694Z: Introduction to Data Mining: Midterm: Autumn 2013

Instructions **NAME**_____

Answer all questions
State and underline any assumptions
Use backside of sheets for rough work

I. Data Preprocessing (15 minutes – 15 points)

A For a convergence threshold of 0.2 solve the following missing data problem using the EM algorithm. You are asked to compute an estimate of the mean for the following dataset containing eight elements of which 3 are missing. The data is {1, 5, 9, 4, 20,x,y,z}. Your initial guess for the mean should be 5. Show all steps.

$\hat{\mu}_1 = \frac{1+5+9+4+20}{8} + \frac{5+5+5}{8} = 6.75$

$\hat{\mu}_2 = \quad " \quad + \frac{6.75+6.75+6.75}{8} = 7.4065$

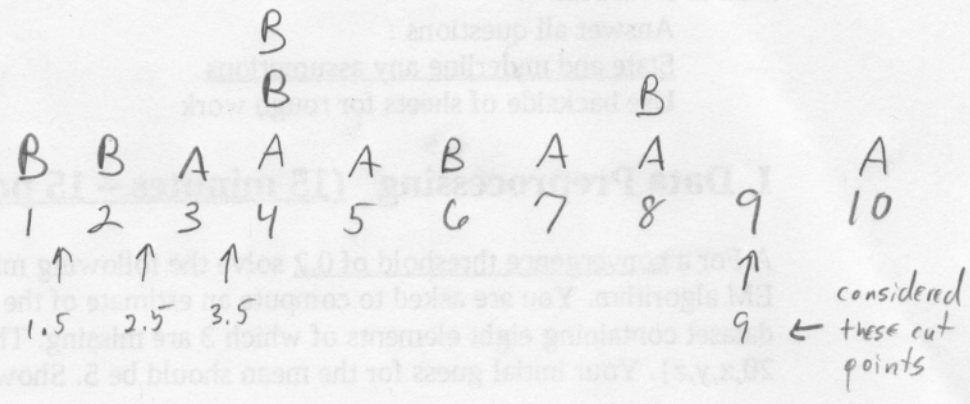
$\hat{\mu}_3 = \quad " \quad + \frac{(7.4065 * 3)}{8} = 7.6523$

$\hat{\mu}_4 = \quad " \quad + \frac{(7.6523 * 3)}{8} = 7.7496$

$(\hat{\mu}_4 - \hat{\mu}_3) = 0.09 < 0.2$.

B. For the following dataset:

AttrI	Class
4	A
6	B
8	A
7	A
4	B
3	A
2	B
1	B
10	A
8	B
4	B
5	A



Find the optimal cutpoint (split into two intervals) for AttrI using entropy as your basis for discretization.

We can see H(split 1.5) = H(split 9) because one node A:6, B:5 or vice-versa. Both have more entropy than H(split 2.5), so try that:

split at 2.5: <2.5 A:0, B:2 $\quad >=2.5$ A:6, B:4

$$H(\text{split } 2.5) = -\frac{10}{12}\left(\frac{6}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2\left(\frac{4}{10}\right)\right) = .809$$

weight for node size

split at 3.5: <3.5 A:1, B:2 $\quad >=3.5$ A:5, B:4

$$H(\text{split } 3.5) = -\frac{3}{12}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) - \frac{9}{12}\left(\frac{5}{9}\log_2\left(\frac{5}{9}\right) + \frac{4}{9}\log_2\left(\frac{4}{9}\right)\right)$$

$$= .973$$

Looks like 2.5 is the best cut point

II. Classification (35 minutes 35 points)

You are on an island. To survive you must eat mushrooms indigenous to the island. You landed as a party of eight, five of your party are ill. You have the following data to consider. 1 represents yes and 0 represents no in this table.

Example	IsHeavy	IsSmelly	IsSpotted	IsSmooth	IsPoisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A-H are poisonous, but you do not know about U through W.

A. What is the entropy of IsPoisonous?

Entropy $= -\frac{3}{8}\log\frac{3}{8} - \frac{5}{8}\log\frac{5}{8}$

~~$= 0.1597 + 0.1276$~~

$= 0.9544$

B. Which attribute do you choose as the root of the decision tree? Hint: you should be able to figure this out on just inspecting the table.

