

CS480/580: Fall 2009

Solutions for HW#6

Q1.

At the root we have [8 Reject, 9 Admit]. So  $p_{\text{Reject}} = 8/17$ ;  $p_{\text{accept}} = 9/17$ . So the entropy at root =  $-8/17 \cdot \log_2(8/17) - 9/17 \cdot \log_2(9/17) = 0.5117 + 0.4876 = 0.9973$

Splitting at Degree, we get MS node: [1,2] and BS node [7,7]. Entropy of BS node =  $-0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1.0$ ; Entropy of MS node =  $-1/3 \log_2(1/3) - 2/3 \log_2(2/3) = 0.9183$ . Gain =  $0.9973 - 0.9183 = 0.0790$

Splitting at Major, we get CS node: [6,6] and non-CS node [2,3]. Entropy of CS node =  $-0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1.0$ ; Entropy of non-CS node =  $-2/5 \log_2(2/5) - 3/5 \log_2(3/5) = 0.9710$ . gain =  $0.9973 - 0.9710 = 0.0263$ .

For GPA and GRE, we need to divide them into ranges: Let us say the ranges selected are:

GPA		GRE		Experience	
GPA1	< 2.5	GRE1	<1150	Exp1	0
GPA2	2.5 <= GPA <3.0	GRE2	1150 <= GRE < 1300	Exp2	1
GPA3	3.0 <= GPA < 3.5	GRE3	1300 <= GRE	Exp3	>1
GPA4	3.5 <= GPA				

GPA1 [0,0]; GPA2 [1,2]; GPA3 [5,5]; GPA4 [2, 2]: Entropy of GPA2 =0.9183; Entropy of GPA3 and GPA4 = 1.0; So Gain due to split with GPA =  $0.9973 - (3/17 \cdot 0.9183 + 10/17 \cdot 1 + 4/17 \cdot 1) = 0.9973 - 0.8764 = 0.1209$ .

GRE1[6,0]; GRE2[2,6]; GRE3[0,3]. Entropy of GRE1=0; Entropy of GRE3 = 0; Entropy of GRE2 =  $-0.25 \log_2(0.25) - 0.75 \log_2(0.75) = 0.5028$ ; Gain =  $0.9973 - 8/17 \cdot 0.5028 = 0.2366$

For recommendation, Strong[1,4]; Average[2,4]; Weak[5,1]; Entropy of Strong= $-0.2 \cdot \log_2(0.2) - 0.8 \cdot \log_2(0.8) = 0.7219$ ; Entropy of Average =  $-1/3 \cdot \log_2(1/3) - (2/3) \cdot \log_2(2/3) = 0.9184$ ; Entropy of Weak= $-(5/6) \cdot \log_2(5/6) - (1/6) \cdot \log_2(1/6) = 0.2630$ . Gain =  $0.9973 - (5/17 \cdot 0.7219 + 6/17 \cdot 0.9184 + 6/17 \cdot 0.2630) = 0.2060$ .

Experience: Exp1[7,5]; Exp2[1,1]; Exp3[0,3]; Entropy of Exp3 =0; Entropy of Exp2 = 1.0; Entropy of Exp1 =  $-7/12 \log_2(7/12) - (5/12) \cdot \log_2(5/12) = 0.9799$ ; Gain =  $0.9973 - (2/17 \cdot 1 + 12/17 \cdot 0.9799) = 0.1880$ .

Maximum gain is possible by splitting at GRE attribute. So GRE is the root of the decision tree.

Since GRE1 and GRE3 are leaf nodes, we don't need to go further. But GRE2 needs to be split. The data we have to deal with:

Instance#	Degree	Major	GPA	GRE	Recommendations	Experience (Years)	Outcome
3	BS	Non-CS	3.5	1250	Average	0	Admit
4	MS	Non-CS	3.0	1200	Average	2	Admit
5	BS	CS	2.9	1250	Strong	1	Admit
6	BS	Non-CS	2.95	1150	Strong	5	Admit
10	BS	CS	3.2	1150	Strong	3	Admit
11	MS	CS	3.1	1250	Average	0	Admit
12	BS	Non-CS	3.4	1200	Weak	0	Reject
16	BS	CS	3.4	1150	Weak	0	Reject

Repeating the above analysis for this data (excluding GRE attribute), we find that Recommendation should be the next attribute to select, since Strong[0,3], Average[0,3], Weak[2,0]. So Gain is maximum here and all these nodes are leaf nodes.

