are to be clustered in 2 clusters. Assume that we use the 2-means algorithm with X_2 and X_5 as initial centroids. How many iterations does 2-means need to converge and what are the resulting centroids and clusters? [6]

   (b) Give a set of 3 points in two-dimensional Euclidean space, two of which are the initial centroids, such that 2-means converges in exactly 2 iterations. [6]

   (c) Give a set of 4 points in two-dimensional Euclidean space, two of which are the initial centroids, such that 2-means converges in exactly 3 iterations. [8]

Solution: *Comprehension – requires student to show understanding of concepts*; *Application – student needs to apply techniques they have learned*

(a) 2-means converges in 1 iteration. The centroids will be $(2, 2)$ and $(5, 1.5)$. The clusters will be $\{X_1, X_2, X_3\}$ and $\{X_4, X_5\}$.

(b) Take the points $X_1 = (0, 0)$, $X_2 = (2, 0)$, $X_3 = (10, 0)$; the first two are the initial centroids. After 1 iteration the clusters will be $\{X_1\}$ and $\{X_2, X_3\}$. After the second iteration the clusters will be $\{X_1, X_2\}$ and $\{X_3\}$.

(c) Take the points $X_1 = (0, 0)$, $X_2 = (1, 0)$, $X_3 = (3, 0)$, and $X_4 = (8.9, 0)$; the first two are the initial centroids. After 1 iteration the clusters will be $\{X_1\}$ and $\{X_2, X_3, X_4\}$. After the second iteration the clusters will be $\{X_1, X_2\}$ and $\{X_3, X_4\}$. After the third iteration the clusters will be $\{X_1, X_2, X_3\}$ and $\{X_4\}$.

5. The following questions concern the Furthest Point Clustering (FPC) algorithm with Euclidean distance and $k = 2$ (i.e., 2 clusters are to be constructed). (*Note: all answers must be briefly justified!*) [20]

   (a) Prove that FPC yields an optimal clustering, in terms of the 2-center objective, provided that there are 3 points in two-dimensional Euclidean space. [6]

   (b) Give a set of 4 points in two-dimensional Euclidean space, one of which is the initial center, such that FPC results in an exact 2-approximation (the maximum diameter of the 2 clusters is twice as much as the optimal diameter). [6]

(c) Give a set of 4 points in two-dimensional Euclidean space, one of which is the initial center, such that FPC results in an exact 1.5-approximation (the maximum diameter of the 2 clusters is 1.5 larger than the optimal diameter). [8]

Solution: *Comprehension – requires student to show understanding of concepts*; *Application – student needs to apply techniques they have learned*

(a) Assume the points X_1, X_2, X_3; without loss of generality assume that $d(X_1, X_2) \geq d(X_2, X_3) \geq d(X_3, X_1)$. If either X_1 or X_2 are the initial centers then the other will be the next center, and thus the diameter is $\frac{d(X_1,X_3)}{2}$. If X_3 is the initial center then X_2 will be the next center, in which case again the diameter will be $\frac{d(X_1,X_3)}{2}$ (the optimal diameter).

(b) Take the points $X_1 = (0, 0)$, $X_2 = (2, 0)$, $X_3 = (4 - \varepsilon, 0)$, and $X_4 = (6, 0)$, and choose X_2 as the initial center. X_4 will be chosen as the second center and thus the clusters will be $\{X_1, X_2, X_3\}$ and $\{X_4\}$. The resulting diameter is 2 (by making ε arbitrarily small). However, the optimal diameter is 1 (in the case of the clustering $\{X_1, X_2\}$ and $\{X_3, X_4\}$.

(c) Take the points $X_1 = (\varepsilon, 0)$, $X_2 = (4, 0)$, $X_3 = (6 - \varepsilon, 0)$, and $X_4 = (8, 0)$, and choose X_2 as the initial center. X_4 will be chosen as the second center and thus the clusters will be $\{X_1, X_2, X_3\}$ and $\{X_4\}$. The resulting diameter is 3 (by making ε arbitrarily small). However, the optimal diameter is 2 (in the case of the clustering $\{X_1, X_2\}$ and $\{X_3, X_4\}$.

End