Max Shi

CS 559

Professor Wang

Homework 3

1. (1) $Gain = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$

$$Entropy(p) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.991$$

Attribute 1:

True

Positive	3
Negative	1

False

Positive	1
Negative	4

$$Entropy(True) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$$

$$Entropy(False) == -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$$

$$Gain = 0.991 - \frac{4}{9}(0.811) - \frac{5}{9}(0.722) = 0.229$$

Attribute 2:

	<=1.0	>1.0
Positive	1	3
Negative	0	5

$$Entropy(<= 1.0) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$$

$$Entropy(> 1.0) == -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.954$$

$$Gain = 0.991 - \frac{1}{9}(0) - \frac{8}{9}(0.954) = 0.143$$

	<=3.0	>3.0
Positive	1	3
Negative	1	4

$$Entropy(<= 3.0) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Entropy(> 3.0) == -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.923$$

$$Gain = 0.991 - \frac{2}{9}(1) - \frac{7}{9}(0.923) = 0.051$$

	<=4.0	>4.0
Positive	2	2
Negative	1	4

$$Entropy(<= 4.0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$$
$$Entropy(> 4.0) == -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.918$$
$$Gain = 0.991 - \frac{3}{9}(0.918) - \frac{6}{9}(0.918) = 0.0727$$

	<=5.0	>5.0
Positive	2	2
Negative	3	2

$$Entropy(\leq 5.0) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$
$$Entropy(> 5.0) == -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$
$$Gain = 0.991 - \frac{5}{9}(0.971) - \frac{4}{9}(1) = 0.00714$$

	<=6.0	>6.0
Positive	3	1
Negative	3	2

$$Entropy(<= 6.0) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$
$$Entropy(> 6.0) == -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$
$$Gain = 0.991 - \frac{6}{9}(1) - \frac{3}{9}(0.918) = 0.018$$

	<=7.0	>7.0
Positive	4	0
Negative	4	1

$$Entropy(<= 6.0) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$$
$$Entropy(> 6.0) == -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} = 0$$
$$Gain = 0.991 - \frac{8}{9}(1) - \frac{1}{9}(0) = 0.102$$

Thus, the best candidate for the first split is the split between true and false on attribute 1.

(2) If we use instance as an attribute, the decision tree will split based on the index of the row entry used. This is undesirable, as presumably, all subsequent entries will use increasing instance numbers, which will cause each subsequent point to use a decision tree based off a subset of the data instead of all the data, which is very undesirable. For example, if the split took place at instance value 5, the resulting decision tree would only take into account all values after instance 5, and all subsequent decisions made using the tree would only be trained with those values.

2. (1)$Gain = GINI(p) - \sum_{i=1}^{k}\frac{n_i}{n}GINI(i)$

$$GINI(p) = 1 - \frac{35}{100}^2 - \frac{65}{100}^2 = 0.455$$

Attribute A:

	A=T	A=F
Positive	20	15

Negative	30	35

$$GINI(A = T) = 1 - \frac{20^2}{50} - \frac{30^2}{50} = 0.48$$

$$GINI(A = F) = 1 - \frac{15^2}{50} - \frac{35^2}{50} = 0.42$$

$$Gain = 0.455 - \frac{50}{100} * 0.48 - \frac{50}{100} * 0.42 = 0.005$$

(2) Cost of matrix:

$$20 * -1 + 30 * 0 + 15 * 100 + 35 * -10 = 1130$$

Attribute B:

	B=T	B=F
Positive	15	20
Negative	20	45

$$GINI(B = T) = 1 - \frac{15^2}{35} - \frac{20^2}{35} = 0.490$$

$$GINI(B = F) = 1 - \frac{20^2}{65} - \frac{45^2}{65} = 0.426$$

$$Gain = 0.455 - \frac{35}{100} * 0.490 - \frac{65}{100} * 0.426 = 0.0066$$

(2) Cost of matrix:

$$15 * -1 + 20 * 0 + 20 * 100 + 45 * -10 = 1535$$

B should thus be chosen as the first splitting attribute based on the GINI index, but based on cost, A should be chosen as the first splitting attribute.

