CPSC 340 Midterm Examination
(50 minutes)

Allowed: scientific calculator
Total Points: 20
Number of pages: 2

1. Consider the following training relation:

Salary	Education	Label
10,000	Undergraduate	Reject
40,000	Undergraduate	Accept
19,000	Graduate	Accept
18,000	Undergraduate	Reject
75,000	Graduate	Accept
15,000	Graduate	Accept

(a) (3 points) Using gini index as the impurity function, show the reduction of impurity if the split point is "**Salary** ≤ 15,000" .

(b) (1 point) If the attribute **Salary** is to be used as the splitting attribute, which value will give the highest reduction? No need to show derivations; just state your answer.

Answer:

(a) Let p1 be accept and p2 be reject. Without the split, p1 = 2/6 and p2 = 4/6.
Gini = 1 – 2/6 * 2/6 – 4/6 * 4/6 = 4/9.
With the split at 15,000, for the left branch:
Gini = 1 – ½ * ½ - ½ * ½ = ½
For the right branch, gini = 1 – ¾ * ¾ - ¼ * ¼ = 3/8
Gini reduction = 4/9 – ½ * 2/6 – 3/8 * 4/6 = 1/36

(b) between 18,000 and 19,000. Use 18,500.

2. Consider the training relation from the previous question. This time we use 2-fold cross validation to estimate the sensitivity and precision of the tree built by the training relation. In this case, the positives correspond to Label = Accept, and the negatives correspond to Reject.

(a) (3 points) The first fold consists of the first 3 rows. Show the decision tree built by the first fold using gini index as the impurity function. (You don't need to show the derivations; just show the tree.) Show how the confusion matrix looks like when applied to the second fold consisting of the last 3 rows.

(b) (2 points) Complete the 2-fold cross validation by repeating (a) on the second fold. Show the final confusion matrix.

(c) (2 points) Give the 2-fold cross validation sensitivity and precision of the tree built by the entire training relation.

Answer:

(a) For the first fold, the tree consists of a single node, which is salary <= 14500, reject. Otherwise, accept. 4th row is a FP. 5th and 6th row are TP.

(b) For the second fold, the tree consists of a single node, which is education = Graduate, accept. Otherwise, reject. First row is TN. Second row is FN. 3rd row is TP.

(c) sensitivity = TP / (TP + FN) = ¾
 precision = TP /(TP + FP) = ¾

3. The following table gives the pairwise distance between 8 data objects.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	0	4	6	14	18	13	28	20
B	4	0	4	10	14	9	24	16
C	6	4	0	8	12	11	22	14
D	14	10	8	0	6	13	14	14
E	18	14	12	6	0	7	10	8
F	13	9	11	13	7	0	15	7
G	28	24	22	14	10	15	0	8
H	20	16	14	14	8	7	8	0

(a) (3 points) Suppose in the first iteration of the **k-medoids** clustering algorithm, the 3 points C, G, H are chosen as the medoids to form the initial clusters. What are the clusters corresponding to C, G and H?

(b) (2 points) For the initial clusters in (a), what are the new medoids? Show your derivations.

Answer:

(a) For A: C
For B: C
For D: C
For E: H
For F: H
So clusters are {A, B, C, D}, {E, F, H} and {G}.

(b) For the {A, B, C, D} cluster:
A as medoid = 4 + 6 + 14 = 24
B as medoid = 4 + 4 + 10 = 18
C as medoid = 6 + 4 + 8 = 18
D as medoid = 14 + 10 + 8 = 32
Thus, either B and C can be the medoid for the first cluster.

For the {E, F, H) cluster:
E as medoid = 15, F as medoid = 14, H as medoid = 15. So F is the medoid.

For the{G} cluster, G is the medoid.

4. Consider the following 6 data objects with a numeric attribute X and a categorical attribute Y.

Object	**X**	**Y**
A	1	**β**
B	2	**α**
C	4	**β**
D	5	**α**
E	7	**β**
F	8	**α**

(a) (2 points) For k = 2, use **k-means** clustering to form clusters based on attribute X only. What are the two clusters formed if the initial clusters are based on A and F? Compute the entropy based on attribute Y.
(b) (1 point) For k = 3, use **k-means** clustering to form three clusters based on attribute X only. What are the three clusters if the initial clusters are based on A, D and F? Compute the entropy based on attribute Y.
(c) (1 point) Does a larger value of k always give a smaller entropy for k-means clustering? What about hierarchical clustering? Provide explanations to your answers.

Answer:
(a) {A,B,C} {D,E,F} form 2 clusters and is a local optimum.
For {A,B,C}, entropy = - (0.67 log 0.67 + 0.33 log 0.33) = 0.92
For {D,E,F}, entropy = 0.92
Thus, for k = 2, the combined entropy is 3/6 * 0.92 + 3/6 * 0.92 = 0.92
(b) {A,B} {C,D} {E,F} form 3 clusters and is a local optimum.
For each of the 3 clusters, the entropy = - (0.5 log 0.5 + 0.5 log 0.5) = 1
Thus, for k = 3, the combined entropy is 2/6 * 1 + 2/6 * 1 + 2/6 *1 = 1
(c) No, k = 2 gives a smaller entropy than k = 3 in the above example. In hierachical clustering, it is different because it chooses a cluster to split in each step, always making entropy non-increasing. In k-means, the new cluster need not be a subset of any cluster from the previous configuration.