Student ID:

Name:

Question 1: Consider the following data set:

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

(a) We are trying to predict ’profitable’, please illustrate the steps to select the root in a decision tree if we use multi-way splits and the Gini index impurity measure? (8 points)

(b) For the same data set, suppose we decide to construct a decision tree using binary splits and the entropy impurity measure. Which among the following feature and split point combinations would be the best to use as the root node assuming that we consider each of the input features to be unordered? We only consider the following four choices ((1)price - {low, med}|{high} (2) maintenance - {high}|{med, low} (3) maintenance - {high, med}|{low} (4) capacity - {2}|{4, 5}) (8 points)

Solution 1:

(a) $gini_{price}(D) = \frac{4}{11}[1-(\frac{2}{4})^2-(\frac{2}{4})^2] + \frac{4}{11}[1-(\frac{2}{4})^2-(\frac{2}{4})^2] + \frac{3}{11}[1-(\frac{2}{3})^2-(\frac{1}{3})^2]$

$= \frac{2}{11} + \frac{2}{11} + \frac{4}{33} = \frac{16}{33} = 0.485$

$gini_{maintenance}(D) = \frac{2}{11}[1-(\frac{2}{2})^2-(\frac{0}{2})^2] + \frac{4}{11}[1-(\frac{2}{4})^2-(\frac{2}{4})^2] + \frac{5}{11}[1-(\frac{2}{5})^2-(\frac{3}{5})^2]$

$= \frac{2}{11} + \frac{12}{55} = \frac{22}{55} = \frac{2}{5} = 0.4$

$gini_{capacity}(D) = \frac{3}{11}[1-(\frac{1}{3})^2-(\frac{2}{3})^2] + \frac{6}{11}[1-(\frac{3}{6})^2-(\frac{3}{6})^2] + \frac{2}{11}[1-(\frac{2}{2})^2-(\frac{0}{2})^2]$

$= \frac{4}{33} + \frac{3}{11} = \frac{13}{33} = 0.394$ ✓

$gini_{airbag}(D) = \frac{5}{11}[1-(\frac{3}{5})^2-(\frac{2}{5})^2] + \frac{6}{11}[1-(\frac{3}{6})^2-(\frac{3}{6})^2] = \frac{12}{55} + \frac{3}{11} = \frac{27}{55} = 0.491$

Hence, I would select the attribute of 'capacity' as the root in a decision tree.

(b) cross entropy$_{price(\{low, med\}|\{high\})}(D) = \frac{8}{11}(-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}) + \frac{3}{11}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) = 0.9777$

cross entropy$_{maintenance(\{high\}|\{med, low\})}(D) = \frac{5}{11}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{6}{11}(-\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}) = 0.9422$

cross entropy$_{maintenance(\{high, med\}|\{low\})}(D) = \frac{9}{11}(-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}) + \frac{2}{11}(-1\log 1 - 0\log 0) = 0.8109$ ✓

cross entropy$(\{2\}|\{4,5\})(D) = \frac{3}{11}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) + \frac{8}{11}(-\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8}) = 0.9446$

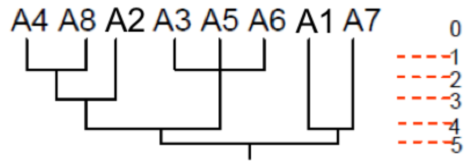
Hence, the 3rd choice is the best one.

<u>Question 2:</u> Use single-link, complete-link, average-link agglomerative clustering to cluster the following 8 examples: A1=(2,5), A2=(2,10), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Show the steps and dendrograms. (18 points)

Solution 2:

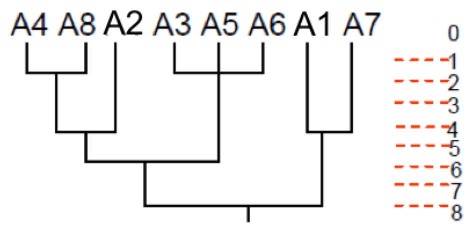
Single Link:

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A2}, {A3, A5, A6}, {A1}, {A7}
4	2	{A2, A3, A4, A5, A6, A8}, {A1, A7}
5	1	{A2, A3, A4, A5, A6, A8, A1, A7}



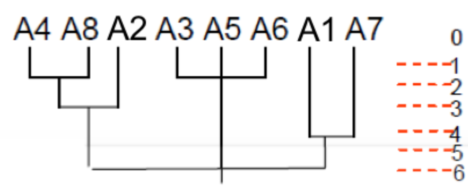
Complete Link

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A2}, {A3, A5, A6}, {A1, A7}
5	3	{A4, A8, A2}, {A3, A5, A6}, {A1, A7}
6	2	{A4, A8, A2, A3, A5, A6}, {A1, A7}
7	2	{A4, A8, A2, A3, A5, A6}, {A1, A7}
8	1	{A4, A8, A2, A3, A5, A6, A1, A7}



Average Link

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A2}, {A3, A5, A6}, {A1}, {A7}
4	3	{A4, A8, A2}, {A3, A5, A6}, {A1, A7}
5	3	{A4, A8, A2}, {A3, A5, A6}, {A1, A7}
6	1	{A4, A8, A2, A3, A5, A6, A1, A7}

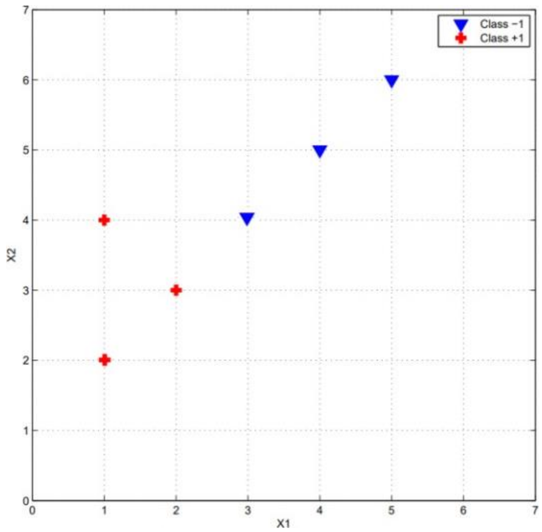


Average distance from {A3, A5, A6} to {A2, A4, A8} is 5.53 and is 5.75 to {A1, A7}

Question 3: Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 6 points shown in the following Figure. This dataset consists of three examples with class label -1 (denoted with plus), and three examples with class label +1 (denoted with triangles).

(a) Find the weight vector w and bias b. What's the equation corresponding to the decision boundary? (10 points)

(b) Circle the support vectors and draw the decision boundary. (6points)



Solution 3: (a) SVM tries to maximize the margin between two classes. Therefore, the optimal decision boundary is diagonal, and it crosses the point (2.5, 3.5). It is perpendicular to the line between support vectors (3,4) and (2,3), hence it is slope is m = -1. Thus the line equation is $(x2 - 3.5) = -1(x1 - 2.5) = x1 + x2 = 6$. From this equation, we can deduce that the weight vector has to be of the form (w1, w2), where w1 = w2. It also has to satisfy the following equations:

$2w_1 + 3w_2 + b = 1$ and

$3w_1 + 4w_2 + b = -1$

Hence $w_1 = w_2 = -1$ and $b = 6$

(b) Circle the support vectors and draw the decision boundary.