MCAST

Malta College of Arts, Science & Technology

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

Statistics for Computer Science

Assignment Guidelines

Read the following instructions carefully before you start the assignment. If you do not understand any of them, ask your lecturer.

- The assignment coversheet should be the first sheet in your assignment. Moreover, the coversheet should be fully completed with all the necessary details.
- All text\code must be properly referenced. In the absence of proper referencing, the assignment will be regarded as plagiarised.
- Copying is strictly prohibited and will be penalized in line with the College's disciplinary procedures.
- The deadline should be specified by your lecturer. When the deadline is due, you shall hand in 2 deliverables:
    - A printed, typed, document containing the answers for all Tasks, submitted neatly in a flat file or bound.
    - A CD containing the application, all source code, and a digital copy of the documentation. The CD must be properly attached to the flat file.
- You are also required to submit your assignment via Turnitin on the same deadline as for the hard copy. Your lecturer will forward you details in order to submit your assignment via Turnitin.
- The lecturer may hold a post-submission interview. Attendance to such interview is mandatory. Moreover, marks assigned to the criteria will be affected by the interview performance.
- **All work that has been carried out, must be written down and included within the assignment as evidence. No marks will be awarded for work that is not presented.**
- Please refer to the cover sheet for the assignment deadline.

Part 1
Analysis of a grain data set

A group of kernels from three different wheat varieties, Kama, Rosa and Canadian, has been collected (available at: https://archive.ics.uci.edu/ml/datasets/seeds). There are 70 kernels for each wheat variety. For each individual kernel, the following information has been obtained:

1. Area
2. Perimeter
3. Compactness
4. Length of Kernel
5. Width of Kernel
6. Asymmetry Coefficient
7. Length of Kernel Grove
8. Variety.

Attributes 1 to 7 are real-valued while attribute 8 is the wheat variety.

The aim is to be able to identify the wheat variety from the kernel, using the same information as above, when the actual kernel variety is unknown.

Section 1

Investigate the kernel data set and perform dimensionality reduction. (AA3.2, 7 marks)

The attributes have 7 dimensions (excluding the category attribute). Can the number of dimensions be reduced without losing too much information about the kernel?

You are required to use PCA to perform dimensionality reduction.

Task:

a) Perform PCA on the first 7 attributes, write down the results. (2 marks)

b) Write down the eigenvalues for each of the attributes, sorted in descending order. (1 mark)

c) Create a Scree plot using the eigenvalues found in (b). (1 mark)

d) Chose the number of dimensions to keep, given a sound explanation as to how you arrived at this decision.
Which of the dimensions did you choose? Why?
Save an updated file with the new data. Paste the first 10 rows of the data. (3 marks)

Grading guideline information				
AA3.2	Inadequate Work 0 marks		Superior Work 7 marks	Score Achievement
Complete the above task.	All answers incorrect.		All answers correct, sound and well explained.	

Section 2

Find and analyse clusters in the data set. (KU1.2, 5 marks) (KU3.1, 5 marks) (AA4.2, 7 marks) (SE4.3, 10 marks)

Task 1:

Research and compare different clustering algorithms that can be applied to your data set.

You must compare *at least* 1 randomised and 1 non-randomised clustering algorithm. You must compare *at least* 3 clustering algorithms in total.

For each of the chosen algorithms, give information like expected results, advantages and disadvantages of the algorithm with respect to this data set. (5 marks)

Grading guideline information				
KU1.2	Inadequate Work 0 marks	Inferior Work 3 marks	Superior Work 5 marks	Score Achievement
Complete the above task.	Comparison does not make sense. Only 1 clustering algorithm mentioned.	3 clustering algorithms compared, but the written work is superficial.	All answers correct, sound and well explained.	

Task 2:

Research, the clustering algorithm k-means and 1 or more variation of the k-means algorithm (e.g. k-medoids, k-medians, etc…). Based on your research, choose, a specific variation. You may choose the original k-means itself.

Write and explain the pseudo-code for the chosen variation. (2 marks)

Identify specific issues and problems that you will face when applying the chosen variation. (2 marks)

Discuss how you will choose the value of k (the number of clusters), justifying your choice. (1 marks)

Discuss and describe how you will initialise the algorithm. Give justification of your initialisation method. (1 marks)

Discuss the metric that you will be using for the analysis. Give justification of the metric used. (1 marks)

Grading guideline information				
AA4.2	Inadequate Work 0 marks	Inferior Work 3 marks	Superior Work 7 marks	Score Achievement
Complete the above task.	Missing work or incorrect reasoning.	All answers correct, but the differences and issues in the chosen variations is not sufficiently explained or justification is poor.	All answers correct, sound and well justified and explained.	

