PAPER CODE NO.
COMP336

EXAMINER: Dominik Wojtczak Tel. No. dkw@liverpool.ac.uk
DEPARTMENT: Computer Science

UNIVERSITY OF LIVERPOOL

FIRST SEMESTER EXAMINATIONS 2022/23

Big Data Analytics

TIME ALLOWED : TWO Hours

INSTRUCTIONS TO CANDIDATES

Answer **ALL** questions.

The exam consists of 4 questions.

The numbers in the right hand margin represent an approximate guide to the marks available for that part of a question.

Total marks available are 100.

Question 1 is worth 30 marks.

Question 2 is worth 14 marks.

Question 3 is worth 26 marks.

Question 4 is worth 30 marks.

Calculators are permitted.

1. **(a)** Consider the following dataset: $\mathbf{x} = \{1, 5, 5, 5, 5, 5, 5, 5, 30\}$.

i. Calculate the mean and variance of this dataset. **[7 marks]**

ii. Do you think that the mean and variance are representative of the location and spread respectively? Justify your answer. **[3 marks]**

iii. What other summary statistics could you use to describe the location and spread? **[4 marks]**

(b) The residual sum-of-squares for two statistical models are 24 for Model A and 42 for Model B.

i. Is Model B necessarily better than Model A? Give an explanation. **[3 marks]**

ii. How might you check the accuracy of the models? **[3 marks]**

(c) What is more likely when tossing a fair coin 6 times: there will be 4–2 split (i.e., 4 tails and 2 heads, or 4 heads and 2 tails) or there will be 3–3 split (i.e., 3 heads and 3 tails)? Justify your answer and calculate the probability of these events. **[6 marks]**

(d) What is the Markov property of a model and how is it used in the Kalman filter? **[4 marks]**

2. **(a)** What are the four Vs of Big Data and what do they mean? **[4 marks]**

(b) Name three clustering algorithms. **[3 marks]**

(c) What is the role of NameNode in Hadoop cluster? **[3 marks]**

(d) Suppose we have a cluster with 3 data nodes: node DN1, DN2 and DN3. We want to distribute 2 files from a local file system to this HDFS cluster with replication factor 2. File 1 is divided into three blocks: B1, B2 and B3 and the other File 2 is divided into three blocks: B4, B5 and B6. Your task is to distribute all blocks across our data nodes DN1, DN2 and DN3 as evenly as possible.

[4 marks]

3. Consider the following dataset, which consists of a set of records with the format:

(price, amount of product sold, day of the month, month)

The dataset we are going to use is the following: (70, 7, 29, 1), (30, 9, 30, 1), (40, 11, 31, 1), (70, 8, 1, 2), (80, 6, 2, 2), (50, 10, 3, 2).

Write down a Map and Reduce data flow steps/diagram to solve the following problem where you should randomly split the input data into two equal parts. Your solution should be data efficient, i.e., retain as little data as needed after the Map operation.

(a) For each month output the amount of product sold (in total that month).

[6 marks]

Now for the same dataset write down PySpark code that just uses the standard RDDs' actions and transformations to output what each of the questions **(b), (c), (d), (e)** below asks for. (We are only interested in the code, not its output.) You should assume that everything is already setup, and the data is loaded into variable `input` which consists of 4-tuples (i.e., not a DataFrame). In other words, if x is a single row, then $x[0]$ is the month, $x[1]$ is day of the month, etc. In each of the solutions you have to use `reduceByKey` transformation. You can make use of the standard operator `add` which was already imported as `from operator import add`.

(b) For each month output the the total sales value (i.e., the sum over all days that month of the amount of product sold times its price).

[5 marks]

(c) Output the month for which the sales value is the highest.

[5 marks]

(d) For each (month, day) pair output the total amount of product sold that day.

[5 marks]

(e) Output the (month,day) pair for which the amount of product sold is the highest.

[5 marks]

4. Consider the following dataset, which consists of a set of records with the format:

(class, feature 1, feature 2)

The dataset we are going to use is the following: (0, 1, 2), (0, -1, 1), (0, -2, -1), (1, 2, 1), (1, 1, -1), (1, -1, -2).

(a) Write down a design matrix X for this dataset.

[2 marks]

(b) Draw a plot of the features, where the plot's horizontal axis (x_1) corresponds to feature 1 and the vertical axis (x_2) corresponds to feature 2. Represent the class 0 samples with filled-in dots and the class 1 samples with crosses.

[2 marks]

(c) The maximal-margin classifier for this dataset has a separating hyperplane at $x_2 = x_1$. Add this as a thick solid line on the plot.

[1 marks]

(d) Draw the margins for this classifier as a pair of dotted lines (assuming hard margins - i.e. that no point is allowed to cross its respective margin).

[4 marks]

(e) Mark the support vectors on the plot with open squares.

[4 marks]

(f) For the linear SVC described above, identify the values of the unit vector β and the scalar β_0.

[7 marks]

(g) For linear SVCs in general, β has the property that it can be written as a weighted sum of the feature vectors, where a given weight is non-zero if and only if the corresponding feature vector is a support vector.

i. Find a set of weights that satisfy this property. **[6 marks]**

ii. Write your weights as a vector α and verify that $X^T\alpha = \beta$ for the design matrix X that you provided in part (a). **[4 marks]**