So the decision tree looks like this: (I do not have a graphic software at home. So I will just explain it in words):

GRE < 1150: **Reject**

GRE >= 1300 : **Admit**

1150 <= GRE < 1300: Recommendation: Average: **Admit**

Recommendation: Strong: **Admit**

Recommendation: Weak: **Reject**

Now, let us validate the tree based on the validation data: For each entry, predict the outcome based on the above tree.

Instance#	Degree	Major	GPA	GRE	Recommendations	Experience (Years)	Actual Outcome	Predicted Outcome
1	BS	CS	2.9	1150	Weak	0	Reject	Reject
2	BS	CS	3.5	1210	Average	0	Admit	Admit
3	BS	Non-CS	3.5	1150	Strong	0	Admit	Admit
4	BS	Non-CS	3.0	1100	Average	2	Reject	Reject
5	MS	CS	3.1	1150	Average	0	Admit	Admit
6	MS	Non-CS	3.5	1250	Average	1	Admit	Admit
7	BS	CS	3.2	1050	Strong	0	Reject	Reject
8	BS	CS	3.3	1450	Average	0	Admit	Admit
9	BS	CS	3.3	1100	Average	4	<b>Admit</b>	<b>Reject</b>

So, among the 9, there are: 5 True positives; 3 True Negatives; 1 false Negative;

We will now use it on the test data.

GRE < 1150: **Reject**

GRE >= 1300 : **Admit**

1150 <= GRE < 1300: Recommendation: Average: **Admit**

Recommendation: Strong: **Admit**

Recommendation: Weak: **Reject**

Instance#	Degree	Major	GPA	GRE	Recommendations	Experience (Years)	Outcome
1	BS	CS	3.5	1150	Weak	0	<b>Reject</b>
2	BS	CS	2.9	1110	Strong	6	<b>Reject</b>
3	BS	CS	3.7	950	Average	0	<b>Reject</b>
4	BS	CS	3.2	1150	Weak	5	<b>Reject</b>
5	MS	CS	3.0	1150	Average	0	<b>Admit</b>
6	MS	CS	3.1	1250	Weak	0	<b>Reject</b>
7	MS	Non-CS	3.2	1250	Average	1	<b>Admit</b>
8	BS	Non-CS	3.7	1050	Average	0	<b>Reject</b>
9	BS	Non-CS	3.3	1300	Average	2	<b>Admit</b>
10	BS	Non-CS	3.0	1450	Strong	0	<b>Admit</b>

Q2. Let us find cosine similarity between pairs from different clusters:

	<2,4,4,4>	<6,5,6,4>	<5,6,6,4>
<3, 5, 7, 3>	0.9542	0.9513	0.9710
<4, 6, 5, 4>	0.9788	0.9755	0.9951

$$\langle 3,5,7,3 \rangle \text{ and } \langle 2,4,4,4 \rangle: CS = \frac{(6+20+28+12)}{(\sqrt{92}) * \sqrt{52}} = \frac{66}{69.17} = 0.9542$$

$$\langle 3,5,7,3 \rangle \text{ and } \langle 6,5,6,4 \rangle: CS = \frac{(18+25+42+12)}{(\sqrt{92}) * \sqrt{113}} = \frac{97}{101.96} = 0.9513$$

$$\langle 3,5,7,3 \rangle \text{ and } \langle 5,6,6,4 \rangle: CS = \frac{(15+30+42+12)}{(\sqrt{92}) * \sqrt{113}} = \frac{99}{101.96} = 0.9710$$

$$\langle 4,6,5,4 \rangle \text{ and } \langle 2,4,4,4 \rangle: CS = \frac{(8+24+20+16)}{(\sqrt{93}) * \sqrt{52}} = \frac{68}{69.54} = 0.9778$$

$$\langle 4,6,5,4 \rangle \text{ and } \langle 6,5,6,4 \rangle: CS = \frac{(24+30+30+16)}{(\sqrt{93}) * \sqrt{113}} = \frac{100}{102.5} = 0.9755$$

$$\langle 4,6,5,4 \rangle \text{ and } \langle 5,6,6,4 \rangle: CS = \frac{(20+36+30+16)}{(\sqrt{93}) * \sqrt{113}} = \frac{102}{102.5} = 0.9951$$

Single link cosine similarity = Max of similarity = 0.9951

Complete link similarity = Min of similarity = 0.9513

Q3: Here  $K = 3$ . We need to randomly choose three seeds for the three clusters:

Let us choose:  $\langle 2, -5, -4 \rangle$  for cluster 1;  $\langle 7, 7, 7 \rangle$  for cluster 2; and  $\langle 0, 0, 2 \rangle$  for cluster 3.

