The Hong Kong Polytechnic University
Department of Computing
COMP5511 – Artificial Intelligence Concepts

Assignment 1
Due Date: October 22, 2018

1. SunglassesCrafters, which sells various designer-brand sunglasses, is expanding rapidly and is looking into starting three new retail outlets in such popular locations as Mongkok, Tsimshatsui and the Causeway Bay. To better market their sunglasses, they would like to make sure that the new salespersons they hire be able to recommend the right products to right customers. Knowing that you are taking COMP5511 at PolyU and have been taught about the Machine Learning. Could you help them to identify the preferences of customers?
   a. Use a decision tree model, e.g. ID3 to discover in a sample of their customer database (shown in Table 1) what best recommendation to make to which kind of customers.
   b. You are given a testing data set (shown in Table 2) as follows, how much should you trust the recommendations made according to the rules discovered by the model?
   c. Use the Scikit-Learn package of Python to generate a decision tree model (see Python_decision_tree.html for an example of the use of Scikit-Learn). Compare the results generated by Scikit-Learn with that obtained by you. What are the similarities and differences? Please discuss.

(40 marks)

Table 1: Training Data Set

Tear Production Rate	Sex	Age	Spectacle Prescription	Astigmatism	Recommendation
Reduced	M	Young	Myope	Yes	Lifestyle
Reduced	F	Old	Myope	No	Street
Normal	M	Old	Hypermetrope	Yes	Lifestyle
Normal	M	Young	Hypermetrope	No	Polarized
Reduced	F	Middle	Myope	No	Street
Normal	F	Middle	Hypermetrope	Yes	Lifestyle
Normal	M	Young	Hypermetrope	No	Polarized
Reduced	M	Young	Hypermetrope	Yes	Lifestyle
Normal	M	Old	Myope	No	Street
Normal	M	Middle	Myope	Yes	Polarized
Reduced	F	Middle	Myope	No	Street
Reduced	F	Old	Hypermetrope	Yes	Polarized
Normal	M	Young	Myope	No	Lifestyle
Reduced	F	Old	Hypermetrope	Yes	Polarized
Normal	F	Old	Hypermetrope	Yes	Lifestyle

Table 2: The Testing Data Set

Tear Production Rate	Sex	Age	Spectacle Prescription	Astigmatism	Recommendation
Reduced	F	Young	Hypermetrope	Yes	Lifestyle
Normal	M	Old	Hypermetrope	No	Street
Reduced	F	Old	Myope	Yes	Polarized
Normal	F	Young	Hypermetrope	No	Polarized
Reduced	M	Middle	Myope	No	Lifestyle

2. The customer service department of a local departmental store offers three types of membership to their customers: Gold, Silver, and Bronze. When shopping at any of the 30 store outlets, Gold members receive larger discounts than Silver and Silver than Bronze, etc. In return, Gold members need to pay larger membership dues than Silver and Silver than Bronze, etc. Having been operating for over three years, it is discovered that some Silver members would decide to upgrade to Gold whereas some downgrade to Bronze. To predict if a Silver member may remain, the following data set is obtained.

Customer No.	**Sex**	**Average No. of Transactions**	**Average Monthly Payment**	**Average No. of months in Silver**	**Decision**
1	F	8	301	4	Remain
2	M	18	448	8	Downgrade
3	F	5	305	9	Remain
4	M	3	309	6	Downgrade
5	F	11	522	10	Remain
6	M	3	650	13	Downgrade
7	F	9	490	5	Upgrade
8	M	10	300	7	Upgrade
9	F	7	274	12	Downgrade
10	F	20	575	15	Upgrade
11	M	22	530	9	Downgrade
12	F	14	363	6	Upgrade
13	M	10	409	8	Remain
14	M	15	479	7	Remain
15	M	13	445	11	Remain

a) Assume that $k = 5$, using the k-NN algorithm, what do you expect the decision of a customer, who is a male and has an average no. of transactions of 9, an average monthly payment of 410 and an average no. of months in Silver of 5, to be?

b) Assume again that $k = 5$ and ignoring the "decision" of the customers. Using the k-NN algorithm, what do you expect the average no. of transactions of a customer to be, given that her average monthly payment of 380.072 and an average no. of months in Silver of 9.12?

c) If you are free to choose the value of k, what will your choice be, for example, for part (a)? Why?

(30 marks)

3. A study has been carried out to determine if a drug could be used effectively in a test for measuring a patient's risk of having a heart attack or cardiac event. A typical test of this risk takes various measurements, such as heart rate and blood pressure, as well as more complicated measurements of the heart. Assume you are hired as a data mining consultant for this research, you have been asked to determine if the death of a patient can be predicted accurately. To help you with your job, you are given some sample data in the attached database file (*cardiac.txt*). The data file contains a sample of the data obtained from a much larger population. The details of the data attributes are explained below.

Attribute	Description
bhr	BASAL HEART RATE
basebp	BASAL BLOOD PRESSURE
basedp	BASAL DOUBLE PRODUCT (= bhr x basebp)
pkhr	PEAK HEART RATE
sbp	SYSTOLIC BLOOD PRESSURE
dp	DOUBLE PRODUCT (= pkhr x sbp)
dose	DOSE OF DOBUTAMINE GIVEN
maxhr	MAXIMUM HEART RATE
%mphr(b)	% OF MAXIMUM PREDICTED HEART RATE ACHIEVED BY PATIENT
mbp	MAXIMUM BLOOD PRESSURE
dpmaxdo	DOUBLE PRODUCT ON MAXIMUM DOBUTAMINE DOSE
dobdose	DOBUTAMINE DOSE AT WHICH MAXIMUM DOUBLE PRODUCT OCCURED
age	PATIENT'S AGE
gender	PATIENT'S GENDER (male = 0)
baseEF	BASELINE CARDIAC EJECTION FRACTION (a measure of the heart's pumping efficiency)
dobEF	EJECTION FRACTION ON DOBUTAMINE
chestpain	0 MEANS THE PATIENT EXPERIENCED CHEST PAIN
posECG	SIGNS OF HEART ATTACK ON ECG
equivecg	ECG IS EQUIVOCAL
restwma	CARDIOLOGIST SEES WALL MOTION ANAMOLY ON ECHOCARDIOGRAM
posSE	STRESS ECHOCARDIOGRAM WAS POSITIVE
newMI	NEW MYOCARDIAL INFARCTION, OR HEART ATTACK
newPTCA	RECENT ANGIOPLASTY
newCABG	RECENT BYPASS SURGERY
death	THE PATIENT DIED
hxofHT	PATIENT HAS HISTORY OF HYPERTENSION
hxofdm	PATIENT HAS HISTORY OF DIABETES
hxofcig	PATIENT HAS HISTORY OF SMOKING
hxofMI	PATIENT HAS HISTORY OF HEART ATTACK
hxofPTCA	PATIENT HAS HISTORY OF ANGIOPLASTY
hxofCABG	PATIENT HAS HISTORY OF BYPASS SURGERY

a) Without pre-processing your data, use a decision-tree based algorithm provided by Scikit-Learn package, what can you discover in it?

b) Now preprocess the data set with your own observations. What preprocessing steps would you go through to improve accuracy of the classification process? Why? Is the result obtained indeed better than that without pre-processing as in a)?

(30 marks)

**** END ****

*** *Please submit your report in either MS Word or PDF format. Other format will not be accepted.*