PAPER CODE NO.	EXAMINER:	Tel. No.
BigDataExamQuestions	DEPARTMENT: Computer Science	

UNIVERSITY OF LIVERPOOL

EXAMINATIONS

SAMPLE QUESTIONS

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.

UNIVERSITY OF LIVERPOOL

1. **(a)** Consider the following dataset: $\mathbf{x} = \{1, 4, 5, 5, 5, 5, 5, 30\}$.

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. **[5 marks]**

The mean is 60/8=15/2=7.5 and the variance is $((1 - 15/2)^2 + (15/2 - 4)^2 + 5 \cdot (15/2 - 5)^2 + (30 - 15/2)^2)/(8 - 1) = 592/7 = 84.57$. They are not representative because there are two big outliers in the data set; the mean is only slightly off but the variance is very large.
for the correct mean [2pt], for the correct variance [5pt], division by 8 instead of (8-1) in the variance [-2pt], for correct explanations that these are not representative statistics [5pt].

(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? **[4 marks]**

(i) No; it may be over-fit, i.e., may not generalise to other inputs. (ii) Any one of: gather some more data and re-calculate the error, leave some data out and re-calculate the model, or use a re-sampling technique will be valid.
Correct answer to (i) [1 mark] mention of over-fitting in (i) [2 mark] One of the valid answers for (ii) [4 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. **[7 marks]**

It is more likely that 4–2 split will occur. [1pt] The probability of 4-2 is 6 choose 4 + 6 choose 2 = 15+ 15 = 30 divided by 2^6 = 64 which gives 30/64 = 46.875% and the probability of 3-3 is 6 choose 3 = 20 / 64 = 31.25% [6 marks]

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

The Markov property is that the state of a system can be calculated from the previous step alone without knowledge of any previous steps. The Kalman filter uses this to write a single update equation for the state at each step. Knowledge of the Markov property[2]. Justifies how this is used to update based on the previous step alone[2].

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

Volume - Scale of Data, Variety- Different Forms of Data, Velocity-Analysis of Streaming Data, Veracity - Uncertainty of Data

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

Hierachical Clustering, K-means clustering, DBSCAN

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

i. Bookkeeper for HDFS
ii. Keeps track of where blocks constituting copies of each file are stored
iii. Monitors health of the distributed file-system

(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]

One possible solution:
DN1: B1 B2 B3 B4
DN2: B3 B4 B5 B6
DN3: B5 B6 B1 B2
Wrong replication factor: -2pt
Not evenly distributed: -2pt

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

(store ID, amount of product sold, day of the month, month)

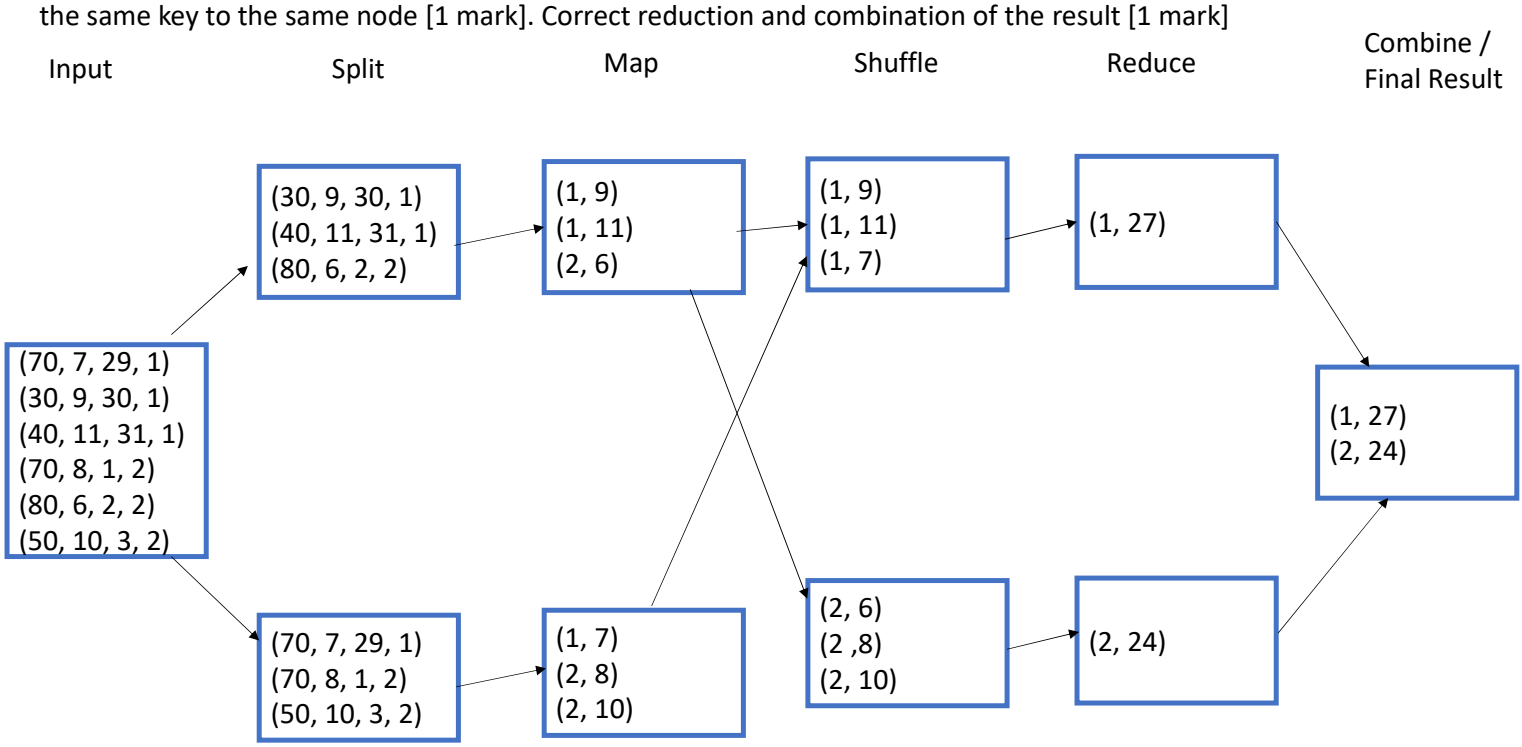
The dataset we are going to use is the following: (70, 7, 29, 1), (30, 9, 30, 1), (40, 11, 31, 1), (70, 8, 1, 2), (80, 6, 2, 2), (50, 10, 3, 2).

