Imperial College Business School

MSc Business Analytics Examinations 2017/2018

For internal Students of Imperial College of Science Technology and Medicine.
This paper also forms part of the examination for the Associateship.

MACHINE LEARNING (BS1806)

Day date month year; xx:xx – xx:xx

CLOSED BOOK

Instructions

Answer all questions. **Please submit both your hand-written response as well as the code that you have produced!**

Please indicate if the use of college-approved calculators is allowed: YES.

Please indicate the weighting(s) of the questions: SEE BELOW.

Question 1 (*k*-Nearest Neighbours – By Hand): 34 points

Consider the following dataset, which records whether loans have been paid back or not ("default?") depending on characteristics of the borrower (age – *in whole numbers*, income – *in intervals of £10k*, marital status – *yes/no* and education level – *high school, UG studies or PG studies*; missing values are shown as "—"):

DEFAULT?	AGE	INCOME	MARRIED?	EDUCATION
YES	25	£20k-£30k	No	UG studies
NO	37	£80k-£90k	No	PG studies
NO	42	£60k-£70k	Yes	PG studies
YES	21	£10k-£20k	No	High school
??	30	£40k-£50k	—	PG studies
??	50	—	No	UG studies

Use a 2-Nearest Neighbours classifier to predict whether the last two applicants will pay back their loans or not. Do the calculations "by hand" and explain all of your calculations! To this end:

(a) Suggest a suitable normalisation for your dataset, if necessary. You can expect that all applicants are 20-65 years old, earn between £10k and £100k and have attended high school, UG (undergraduate) studies and/or PG (postgraduate) studies. **[12 points]**

(b) Determine the two nearest neighbours for each of the two last records. Explain how you deal with the missing values, and justify your approach. Predict for each of the two last records whether the applicants will default or not. **[12 points]**

Critically assess your classifier. What would you change (if anything) and why? **[10 points]**

Question 2 (Classification and Regression Trees – In Python): 33 points

The "heart.csv" dataset contains 303 patient records: In each record, the first thirteen fields describe the input parameters in the following order:

1. **Age:** age in years.
2. **Sex:** 1 (male) or 0 (female).
3. **Type of chest pain:** 1 (typical angina), 2 (atypical angina), 3 (non-anginal pain) or 4 (asymptomatic).
4. **Resting blood pressure:** in mm Hg on admission to the hospital.
5. **Cholesterol:** serum cholestoral in mg/dl.
6. **Blood sugar:** 1 (above 120 mg/dl) or 0 (below 120 mg/dl).
7. **Resting electrocardiographic results:** 0 (normal), 1 (ST-T wave abnormality), 2 (probable or definite left ventricular hypertrophy)
8. **Maximum heart rate achieved:** numeric.
9. **Exercise induced angina:** 1 (yes) or 0 (no).
10. **ST depression induced by exercise:** numeric.
11. **Slope of the peak exercise ST segment:** 1 (upsloping), 2 (flat) or 3 (downsloping).
12. **Number of major vessels colored by flourosopy:** 0, 1, 2 or 3.
13. **Thalassemia:** 3 (normal), 6 (fixed defect) or 7 (reversable defect).

The fourteenth (final) field in each record is the output parameter, which describes whether a heart disease is diagnosed (value 1) or not (value 0).

(a) Load the dataset into Python. Keeping in mind that you will construct a classification tree, would you suggest to apply any normalisation to the dataset? If so, conduct the normalisation (in Python). Either way, justify your answer! **[10 points]**

(b) Shuffle the dataset and split it into a training (70%) and a validation (30%) set (in Python). **[5 points]**

(c) Train a classification tree on the training set, and use the validation set to predict the generalisation error of your classifier (in Python). **[8 points]**

(d) How could you improve the performance of your classifier? Discuss the potential advantages and shortcomings in the light of the application area! **[10 points]**

Question 3 (Hierarchical Clustering – By Hand): 33 points

Consider the following comparison table of different oil types:

Oils/Fats	*Saturated Fatty Acids*	*Mono-Unsaturated Fatty Acids*	*Omega-6 : Omega-3 Ratio*
Mustard Oil	4	70	1.2 : 1
Desi Ghee	65	32	3 : 1
Soya Bean Oil	15	27	10.6 : 1
Olive Oil	13	76	20 : 1

Conduct a Hierarchical Clustering by hand. Explain all of your calculations! To this end:

(a) Normalise the table column by column using a min-max normalisation. Round after the second digit (*e.g.*, 0.27 or 0.98). **[7 points]**

(b) Conduct a hierarchical clustering on the dataset (using the Centroid distance and the 2-norm) until only two clusters are left. Again, round all calculations after the second digit. **[20 points]**

(c) Draw (by hand) a dendrogram (it does not have to be fully to scale). Explain your steps in creating the dendrogram. Provide a brief interpretation of your result. **[6 points]**