# The Chinese University of Hong Kong
# Dept of Decision Sciences & Managerial Economics
# DSME6650: Individual Exercise

Date Due: 10 Feb 2019 (23:55)

Submission: The exercise is in individual basis. Every student should submit his/her own assignment. You should follow the University guidelines as mentioned in http://www.cuhk.edu.hk/policy/academichonesty/ to have an electronic submission. Please submit it to (i) VeriGuide and (ii) the blackboard. Please contact your tutor if you have any problem during your submission.

**Question 1**

(a) 15 data points are arranged and settled into three clusters as follows:

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| Centroid at<br>(15.5, 28.5) | Centroid at<br>(43.75, 19.75) | Centroid at<br>(10.4, 10.6) |
| (11, 34) | (42, 25) | (5, 12) |
| (16, 31) | (47, 21) | (10, 18) |
| (18, 24) | (45, 13) | (15, 8) |
| (20, 23) | (41, 20) | (9, 4) |
| (10, 27) | | (13, 11) |
| (18, 32) | | |

(i) Which is the nearest cluster of Cluster 1?

(ii) What are the average Silhouette coefficients of the three clusters?

(iii) What is the average Silhouette coefficient of the entire dataset?

(iv) Which cluster is more cohesive than others?

(v) Justify whether the choice k=3 for the clustering is appropriate or not.

(b) Apply the fuzzy c-means technique on the dataset with three initial centroids, (15, 28), (40, 20) and (10, 8) for three iterations. Compare your result in (a).

**Question 2**

Consider the training instances shown in Table Q2 for a dichotomous decision problem.

(a) What is the entropy of this collection of customer instances with respect to the target attribute with value "Stay"?

(b) What are the information gains of the two attributes x1 and x2 relative to the customer instances?

(c) For the attribute x3, identify the best threshold by computing the information gain for every possible split.

(d) What is the best split, among x1, x2, and x3, according to the information gain?

(e) What is the best split, between x1 and x2, according to the Gini index?

(f) Devise two different but complete decision trees for the instances shown in the table using information gain.

| CID | x1 | x2 | x3 | Target attribute |
|---|---|---|---|---|
| 01 | M | HIGH | 1.0 | Stay |
| 02 | M | HIGH | 6.0 | Stay |
| 03 | M | LOW | 5.0 | Churn |
| 04 | F | LOW | 4.0 | Stay |

| | | | | |
|---|---|---|---|---|
| 05 | F | HIGH | 7.0 | Churn |
| 06 | F | HIGH | 3.0 | Churn |
| 07 | F | LOW | 8.0 | Churn |
| 08 | M | LOW | 7.0 | Stay |
| 09 | F | HIGH | 5.0 | Churn |

Table Q2: Instances show whether the customers stay and churn in a brand

**Question 3**

Risk assessment is important for financial institution, especially in loan applications. Some have already implemented their own credit-scoring mechanisms to evaluate their clients' risk and make decisions based on this indicator. In fact, the data gathered by financial institutions is valuable source of information to create information assets, from which credit-scoring mechanisms can be developed. The purpose of this question is to, from information assets, create a decision mechanism that is able to evaluate a client's risk. The attached data files contain 700 training instances and 300 test instances. The training instances are used to train decision models while the test instances are used for the evaluation. The data set itself is a combination of personal, social and financial information about past bank clients. The complete list of attributes is shown in Table Q3.

| No. | Attribute | Description |
|---|---|---|
| 1 | Status | Status of checking account |
| 2 | Duration | Loan duration in months |
| 3 | Credit history | Client's credit history |
| 4 | Purpose | Purpose of the loan |
| 5 | Credit amount | Loan amount |
| 6 | Savings | Saving accounts or bonds |
| 7 | Employment duration | Duration of the present employment |
| 8 | Instalment rate | Instalment rate of disposable income |
| 9 | Personal status | Personal marital status and sex |
| 10 | Debtors | Other debtors/guarantors |
| 11 | Residence | Present residence since |
| 12 | Property | Property owned |
| 13 | Age | Age in years |
| 14 | Instalment plans | Other instalment plans |
| 15 | Housing | House status |
| 16 | Existing credits | Number of existing credits at the bank |
| 17 | Job | Job category |
| 18 | Liable people | Number of people liable to |
| 19 | Telephone | Existence of telephone number |
| 20 | Foreign worker | Native or foreign worker |
| 21 | Classification | Credit classification |

Table Q3: Description of attributes

(a) Explain what are the pre-processing procedure(s), if any, that should be taken before the data mining.

(b) Without any data preprocessing, compare the training and test results on the following major classification algorithms using Weka. You have to produce a short report, with not more than 800 words, to explain your findings.
   (i) Logistic regression
   (ii) Ridge logistic regression for the optimal ridge

(iii) k-NN, for the optimal value of k.
(iv) Decision Tree, using J48
(v) Random Forest

**Question 4**

Study the business case “Customer Segmentation” carefully.

(a) Produce a short summary, not more than 300 words, explain how the customer segmentation can be achieved.

(b) Using not more than 500 words, list out all the major steps in the KDD approach. In each step, describe clearly, in form of a table,

(i) what is/are the objective(s) of the step.
(ii) what is/are the major technique(s) or data mining algorithm(s) involved.
(iii) what is/are the precaution(s) that should be taken when the techniques or algorithms are applied.

*** End ***