Imperial College Business School

MSc Business Analytics Examinations 2018/2019

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.

©Imperial College London 2018/2019

Solution to Question 1 (a):

Since all input parameters are ordinal, we can translate them into numbers that range between 0 and 1 as follows:

DEFAULT?	AGE	INCOME	MARRIED?	EDUCATION
YES	0.11	0.11	0.00	0.50
NO	0.38	0.78	0.00	1.00
NO	0.49	0.56	1.00	1.00
YES	0.02	0.00	0.00	0.00
??	0.22	0.33	—	1.00
??	0.67	—	0.00	0.50

Here we applied a minmax normalisation, which is justified if the data is more or less uniformly distributed. If you would argue that this is not the case, then you should use a Z-score transformuation.

Solution to Question 1 (b):

There are different ways to deal with the missing values. One way is to just ignore them in the calculations. This can be justified along the lines of our in-class discussion in the Introduction module. We then obtain the following distances for the penultimate record (using the 2-norm):

0.56 0.48 0.35 1.07; nearest neighbours: samples 2 and 3.

Thus, the prediction for the penultimate sample is “no”. For the ultimate sample:

0.56 0.58 1.13 0.82; nearest neighbours: samples 1 and 2.

In this case, we have a tie between “yes” and “no”.

Solution to Question 1 (c):

Several things can be discussed here: transparency (perhaps a classification tree would be better in this regard); other ways to treat missing values; other distance measures (rather than transforming all categorical values to numeric ones); further input parameters; …

©Imperial College London 2018/2019

Solution to Question 2 (a):

Normalisation is not needed for classification trees; you can argue along the lines of our discussion in Module 4 here.

```
import numpy as np
import pandas as pd
from sklearn.utils import shuffle
from sklearn import tree

df=pd.read_csv("heart_proc.csv",names=["age","sex","paintype","bloodpressure","Cholestoral",
"sugar","REC","MHRA","EIA","STDIE","SOPE","NMVCF","Thal","DIAG"])
```

Solution to Question 2 (b): *(NB: This is Python 2.7 code for illustration purposes only. In the exam, you will be asked to produce Python 3 code as learnt in Heikki's class.)*

```
data=np.array(df)
X=data[:,:-1]
y=data[:,-1]
X,y=shuffle(X,y)

#split to train/val/test 50/30/20
datapoints=X.shape[0]
trainsize=int(0.7*datapoints)

X_train=X[:trainsize,:]
X_val=X[trainsize:,:]
y_train=y[:trainsize]
y_val=y[trainsize:]
```

Solution to Question 2 (c): *(NB: This is Python 2.7 code for illustration purposes only. In the exam, you will be asked to produce Python 3 code as learnt in Heikki's class.)*

```
clf=tree.DecisionTreeClassifier()
clf=svm.LinearSVC()
clf=clf.fit(X_train,y_train)
print "in sample error", clf.score(X_train,y_train),"validation error",clf.score(X_val,y_val)
from sklearn.metrics import confusion_matrix
y_pred=clf.predict(X_val)
print "confusion matrix : "
print confusion_matrix(y_val, y_pred,labels=[0,1])
```

Solution to Question 2 (d):

One can discuss the benefits and shortcomings of tree ensembles (bagging, boosting, random forests) here: better classification performance but less transparency. One could also discuss oversampling or weighting to balance type-1 and type-2 errors.

©Imperial College London 2018/2019

Solution to Question 3 (a):

After normalisation, the table looks as follows:

Oils/Fats	*Saturated Fatty Acids*	*Mono-Unsaturated Fatty Acids*	*Omega-6 : Omega-3 Ratio*
Mustard Oil	0.00	0.88	0.00
Desi Ghee	1.00	0.10	0.10
Soya Bean Oil	0.18	0.00	0.50
Olive Oil	0.15	1.00	1.00

NB: I have not provided the formulas here; you should do so in the exam. The formula for the normalisation can be found in the slides for Module 2.

Solution to Question 3 (b):

The initial pairwise distances are as follows:

	Mustard Oil	*Desi Ghee*	*Soya Bean Oil*	*Olive Oil*
Mustard Oil	0.00	1.27	1.03	1.02
Desi Ghee		0.00	0.92	1.53
Soya Bean Oil			0.00	1.12
Olive Oil				0.00

Based on these distances, we cluster *Desi Ghee* and *Soya Bean Oil*. The new cluster centre is (0.59, 0.05, 0.30), and the new distances are:

	Mustard Oil	*{ DG, SBO }*	*Olive Oil*
Mustard Oil	0.00	1.06	1.02
{ DG, SBO }		0.00	1.26
Olive Oil			0.00

We now cluster *Mustard Oil* and *Olive Oil*. The new cluster centre is (0.08, 0.94, 0.50), and the new distances are:

	{ MO, OO }	*{ DG, SBO }*
{ MO, OO}	0.00	1.05
{ DG, SBO }		0.00

NB: I have not provided the formulas here; you should do so in the exam.

Solution to Question 3 (c):

The dendrogram will be of the following shape:

©Imperial College London 2018/2019

Interpretation: DG and SBO are most similar. MO and OO are quite similar, but not as much. One could provide further interpretation based on the linkage used (how could the results differ, e.g. if a minimum or maximum linkage was employed instead).

©Imperial College London 2018/2019