Task 3:

Implement the chosen k-means variation.

If you are using an existing library or tool that supplies the actual code, then for the purposes of the assignment, the implementation is the code/script/formula that calls this method.

Place screenshots of the k-means implementation. There is no need to take screenshots of the supporting code, only the k-means and it initialisation.

Apply your implementation to the data set, and describe your resulting clusters, giving screenshots of the cluster visualisation.

Grading guideline information				
KU3.1	Inadequate Work 0 marks	Inferior Work 3 marks	Superior Work 5 marks	Score Achievement
Complete the above task.	Missing work or implementation not working.	Implementation and screenshots of implementation supplied but the clusters are not visualised.	Working implementation and screenshots visualising the data and explaining the work carried out provided.	

Task 4:

Visually inspect your cluster visualisation. Comment whether the resulting clusters appear to be good, explaining why. (1 mark)

What is the accuracy of the cluster for identifying the correct variety for the given sample? (1 mark).

The accuracy identified above might not reflect the accuracy for new kernels that were not used in the original sample to build the clusters.

Split the original data set into a training set and a test set and perform the same test again.

Find the sensitivity, specificity and accuracy of the clusters. (3 marks)

Apart from the statistics mentioned above, research another method to evaluate the resulting clusters. Write down information about the selected method. (1 mark)

Use the selected method on your clusters. (1 mark)

Perform the following: (3 marks)

- Try different parameters, metrics and initialisations, and compare the resulting clusters.
- For the optimal setup found, repeat the experiment using different training sets and data sets. Discuss how the clusters vary, both in location and shape as well as the other methods used in evaluating clusters (sensitivity, specificity, accuracy and the method chosen).

Grading guideline information				
SE4.3	Inadequate Work 0 marks	Inferior Work 3 marks	Superior Work 10 marks	Score Achievement
Complete the above task.	Missing work or incorrect reasoning.	Poor work overall. Analysis using different parameters not carried out.	Detailed analysis of cluster quality and attempts using different parameters carried out.	

Part 2
The Monte Hall Problem

A local television studio is planning on launching a new variation of the popular TV show, “Let’s Make a Deal”.

In the original show, the player is presented with 3 doors, 1 hiding a car and the other 2 doors are hiding a booby prize (the goats). The player proceeds by selecting a door, without opening that door. The presenter would then open one of the other doors that does not have the car. The player is offered the opportunity of staying with their current choice or switching to the remaining door.

Figure 1, shows a probability tree representing the game; the car is hidden behind Door 1.

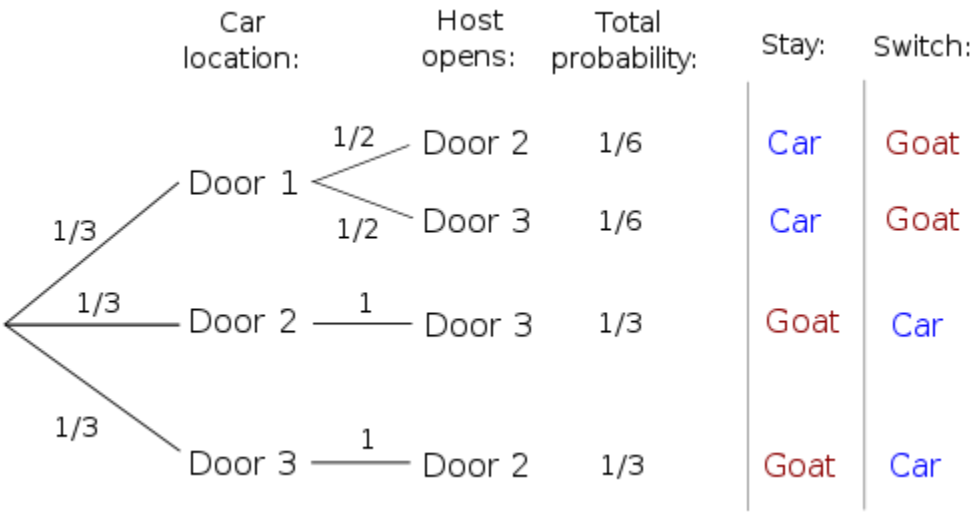

Figure 1 - reproduced from https://commons.wikimedia.org/wiki/File:Monty_tree_door1.svg

The analysis of the game show, with the standard rule set explained above, was very controversial due to the counter-intuitive nature of the problem.