IsSmooth → because it results in 3 misclassification while all others result in 4.

C. What is the resulting information gain from choosing the above attribute (from your answer to part b)

D. Build an ID3 decision tree from the training dataset and classify U, V and W.

$Entropy_{root} = 0.9544$

If split on IsHeavy, we get

IsHeavy	0	1
0	2	1
1	3	2

$Entr_0 = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.9709$

$Entr_1 = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.9183$

$Gain = 0.9544 - \frac{5}{8} \times 0.9709 - \frac{3}{8} \times 0.9183$

$= 0.0032$

If split on IsSmelly, we get

IsSmelly	0	1
0	2	1
1	3	2

$Entr_0 = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$ = same as IsHeavy

$Entr_1$ = Same as IsHeavy

Gain = Same as IsHeavy

Split on IsSpotted,

IsSpotted	0	1
0	2	1
1	3	2

$Entr_0$, $Entr_1$, Gain → Same as above.

Split on IsSmooth:

IsSmooth	0	1
0	2	1
1	2	3

$Entr_0 = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

$Entr_1 = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.8113$

$Gain = 0.9544 - \frac{4}{8} \times 1 - \frac{4}{8} \times 0.8113 = 0.0487$ Max Gain

So split root with IsSmooth.

On the left child, Entropy = $Entr_0 = 1$

Split left child on IsHeavy:

	IsHeavy 0	IsHeavy 1
0	2	0
1	1	1

$$Entr_0 = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9183$$

$Entr_1 = 0$

$$Gain = 1 - \frac{3}{4} \times 0.9183 = 0.3113$$

Split left child (IsSmooth) on IsSmelly:

	IsSmelly 0	IsSmelly 1
0	2	0
1	0	2

$Entr_0 = 0$, $Entr_1 = 0$, $Gain = 1$ ← max Gain

Split on IsSpotted:

	0	1
0	1	1
1	1	1

* we can understand the max. gain will be from IsSmelly

So we split left child on IsSmelly.

The entropy of the resulting children are 0.

(entr: 0.8113)

Now, for right child, at the split on IsSmooth:

Split on IsHeavy:

	0	1
0	0	1
1	2	1

$Entr_0 = 0$, $Entr_1 = 1$, $Gain = 0.8113 - \frac{2}{4} \times 0 - \frac{2}{4} \times 1$

$= 0.3113$

Split on IsSmelly:

