CS9100_C

THE UNIVERSITY OF WARWICK

Examination: Summer 2022

Paper Code: CS9100_C

Foundations of Data Analytics

Time allowed: 2 hours.

Exam Type: Standard Examination

Solve **FIVE** problems only. Answer **BOTH** questions from Section A and **THREE** questions 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 Answer **BOTH** questions.

1. Command line tools are often simple tools which can be initiated via a command line interface.

   (a) You discover a specific tool that you believe would be well suited for your data analysis task. Give one way, other than from the internet or from a book, that you could find more information regarding this specific command line tool. [1]

   (b) Give two benefits of command line tools over languages such as R or Python. [3]

   (c) Imagine you are given the following data set, which represents fictional, past students at the University of Warwick, stored in the file students.csv. Only the top 4 records are shown below, but you can assume the data consists of multiple thousand records. The attribute names are not included in the data file. If you have any assumptions about the format of the data, please state them.

Age	Sex	HighestQualification	GraduationYear	NativeCountry
42	M	BSc	2015	UK
35	F	PhD	2018	Germany
56	F	MSc	2014	India
24	M	PhD	2020	Egypt
...				

   Describe briefly (in words) the output of the following:

   i. tail students.csv [1]

   ii. cut -f1,3 -d, students.csv [1]

   iii. grep BSc students.csv | head [1]

   iv. grep y$ students.csv [1]

   v. sed 's/BSc/BEng/g' students.csv > revised_qual [2]

   vi. cat students.csv | uniq [2]

   (d) A soft drinks company wishes to improve their data analysis practices, and hires you to advise. They currently use spreadsheets for all of their analysis, which includes analysing customer and sales data. They are considering switching to R and/or Python. Discuss the advantages and disadvantages of remaining with spreadsheets, or instead using R and/or Python. [8]

2. This question concerns data basics. The data below represents the number of disease particles present in 15 patients, where each cell represents the number of disease particles found in an individual patient. We are concerned that our data is noisy.

286	314	454	486	512	590	591	596	650	691	710	840	982	1012	1118

(a) Name two smoothing methods that we could use to mitigate the effects of noise. [2]

(b) Apply Equal-Frequency Binning to divide the data into 3 bins, and then demonstrate how both of the smoothing methods described in part (a) could be applied to the 3 bins. [6]

(c) The following data represents fictional students at a university. It contains their percentage mark in an exam, as well as their salary five years after graduation.

	Exam Mark (%)	Salary (£)
1	84	56,000
2	43	39,000
3	68	38,000
4	80	50,000

i. Perform Min-Max normalisation on the two attributes, to ensure the minimum and maximum values of each attribute are 0 and 1 respectively. [6]

ii. The table below shows the Euclidean distance between each pair of student records before Min-Max normalisation was applied.

	1	2	3	4
1	0	17,000	18,000	6,000
2	17,000	0	1,000	11,000
3	18,000	1,000	0	12,000
4	6,000	11,000	12,000	0

Calculate the Euclidean distance between each pair of student records after Min-Max normalisation has been applied. What do you conclude about the effect of normalisation on this data? [6]

Continued

Section B Choose **EXACTLY THREE** questions.

3. This question concerns regression.

   (a) Briefly explain the difference between simple linear regression, multiple linear regression and logistic regression. [3]

   (b) The data set below represents 4 people, and contains their annual income and their house price. The final two attributes are in thousands of pounds.

Person	Annual Income (£,000s) (x)	House Price (£,000s) (y)
1	10	80
2	65	480
3	43	310
4	15	100

   We wish to learn a simple linear regression model for predicting the house price based on the annual income, of the form of:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

   Calculate the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, and write out the full equation in the form above. [14]

   (c) What is the Total Sum of Squares (TSS) of the example in part (b)? [3]

Continued

4. This question concerns classification.

   (a) What are overfitting and underfitting? [4]

   (b) Specificity and precision are defined as follows:

$$\text{Specificity} = \frac{TN}{FP + TN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

   For the following confusion matrix, calculate the specificity and precision. You should assume that a classification of "Yes" is equivalent to a positive classification. Briefly explain what specificity and precision tell us.