Section 1

Implement a Simple Monte Carlo algorithm to solve the Monte Hall problem. (AA1.4, 7 marks)

A collaborator of a collaborator of a collaborator, Paul Erdős, does not believe that switching the door is an optimal strategy. You are required to analyse the above problem and present a Simple Monte Carlo implementation to convince him otherwise.

Task 1:

Analyse the probability tree given above. Writing down your working, show that the "Stay" strategy has 1/3 probability of winning the car, while the "Switch" strategy has 2/3 probability of winning the car. (1 mark)

Task 2:

Implement a Simple Monte Carlo algorithm to verify the probabilities identified above. Add screenshots of your implementation (only the Simple Monte Carlo algorithm itself is required) and of your estimated probability, using a sample size of 10,000. (4 marks)

Task 3:

Using the techniques presented in the Monte Carlo Algorithms section of the assignment, find the variance of the random variable representing the Monte Hall problem (with a 1 representing a car being found, 0 otherwise).

Use this variance to find the sample size for a RMSE error of 0.1.

What is the estimated probability of finding a car, using this sample size and both strategies? (2 marks)

Grading guideline information				
AA1.4	Inadequate Work 0 marks		Superior Work 7 marks	Score Achievement
Answer the questions above.	All answers incorrect.		Correct implementation and all answers correct and well explained.	

Section 2

Design and implement a Simple Monte Carlo algorithm to solve a variation on the Monte Hall problem and evaluate the results. (SE3.3, 10 marks) (SE3.4, 10 marks)

Variation 1

In the first variation of the original game, the new game will have n doors instead of just 3. One door will hide a car, but the others will hide goats.

The player starts by selecting a door.

Following that, the player will select, amongst the other n-1 doors, n-2 doors to open at random (with no knowledge of whether they have a car or a goat behind them). If the car is revealed, the player loses the game and wins the booby prize.

Otherwise, the player ends up with the selected door and one other door (one of which is hiding the car).

The player is given to option to “Stay” or “Switch”.

Variation 2

The game will be played like variation 1.

Instead of cars and goats, the boxes will contain cash prizes and the game will be played until the last box is opened.

The prizes on the boxes are marked with the following prizes:
0c, €1, €10, €20, €50, €100, €500, €1000, €2500, €10000, €25000, €100000.

Task 1:

Consider variation 1.

a) Implement a Simple Monte Carlo algorithm to analyse variation 1. (5 marks)

b) Plot the probability of winning the car, if the player uses the optimal strategy, for n = 3 to 8, where n is the number of boxes. The probabilities error must have a RMSE of 0.1. (3 marks)

c) Find the smallest value of n, such that the probability of winning the car, using an optimal strategy, is smaller than 0.2. (1 mark)

Grading guideline information				
SE3.3	Inadequate Work 0 marks		Superior Work 5 marks	Score Achievement
Answer task 1, part a from above.	All answers incorrect.		Correct implementation and implementation demonstrated to be working correctly.	

Grading guideline information				
SE3.4	Inadequate Work 0 marks		Superior Work 4 marks	Score Achievement
Answer task 1, parts b and c from above.	All answers incorrect.		Answers correct and well explained.	

Task 2:

Consider variation 2.

a) Implement a Simple Monte Carlo algorithm to analyse variation 2. (5 marks)

b) Using your implementation, perform a simulation of the problem and identify whether an optimal strategy exists and if it does, what is the optimal strategy? Show how you arrived at your solution. (3 marks)

c) Using the optimal strategy found, if any, model the system as a random variable. Find the expected winnings of the game to a RMSE of €1. (3 marks)

Grading guideline information				
SE3.3	Inadequate Work 0 marks		Superior Work 5 marks	Score Achievement
Answer task 2, part a from above.	All answers incorrect.		Correct implementation and implementation demonstrated to be working correctly.	

Grading guideline information				
SE3.4	Inadequate Work 0 marks		Superior Work 6 marks	Score Achievement
Answer task 2, parts b and c from above.	All answers incorrect.		Answers correct and well explained.	

Part 3
Online quizzes

KU1.1, KU1.3 and KU2.1 are assessed through quizzes on Moodle.

Grading guideline information				
KU1.1	Inadequate Work 0 marks		Superior Work 5 marks	Score Achievement
Answer questions on the Moodle quiz.				

Grading guideline information				
KU1.3	Inadequate Work 0 marks		Superior Work 5 marks	Score Achievement
Answer questions on the Moodle quiz.				

Grading guideline information				
KU2.1	Inadequate Work 0 marks		Superior Work 5 marks	Score Achievement
Answer questions on the Moodle quiz.				