Example: Euclidean distance between  $\langle 1, 5, 6 \rangle$  and C1 =  $\sqrt{1+100+100} = 14.18$

Fill the rest

	C1 $\langle 2, -5, -4 \rangle$	C2 $\langle 7, 7, 7 \rangle$	C3 $\langle 0, 0, 2 \rangle$	Min distance cluster
$\langle 1, 5, 6 \rangle$	14.18	6.40	6.48	C2
$\langle -2, 9, 8 \rangle$	18.87	9.27	11	C2
$\langle 2, -2, 3 \rangle$	7.62	11.05	3	C3
$\langle 3, 3, 3 \rangle$	10.54	6.93	9.06	C2
$\langle 4, 1, 2 \rangle$	8.72	8.37	4.58	C3
$\langle 6, 5, 4 \rangle$	13.42	3.74	8.06	C2
$\langle -2, 4, 5 \rangle$	13.34	9.70	5.39	C3
$\langle 1, 1, 1 \rangle$	7.87	10.39	1.73	C3
$\langle 2, -2, 2 \rangle$	6.71	11.45	3	C3
$\langle 0, 1, 1 \rangle$	8.06	11	1.41	C3
$\langle -1, -1, -1 \rangle$	5.83	13.86	3.32	C3

New cluster C1:  $\{ \langle 2, -5, -4 \rangle \}$

Cluster C2:  $\{ \langle 7, 7, 7 \rangle, \langle 1, 5, 6 \rangle, \langle -2, 9, 8 \rangle, \langle 3, 3, 3 \rangle, \langle 6, 5, 4 \rangle \}$ ; Its new centroid: Mean of all values:  $(15/5, 29/5, 28/5) = \langle 5, 5.8, 5.6 \rangle$

Cluster C3:  $\{ \langle 0, 0, 2 \rangle, \langle 2, -2, 3 \rangle, \langle 4, 1, 2 \rangle, \langle -2, 4, 5 \rangle, \langle 1, 1, 1 \rangle, \langle 2, -2, 2 \rangle, \langle 0, 1, 1 \rangle, \langle -1, -1, -1 \rangle \} = \langle 6/8, 2/8, 15/8 \rangle = \langle 0.75, 0.25, 1.875 \rangle$

We have to repeat the above step with the new centroids:

	C1 $\langle 2, -5, -4 \rangle$	C2 $\langle 5, 5.8, 5.6 \rangle$	C3 $\langle 0.75, 0.25, 1.875 \rangle$	Min distance cluster
$\langle 2, -5, -4 \rangle$	0			C1
$\langle 7, 7, 7 \rangle$				C2
$\langle 0, 0, 2 \rangle$				C3
$\langle 1, 5, 6 \rangle$				C2
$\langle -2, 9, 8 \rangle$				C2
$\langle 2, -2, 3 \rangle$				C3
$\langle 3, 3, 3 \rangle$				C2
$\langle 4, 1, 2 \rangle$	8.72	6.08	3.34	C3
$\langle 6, 5, 4 \rangle$				C2
$\langle -2, 4, 5 \rangle$	13.34	8.51	5.60	C3
$\langle 1, 1, 1 \rangle$				C3
$\langle 2, -2, 2 \rangle$	6.71	9.10	2.58	C3
$\langle 0, 1, 1 \rangle$				C3
$\langle -1, -1, -1 \rangle$	5.83	11.22	3.59	C3

New cluster C1:  $\{ \langle 2, -5, -4 \rangle \}$

New cluster C2:  $\{ \langle 7, 7, 7 \rangle, \langle 1, 5, 6 \rangle, \langle -2, 9, 8 \rangle, \langle 3, 3, 3 \rangle, \langle 6, 5, 4 \rangle \}$

New cluster C3:  $\{ \langle 0,0,2 \rangle, \langle 2,-2,3 \rangle, \langle 4,1,2 \rangle, \langle -2,4,5 \rangle, \langle 1,1,1 \rangle, \langle 2,-2,2 \rangle, \langle -1,-1,-1 \rangle \}$

So there is no change. This is the final clustering.