	0	1
0	0	1
1	3	0

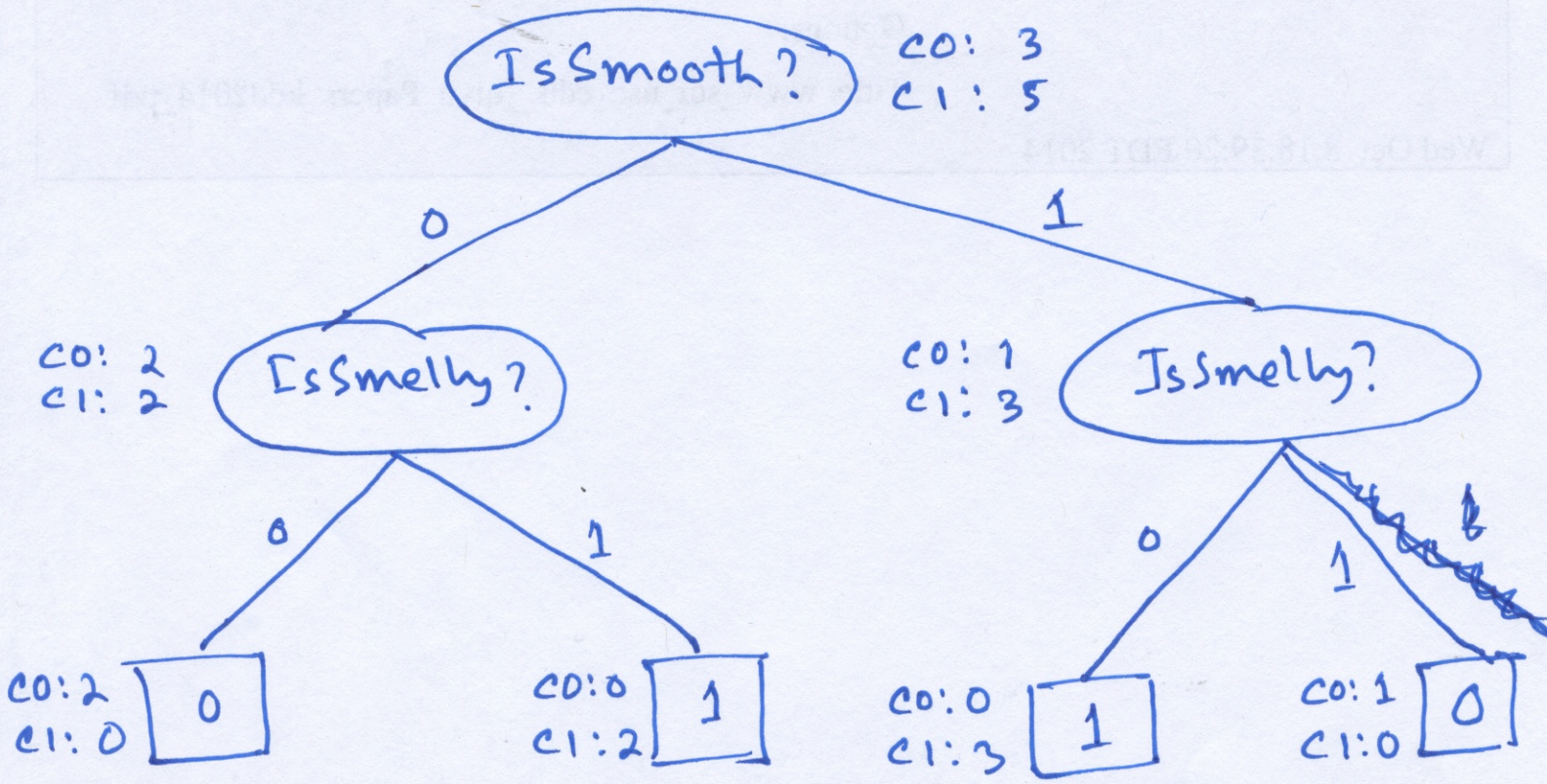
$Entr_0 = 0$, $Entr_1 = 0$, $Gain = 0.8113$

Split on IsSpotted,

	0	1
0	1	0
1	2	1

* Gain will be lower.

So split right child after IsSmooth or IsSmelly.
Now we' have all leaves with entropy 0.
The decision tree:



Classification of U, V, W:

U : 0 (IsPoisonons = 0)

V : 0

W : 1

E. Build a Naïve Bayesian classifier from the training dataset and classify U, V and W.

Heavy Poisonous

$P(H=0 \mid P_o=0) = \frac{2}{3}$, $P(H=1 \mid P_o=0) = \frac{1}{3}$, $P(H=0 \mid P_o=1) = \frac{3}{5}$, $P(H=1 \mid P_o=1) = \frac{2}{5}$

$P(S_{me}=0 \mid P_o=0) = \frac{2}{3}$, $P(S_{me}=1 \mid P_o=0) = \frac{1}{3}$, $P(S_{me}=0 \mid P_o=1) = \frac{3}{5}$, $P(S_{me}=1 \mid P_o=1) = \frac{2}{5}$

$P(S_{po}=0 \mid P_o=0) = \frac{2}{3}$, $P(S_{po}=1 \mid P_o=0) = \frac{1}{3}$, $P(S_{po}=0 \mid P_o=1) = \frac{3}{5}$, $P(S_{po}=1 \mid P_o=1) = \frac{2}{5}$

$P(S_{mo}=0 \mid P_o=0) = \frac{2}{3}$, $P(S_{mo}=1 \mid P_o=0) = \frac{1}{3}$, $P(S_{mo}=0 \mid P_o=1) = \frac{2}{5}$, $P(S_{mo}=1 \mid P_o=1) = \frac{3}{5}$