		Predicted Class	
		Yes	No
True Class	Yes	454	123
	No	96	512

[4]

Continued

(c) Let there be m classes $(C_1, C_2, ..., C_m)$. Using the Naïve Bayes classifier, we can calculate the probability that a data tuple $\mathbf{X}$ ($\mathbf{X} = (x_1, x_2, \ldots, x_n)$) belongs to class C_i using the following:

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

where

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Consider the following data:

age_bracket	favourite_cuisine	salary_band	student_status	attends_restaurant
young	indian_food	high	yes	yes
middle_aged	chinese_food	low	yes	no
middle_aged	chinese_food	low	no	yes
senior	indian_food	high	no	no
young	british_food	low	no	no
middle_aged	british_food	high	yes	yes
senior	indian_food	low	yes	no
young	chinese_food	high	no	no
senior	chinese_food	high	yes	yes
young	chinese_food	low	no	yes
senior	indian_food	high	no	yes

Let's say we want to classify the following individual (i.e. we want to predict if they will or will not attend the restaurant, that is, if attends_restaurant=yes or attends_restaurant=no):

$\mathbf{X} = (age\_bracket = middle\_aged, favourite\_cuisine = indian\_food,$
$salary\_band = high, student\_status = no)$

Showing your working, classify $\mathbf{X}$ using the Naïve Bayes classifier. [12]

---

5. This question concerns clustering.

   (a) Is clustering an example of a supervised or unsupervised method? Why? [3]

   (b) Briefly explain the difference between Hierarchical Agglomerative Clustering, and Hierarchical Divisive Clustering. [3]

   (c) The table below shows two-dimensional data associated with seven objects, and the cluster ID that each object is assigned to.

Object ID	$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$	$\mathbf{o_4}$	$\mathbf{o_5}$	$\mathbf{o_6}$	$\mathbf{o_7}$
x	1	2	0	3	5	9	7
y	6	5	4	1	3	4	7
Cluster ID	C_1	C_1	C_1	C_2	C_2	C_3	C_3

   For object $\mathbf{o_1}$, calculate $a(\mathbf{o_1})$ (the average distance between $\mathbf{o_1}$ and all other objects in the same cluster as $\mathbf{o_1}$), and $b(\mathbf{o_1})$ (the minimum average distance from $\mathbf{o_1}$ to all clusters to which $\mathbf{o_1}$ does not belong). You should use Euclidean distance as your distance metric.

   Then, calculate the silhouette coefficient of object $\mathbf{o_1}$, denoted $s(\mathbf{o_1})$, which is defined below.

$$s(\mathbf{o_1}) = \frac{b(\mathbf{o_1}) - a(\mathbf{o_1})}{max\{a(\mathbf{o_1}), b(\mathbf{o_1})\}}$$

   [10]

   (d) What information does the silhouette coefficient of an object provide? [2]

   (e) We have so far only calculated the silhouette coefficient for a single object. Explain how the silhouette coefficient can be used to measure the overall clustering quality. [2]

Continued

6. This question concerns recommender systems.

   (a) Explain the difference between user-based and item-based collaborative filtering. [4]

   (b) Briefly explain the following rating types. For each rating type you should provide a justified example of where it would be appropriate in a real-world situation.

      i. Ordinal ratings [2]

      ii. Binary ratings [2]

      iii. Unary ratings [2]

   (c) The table below shows movie review data from five users. Ratings are on a scale of 0 to 10.

Item-ID (⇒) User-ID (⇓)	1	2	3	4	5	Mean Rating	Pearson(i,4) (user-user)
1	9	7	4	8	9	7.4	0.152
2	5	1	7	8	9	6	-0.634
3	6	2	1	4	5	3.6	0.184
4	9	8	?	4	?	7	1
5	8	5	1	2	1	3.4	0.791

   As you can see, user 4 has not provided a rating for movies 3 or 5. Showing your working, calculate these ratings, using the Pearson Product Moment Correlation Coefficient as the similarity function. You should not mean-centre the ratings for part (c). The Pearson Product Moment Correlation values between each user (i) and user 4 have been provided in the table above. You should use the prediction function below:

$$\hat{r}_{uj} = \frac{\sum_{v \in P_u(j)} \text{Sim}(u, v) \cdot r_{vj}}{\sum_{v \in P_u(j)} |\text{Sim}(u, v)|}$$

   where $\hat{r}_{uj}$ is the predicted rating of user u for item j, $P_u(j)$ is the set of k-closest users to the target user u who have observed ratings for item j, Sim(u,v) is the similarity measurement (in our case, Pearson Product Moment Correlation) between users u and v, and $r_{v,j}$ is an observed rating for user v for item j. For this question, choose $k = 2$ for creating the set of k-closest users, $P_u(j)$.

