Student ID_____ Student Name_____

CISC3023 Machine Learning

1. Let us apply the naive Bayes classifier to the problem of classifying days according to whether someone will play tennis. Table below provides a set of 14 historical examples of the target concept PlayTennis, where each day is described by the attributes Outlook, Temperature, Humidity, and Wind. Please use naive Bayes classifier (do not use the "imaginary" samples) and the historical data from this table to classify the following novel instance; the target PlayTennis should be Yes or No? (**Show every step of your calculation**)
   (Outlook = rain, Temperature = mild, Humidity = normal, Wind = strong)

Day	*Outlook*	*Temp*	*Humidity*	*Wind*	*Play Tennis*
Dl	Sunny	Hot	High	Weak	**No**
D2	Sunny	Hot	High	Strong	**No**
D3	Overcast	Hot	High	Weak	**Yes**
D4	Rain	Mild	High	Weak	**Yes**
D5	Rain	Cool	Normal	Weak	**Yes**
D6	Rain	Cool	Normal	Strong	**No**
D7	Overcast	Cool	Normal	Strong	**Yes**
D8	Sunny	Mild	High	Weak	**No**
D9	Sunny	Cool	Normal	Weak	**Yes**
D10	Rain	Mild	Normal	Weak	**Yes**
D11	Sunny	Mild	Normal	Strong	**Yes**
D12	Overcast	Mild	High	Strong	**Yes**
D13	Overcast	Hot	Normal	Weak	**Yes**
D14	Rain	Mild	High	Strong	**No**

P(play=yes)=9/14
P(play=no)=5/14

P(yes)P(rain|yes)P(mild|yes)P(normal|yes)P(strong|yes)
=9/14 X 3/9 X 4/9 X 6/9 X 3/9=0.0211640212
It is larger than:
P(no)P(rain|no)P(mild|no)P(normal|no)P(strong|no)
=5/14 X 2/5 X 2/5 X 1/5 X 3/5=0.00685714286

So the answer is Yes.

Student ID_____ Student Name_____

2. Suppose that training examples are given in the following table. The last column is used to predict whether a particular type of rice will be good for making Sushi or not.

rice #	length	width	weight	color	Good for Sushi?
1	long	wide	light	white	yes
2	short	wide	heavy	yellow	no
3	short	narrow	light	white	yes
4	long	normal	light	red	no
5	long	normal	heavy	white	no

a) Use **Find-S** Algorithm to find the most specific hypothesis that satisfies above training examples **Show every step of your derivation**. (20/100)

S1=<l w l w>
S2=<l w l w>
S3=<? ? l w>
S4=<? ? l w>
S5=<? ? l w>

b) Use **candidate-elimination algorithm** to derive the final **version space** from the given examples. **Show every step of your derivation (20/100)**

S0=<ø ø ø ø > G0=<? ? ? ?>
S1=<l w l w> G1=<? ? ? ?>
S2=<l w l w> G2=<l ? ? ?> <? ? l ?><? ? ? w>
S3=< ? ? l w> G3=<? ? l ?><? ? ? w>
S4=<? ? l w> G4=<? ? ? w>
S5=<? ? l w> G5=<? ? l w>
Version space <? ? l w>

c) Calculate the entropy of current collection **Entropy(S),** and Information Gains: ***Gain(S, weight),*** i.e., the information gain if we set **'weight'** as the root node of a decision tree (20/100)

Entropy(s)=-2/5*log2(2/5)-3/5*log2(3/5)=0.97

Entropy(weight=light)=-1/3*log2(1/3)-2/3*log2(2/3)=0.92

Entropy(weight=heavy)=-0/2*log2(0/2)-2/2*log2(2/2)=0

Gain(S, weight)=Entropy(s)-3/5 Entropy(weight=light)-2/5 Entropy(weight=heavy)=0.42

Student ID_____ Student Name_____

3. Suppose that 5 cases are given in the case base in following table.

Case ID	*x1*	*x2*	*x3*	Output f(X)
X_1	7	5	3	2
X_2	3	9	4	10
X_3	10	3	2	6
X_4	4	9	7	2
X_5	5	2	10	8

a) For a query $\mathbf{x_q = (4, 8, 4)}$, calculate the estimated output $\mathbf{f(x_q)}$ using "Distance-weighted 3-Nearest Neighbor algorithm for discrete-valued target function". (In this case, the last column in above table shows the discrete outputs. Then the outputs like **2** and **10** are discrete class labels, **not the numeric numbers**) (20/100)

$di=|xi-xq|^{1/2}$
So according to di^2, we know 3NN of xq are x1 x2 x4

Weight for x1: $w1= (1/d1)^2 =1/[(7-4)^2+(5-8)^2+(3-4)^2]=1/19$
Weight for x2: $w2= (1/d2)^2 =1/[(3-4)^2+(9-8)^2+(4-4)^2]=1/2$
Weight for x4: $w4= (1/d4)^2 =1/[(4-4)^2+(9-8)^2+(7-4)^2]=1/10$

Outputs for x1, x4 are label 2, their weights together is 1/19+1/10
Output for x2 is label 10, its weight is ½
Then the weight for output 10 is larger, output for xq is 10

b) For a query $\mathbf{x_q = (4, 8, 4)}$, calculate the estimated output $\mathbf{f(x_q)}$ using "Distance-weighted 3-Nearest Neighbor algorithm for real-valued target function". (the last column in above table shows the numeric outputs this time) (20/100)

Output(xq)=[w1 f(x1)+w2 f(x2)+w4 f(x4)]/(w1+w2+w4)=[1/19*2+1/2*10+1/10*2]/(1/19+1/2+1/10)=8.1