$P(P_o=0) = \frac{3}{8}$, $P(P_o=1) = \frac{5}{8}$

$P(U \mid P_o=0)\, P(P_o=0) = \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{3}{8}\right) = .0046$

$P(U \mid P_o=1) P(P_o=1) = \left(\frac{2}{5}\right)\left(\frac{2}{5}\right)\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)\left(\frac{5}{8}\right) = .024$

$P(V \mid P_o=0)\, P(P_o=0) = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{3}{8}\right) = .012$

$P(V \mid P_o=1)\, P(P_o=1) = \left(\frac{3}{5}\right)\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)\left(\frac{3}{5}\right)\left(\frac{5}{8}\right) = .054$

$P(W \mid P_o=0)\, P(P_o=0) = \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{3}{8}\right) = .019$

$P(W \mid P_o=1)\, P(P_o=1) = \left(\frac{2}{5}\right)\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)\left(\frac{5}{8}\right) = .024$

III. Clustering (30 points, 30 minutes)

This question relates to the BUPA Liver Disorders dataset:

Relevant information:
-- The first 5 variables are all blood tests which are thought
to be sensitive to liver disorders that might arise from
excessive alcohol consumption. Each line in the bupa.data file
constitutes the record of a single male individual.

Number of instances: 345

Number of attributes: 7 overall

Attribute information:
1. mcv mean corpuscular volume
2. alkphos alkaline phosphotase
3. sgpt alamine aminotransferase
4. sgot aspartate aminotransferase
5. gammagt gamma-glutamyl transpeptidase
6. drinks number of half-pint equivalents of alcoholic beverages drunk per day
7. selector field used to split data into two sets (those with disorders versus those without disorders)

Missing values: none

To better understand this dataset you are asked to cluster the dataset excluding the seventh attribute. You may assume that the first six attributes are continuous attributes.

SAMPLE FROM LIVER DATASET

	Mcv	alkphos	sgpt	sgot	gamgt	drinks	selector
PT1	85,	92,	45,	27,	31,	0.0,	1
PT2	85,	54,	47,	33,	22,	0.5,	2
PT3	96,	67,	26,	26,	36,	0.5,	2
PT4	98,	99,	57,	45,	65,	20.0,	1
PT5	91,	57,	33,	23,	12,	8.0,	1
PT6	91,	63,	25,	26,	15,	6.0,	1

NOTE: FOR ANSWERING PARTS A AND B YOU ARE SUPPOSED TO IGNORE THE 7^{th} ATTRIBUTE (selector).

A. Distance Metrics

The first step in a clustering algorithm is to propose a distance metric.

i) Compute the normalized Eucledian distance between PT1 and PT4 – use min-max normalization on the sample data such that each attribute has a value between 0 and 1.

The min-max normalization is done on each attribute, and the formula is $value_{new} = [value_{old} - min(value)]/[max(value) - min(value)]$. For example, the normalized Mcv value for PT5 becomes (91 - 85) / (98 - 85) = 0.46.
The data table after normalization (not done on selector) is:

	Mcv	alkphos	sgpt	sgot	gamgt	drinks	selector
PT1	0.00	0.84	0.63	0.18	0.36	0.00	1
PT2	0.00	0.00	0.69	0.45	0.19	0.03	2
PT3	0.85	0.29	0.03	0.14	0.45	0.03	2
PT4	1.00	1.00	1.00	1.00	1.00	1.00	1
PT5	0.46	0.07	0.25	0.00	0.00	0.40	1
PT6	0.46	0.20	0.00	0.14	0.06	0.30	1

Therefore, the Euclidian distance between

PT1 (0.00, 0.84, 0.63, 0.18, 0.36, 0.00)
and
PT4 (1.00, 1.00, 1.00, 1.00, 1.00, 1.00) is

sqrt[(0.00-1.00)^2 + (0.84-1.00)^2 + (0.63-1.00)^2 + (0.18-1.00)^2 + (0.36-1.00)^2 + (0.00-1.00)^2]
= sqrt(3.2458)
= 1.8016

B. Clustering Algorithms

i) Using the Manhattan distance metric cluster the above sample using a hierarchical single-link clustering algorithm. Build the complete dendrogram.