   If we must recommend one movie out of these two (movies 3 and 5) to the user, which movie of the two should be recommended? Do you think this movie will be liked by the user? [5]

(d) The observed ratings have now been mean-centred.

Item-ID (⇒) User-ID (⇓)	1	2	3	4	5	Mean rating (of raw ratings)	Pearson(i,4) (user-user)
1	1.6	-0.4	-3.4	0.6	1.6	7.4	0.152
2	-1	-5	1	2	3	6	-0.634
3	2.4	-1.6	-2.6	0.4	1.4	3.6	0.184
4	2	1	?	-3	?	7	1
5	4.6	1.6	-2.4	-1.4	-2.4	3.4	0.791

Compute the new predicted ratings of user 4 for movies 3 and 5, using the prediction function below:

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} \text{Sim}(u,v) \cdot s_{vj}}{\sum_{v \in P_u(j)} |\text{Sim}(u,v)|}$$

The definitions of individual terms remain the same as in the previous part, with the addition of s_{vj} being the mean-centred rating for user v and item j, and μ_u being the mean rating for user u (of their raw ratings). Discuss the differences that you observe between the ratings calculated in part (c), and in part (d). [5]

Continued

7. This question concerns social network analysis. A brand new Social Networking Site, *DCSBook*, is launched. Users can form symmetric (undirected) relationships with each other (i.e. if A is a friend of B, then B is a friend of A). The data below shows a nodelist and adjacency matrix of four users on the site.

A	Alice	Female	Scottish
B	Bernard	Male	English
C	Charlotte	Female	Welsh
D	Dmitri	Male	Irish

	A	B	C	D
A	0	1	1	0
B	1	0	1	1
C	1	1	0	1
D	0	1	1	0

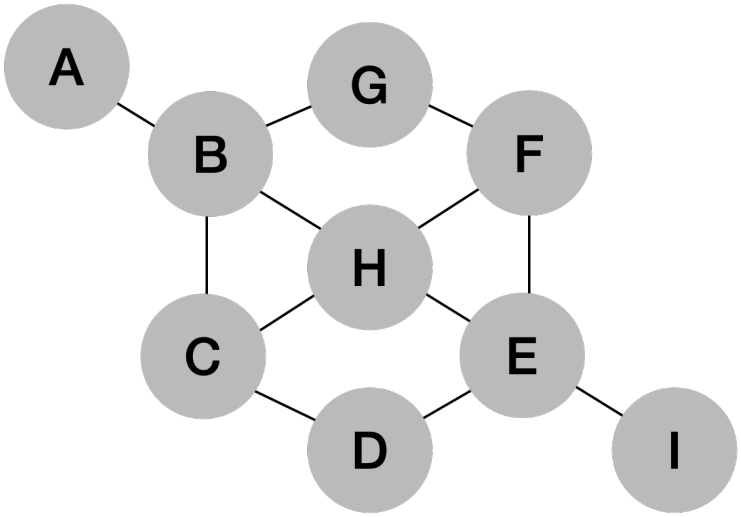
(a) Convert the above nodelist and adjacency matrix into a graph. [3]

(b) Compute the local clustering coefficients of the nodes corresponding to Alice and Bernard.
You may find the following useful: $\binom{n}{r} = \frac{n!}{(r!(n-r)!)}$, where $x! = x \times (x-1) \times (x-2) \times ... \times 2 \times 1$ for any positive integer x, and $0! = 1$. [8]

(c) Compute the Watts-Strogatz network average clustering coefficient for the graph. You can assume that the local clustering coefficient of node Charlotte is equivalent to that of node Bernard, and the local clustering coefficient of node Dmitri is equivalent to that of node Alice. [2]

Continued

(d) For part (d), please refer to the following graph:



i. What is the eccentricty of node B? [2]
ii. What is the degree of node E? [1]
iii. Which node is a central node, and why? [2]
iv. What is the diameter of the graph? [2]

End