3. Weights start at 1/10.

H1: 9 and 10 not classified correctly.

$$err_1 = 0.1 * 2 = 0.2, \alpha_1 = \frac{1}{2}\ln\left(\frac{1 - 0.2}{0.2}\right) = 0.693$$

Correctly classified: $D_2(i) = 0.1 * e^{0.693*1} = 0.2$

Incorrectly classified: $D_2(i) = 0.1 * e^{0.693*-1} = 0.05$

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1
D2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.05	0.05

H2: points 1, 2, 3, 8 classified incorrectly.

$$err_1 = 0.1 * 4 = 0.4, \alpha_1 = \frac{1}{2}\ln\left(\frac{1 - 0.4}{0.4}\right) = 0.2027$$

Correctly classified: $D_2(i) = 0.1 * e^{0.2027*1} = 0.1225$

Incorrectly classified: $D_2(i) = 0.1 * e^{0.2027*-1} = 0.0816$

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1
D2	0.0816	0.0816	0.0816	0.1225	0.1225	0.1225	0.1225	0.0816	0.1225	0.1225

H3: point 9 classified incorrectly.

$$err_1 = 0.1 * 1 = 0.1, \alpha_1 = \frac{1}{2}\ln\left(\frac{1-0.1}{0.1}\right) = 1.099$$

Correctly classified: $D_2(i) = 0.1 * e^{1.099*1} = 0.3$
Incorrectly classified: $D_2(i) = 0.1 * e^{1.099*-1} = 0.033$

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1
D2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.033	0.3

4. (I) The five nearest neighbors to the triangle are the points at (4,4), (6,5), (7,3), (5,1), and (4,1). Thus, as these five nearest neighbors are classified as -, +, +, -, and – respectively, the triangle should be classified as -.
   The three nearest neighbors are the points at (4,4), (6,5) and (7,3). Their Manhattan distances are 1, 2, and 3 respectively. Using the weight of $1/d^2$, and the classification of -, +, and +, respectively, the calculation is:
   $$-\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} = -1 + \frac{1}{4} + \frac{1}{9} < 0$$
   Thus, this point is classified as -.
   (II)

	ID	**x**	**y**	**z**	**actual-class**	**three-nearest-predict**	**euclid-distance-predict**
0	1	8.074807	5.988044	3.844979	1	1	1
1	2	4.952249	5.823205	1.612045	0	0	0
2	3	4.773178	0.078757	4.209442	0	0	0
3	4	9.845919	2.055448	3.525702	1	1	1
4	5	1.612492	1.320515	8.200455	0	0	0
5	6	7.987555	9.188111	7.222228	1	1	1
6	7	0.311558	3.974680	7.897371	0	0	0
7	8	1.219113	0.266045	2.741136	0	0	0
8	9	0.636340	1.831257	6.767459	0	0	0
9	10	0.890168	8.613714	2.884227	0	0	0
10	11	7.226514	9.852794	7.373560	1	1	1
11	12	2.709551	3.719191	5.743540	0	0	0
12	13	2.842368	1.902145	2.216614	0	0	0
13	14	3.610773	4.589548	7.714008	0	0	0
14	15	4.888200	6.720637	7.261562	0	1	1
15	16	8.857224	9.056900	8.862604	1	1	1
16	17	8.239402	9.347802	5.277351	1	1	1

	ID	x	y	z	actual-class	three-nearest-predict	euclid-distance-predict
17	18	3.219759	2.980960	6.646886	0	0	0
18	19	2.146974	5.328725	5.801703	0	0	0
19	20	1.156302	8.542813	1.859447	0	0	0

(I) Find classifications above, and probability estimates below:

```
[[0.         1.        ]
 [0.66666667 0.33333333]
 [1.         0.        ]
 [0.         1.        ]
 [1.         0.        ]
 [0.         1.        ]
 [1.         0.        ]
 [1.         0.        ]
 [1.         0.        ]
 [1.         0.        ]
 [0.         1.        ]
 [1.         0.        ]
 [1.         0.        ]
 [1.         0.        ]
 [0.33333333 0.66666667]
 [0.         1.        ]
 [0.         1.        ]
 [1.         0.        ]
 [1.         0.        ]
 [1.         0.        ]]
```

(II) All predicted labels stay the same as question 1, as seen above.

(III) Both methods give equal performance, as the predicted labels stay the same.