The Manhattan **distance** matrix of those points after normalization (with the smallest non-diagonal distance boldfaced) is:

	PT1	PT2	PT3	PT4	PT5	PT6
PT1	0.00	1.40	2.18	3.97	2.58	2.40
PT2	1.40	0.00	2.37	4.64	1.98	2.07
PT3	2.18	2.37	0.00	4.22	1.79	1.18
PT4	3.97	4.64	4.22	0.00	4.82	4.85
PT5	2.58	1.98	1.79	4.82	0.00	**0.68**
PT6	2.40	2.07	1.18	4.85	**0.68**	0.00

Therefore PT5 and PT6 will be merged first. The distance matrix is updated to:

	PT1	PT2	PT3	PT4	PT5, PT6
PT1	0.00	1.40	2.18	3.97	2.40
PT2	1.40	0.00	2.37	4.64	1.98
PT3	2.18	2.37	0.00	4.22	**1.18**
PT4	3.97	4.64	4.22	0.00	4.82
PT5, PT6	2.40	1.98	**1.18**	4.82	0.00

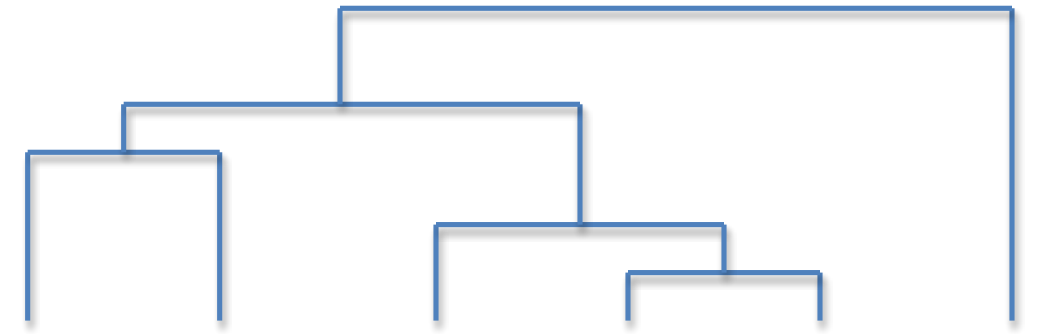
Therefore PT3 will be merged into the cluster of PT5 and PT6. The distance matrix becomes:

	PT1	PT2	PT4	PT3, PT5, PT6
PT1	0.00	**1.40**	3.97	2.18
PT2	**1.40**	0.00	4.64	1.98
PT4	3.97	4.64	0.00	4.22
PT3, PT5, PT6	2.18	1.98	4.22	0.00

The next step is to merge PT1 and PT2, and the new distance matrix is:

	PT1, PT2	PT4	PT3, PT5, PT6
PT1, PT2	0.00	3.97	**1.98**
PT4	3.97	0.00	4.22
PT3, PT5, PT6	**1.98**	4.22	0.00

The two clusters ({PT1,PT2} and {PT3,PT5,PT6}) are merged, and finally PT4 will be merged with others. The final dendrogram looks like (note the ordering of points)



C. Clustering Quality: If the desired number of clusters is two for the above algorithm evaluate the quality of the resulting clusters using entropy/information gain. Give the formulae for the initial entropy (before clustering) and the final entropy (after clustering) and compute the difference to identify the gain.

The two clusters resulting from the hierarchical clustering are {PT4} and {PT1, PT2, PT3, PT5, PT6}.
The entropy before clustering is $-[2/6 * \log(2/6) + 4/6 * \log(4/6)] = 0.92$.
The entropy of cluster {PT4} is 0.
The entropy of cluster {PT1, PT2, PT3, PT5, PT6} is $-[2/5 * \log(2/5) + 3/5 * \log(3/5)] = 0.97$.
The final entropy after clustering is $(1/6 * 0 + 5/6 * 0.97) = 0.81$.
Therefore the information gain in this case is 0.92-0.81 = 0.11.