



Final Exam (Take-Home) Summer - 2020

Subject: Data Science	Max. Marks: 40
Instructor: Dr. Syed Hasan Adil	
Program: BS (CS)/ MS(CS)	

Faculty of Engineering, Sciences & Technology

Please follow the instructions carefully:

1. Write your answers in a Word file and upload the file before the due date on VLE.
2. Write your name and registration ID on the first page of your Word file.
3. **Answer scripts can be uploaded on LMS any time before its deadline. Therefore, do not wait for the last hour to avoid any unforeseen problems.**
4. **Submission of answer copy(ies) will be considered acceptable through LMS only. Therefore, do not submit your document through email or any other medium.**
5. Use 12 pt. font size and Times New Roman font style along with 1-inch page margins.
6. Follow the requirements of the word limit and the marking criteria while writing your answers.
7. Provide relevant, original and conceptual answers, as this exam aims to test your ability to examine, explain, modify or develop concepts discussed in class.
8. Do not copy answers from the internet or other sources. The plagiarism of your answers may be checked through Turnitin.
9. Recheck your answers before the submission on LMS to correct any content or language related errors.
10. Double check your word file before uploading it on LMS to ensure that you have uploaded the correct file with your answers.

Question 1**[5 marks]**

The dataset of bridges in Pittsburgh is given in the bridges.csv file. Use this dataset to answer the following questions,

- a) Address all the missing values.
- b) Look for outliers and smooth noisy data.
- c) Prepare the dataset to establish a relation among:
 - 1) Length of the bridge and its purpose.
 - 2) Number of lanes and its materials.
 - 3) Span of the bridge and number of lanes.

Question 2**[5 marks]**

Write R code using R markdown to perform following SQL queries on the following database schema given below. The database consists of the following four tables,

Employee (person_name, street,city)

Works (person_name, companyname, salary)

Company (company_name, city)

Manages (person_name, manager_name)

- a) Find the name and city of all employees who works for ABC Company?
- b) Find the name and city of all employees who works for ABC Company and earn between \$12,000 and \$20,000?
- c) Find the employee details who work and live in the different city for which they work?
- d) Find the name of the employees who live in the same cities and same street as their manager?
- e) Find all employees in the database who do not work for ABC Company?
- f) Find all employees in the database who earn less than each employee of ABC Company?
- g) Find all employees who earn less than average salary of all employees of their company?
- h) Find the company that has less employees?
- i) Find the company that has the highest payroll?
- j) Find those companies whose employees earn a lowest salary on average than the average salary at ABC Company?

Question 3**[6 marks]**

The AIS Dynamic dataset is given in the nari_dynamic_sar.csv file. Use this dataset to answer the following questions,

- a) How many unique vessels are available in the dataset?

- b) List the number of records available for each vessel in the dataset.
- c) Find out the spatial (latitude and longitude) and temporal coverage of each vessel in the dataset.
- d) Let us use R to understand the relation between speed over the ground and spatial coverage of the vessels that have multiple records.

Question 4

[08 marks]

The Airline Cost dataset is given in the `airlinecost.csv` file. This dataset has the following attributes, among others:

- Airline name
 - Length of flight in miles
 - Speed of plane in miles per hour
 - Daily flight time per plane in hours
 - Customers served in 1000s
 - Total operating cost in cents per revenue ton-mile
 - Total assets in \$100,000s
 - Investments and special funds in \$100,000s
- a) Develop a linear regression model using R language to predict the number of customers each airline serves from its length of flight and daily flight time per plane. Next, build another regression model to predict the total assets of an airline from the customers served by the airline. Do you have any insight about the data from the last two regression models?
- b) Use the gradient descent algorithm in R to find the optimal intercept and gradient to predict the number of customers each airline serves from its length of flight and daily flight time per plane.

Question 5

[06 marks]

The horseshoe crab dataset is given in the `horseshoecrab.csv` file. This dataset has 173 observations of female crabs, including the following characteristics:

- Satellites: number of male partners in addition to the female's primary partner.
- Yes: a binary factor indicating if the female has satellites.
- Width: width of the female crab in centimeters.
- Weight: weight of the female in grams.
- Color: a categorical value having range of 1 to 4, where 1 = light color, and 4 = dark.
- Spine: a categorical variable, valued between 1 and 3, indicating the goodness of spine of the female.

- a) Use Softmax regression to predict the condition of the spine of female crabs based on the remaining features in the dataset and report the accuracy of your predictions.

Question 6

[05 marks]

Consider the following set of training examples:

Instance	Attribute 1 (A1)	Attribute 2 (A2)	Class
1	Yes	Yes	+
2	Yes	Yes	+
3	Yes	No	-
4	No	No	+
5	No	Yes	-
6	No	Yes	-
7	No	No	+
8	Yes	Yes	+

- a) What is the entropy of this collection of training examples with respect to the target function class?
- b) What is the information gain of a1 and a2 relative to these training examples? Which of the feature should be selected as the root node of the decision tree?

Question 7

[05 marks]

Consider the data below:

Case	X1	X2
1	1	1
2	1	2
3	3	6
4	5	7
5	8	5

Cluster this dataset hierarchically and also provide answers to the following:

- a) Compute the matrix of squared Euclidean distances.
- b) Which two cases are closest together?
- c) What are the two clusters?