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]

3a solution. The exact solution will depend on the random split. However, the diagram should more or less look as follows. All stages present and clearly marked [1 mark]. Split is even and is random [1 mark]. Correct keys and values indentified [1 mark]. Only the relevant part of the data is keep after Map operation. [1 mark] Shuffle moves all pairs with the same key to the same node [1 mark]. Correct reduction and combination of the result [1 mark]



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 store ID, $x[1]$ is amount of product sold by this store, 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 total amount of product sold across all stores.

[5 marks]

One possible solution is

```
input.map(lambda x:  (x[3],x[1]).reduceByKey(add).collect()
```

(c) Output the month for which the total amount of product sold across all stores is the highest.

[5 marks]

One possible solution is:

```
input.map(lambda x:  (x[3],x[1])).reduceByKey(add).max(lambda x:  x[1])[0]
```

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

[5 marks]

One possible solution is

```
input.map(lambda x:  ((x[3],x[2]),x[1])).reduceByKey(add).collect()
```

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

[5 marks]

One possible solution is

```
input.map(lambda x:  ((x[3],x[2]),x[1])).reduceByKey(add).max(lambda x:  x[1])[0]
```

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]

The design matrix is given below. It's also OK to swap order of the two columns.

$$X = \begin{bmatrix} 1 & 2 \\ -1 & 1 \\ -2 & -1 \\ 2 & 1 \\ 1 & -1 \\ -1 & -2 \end{bmatrix}$$

Note that there is a column for each feature and a row for each sample. The classes do not appear in the design matrix.

Full marks for a correct answer. Otherwise 0 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]

The plot should be symmetric about the line ($x_2 = x_1$), with a triangle of filled-in dots to the upper left of $x_2 = x_1$ plot and a triangle of crosses to the lower right.

Full marks for a correct answer. Otherwise 0 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]

There should be a horizontal thick solid line at $x_2 = x_1$.

Full marks for a correct answer. Otherwise 0 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]

There should be two horizontal dotted lines at $x_2 = x_1 \pm 1$.

2 marks for each of these two lines.

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

[4 marks]

There are 4 support vectors in the dataset, which all lie on the dotted margin lines: (1, 2), (2, 1), (-1, -2), (-2, -1). These should be marked with open squares.

One mark for each correctly identified support vector (max 4).

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

[7 marks]

The separating hyperplane is defined as the set of all points x such that $x^T\beta = \beta_0$. Hence:

- β can be either $(\sqrt{2}, -\sqrt{2})^T$ or $(-\sqrt{2}, \sqrt{2})^T$, because β is the unit vector orthogonal to the separating hyperplane.
- $\beta_0 = 0$, because the separating hyperplane runs through the origin.

5 marks for $\beta = (\sqrt{2}, -\sqrt{2})^T$ or $(-\sqrt{2}, \sqrt{2})^T$. 2 marks for $\beta_0 = 0$.

(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]**

Any set of weights for the 4 support vectors which are **all non-zero** and add up to either $(\sqrt{2}, -\sqrt{2})^T$ or $(-\sqrt{2}, \sqrt{2})^T$. Weights for the other 2 feature vectors should be 0. (The support vectors are listed in the answer to part (e).)

The two possible answers are $\alpha = \pm\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)^T$.

6 marks for stating a valid α. 4 marks for successfully working through the verification of $X^T\alpha = \beta$ by hand.