Test #1

CS5402 — Intro To Data Mining

Illya Starikov

Due Date: July 15th, 2018

Multiple Choice

1. **e. None of the above**
2. **c. Remove any attribute that has missing values.**
3. **b.** $\frac{1}{2}$
4. **b. wt**
5. **d. Spearman's rank correlation coefficient**
6. **c. Healthland**
7. **b. slice for Time = Q1**
8. **d. roll up on Location = Beijing or Tokyo (i.e., from city to country)**
9. **c. drill down on Time = Q1 (i.e., from quarter to month)**
10. **a. dice for (location = Beijing or Tokyo) and (product = Chain or bracelet) and (time = Q1 or Q2)**

11 Short Answer

Method #1 is the most accurate, because the true positive (y-axis) correctly identified the values, while the false positive (x-axis) incorrectly identified the values. Method #1 had the fastest growing function (with respect to y).

12 1-R Method

Attribute	Attribute Value	# Rows With Attribute Value	Most Frequent Value For sportPref	Errors	Total Errors
ageGroup	youngAdult	3	football (2)	1	3
	middleAge	3	football/hockey/baseball (1/1/1)	2	
	senior	2	baseball (2)	0	
gender	M	5	baseball/football (2/2)	3	5
	F	3	football/hockey/baseball (1/1/1)	2	
petPreference	dog	5	football (3)	2	3
	cat	3	baseball (2)	1	

The rules are as follows:

$$\begin{aligned}
\text{ageGroup} = \textbf{youngAdult} &\implies \text{football} \\
\text{ageGroup} = \textbf{middleAge} &\implies \text{football} \\
\text{ageGroup} = \textbf{senior} &\implies \text{baseball}
\end{aligned}$$

13 Prism

For football, we get the following table:

gender	pet	drink	sport
M	dog	beer	football
F	dog	beer	football

For our P and T values:

	T	P	T/P
gender = M	3	1	1/3
gender = F	4	1	1/4
pet = dog	3	2	2/3
drink = beer	3	2	3/4

Seeing as not T/P values are 1, we must add a clause. We choose pet = dog as the base.

	T	P	T/P
gender = M	1	0	0
gender = F	1	0	0
drink = beer	2	2	1

$$\text{pet} = \textbf{dog} \text{ and drink} = \textbf{beer} \implies \text{football}$$

14 Statistical Modeling

The likelihood would be as follows:

$$\text{likelihood} = {}^4\!/_9 \times {}^2\!/_9 \times {}^6\!/_9 \times {}^3\!/_9 \times {}^9\!/_{14}$$

15 Entropy

(a) entropyBeforeSplit would be as follows:

$$-{}^1\!/_6 \log_2({}^1\!/_6) - {}^2\!/_6 \log_2({}^2\!/_6) - {}^3\!/_6 \log_2({}^3\!/_6)$$

(b) entropyPoor would be as follows:

$$-{}^2\!/_4 \log_2({}^2\!/_4) - {}^2\!/_4 \log_2({}^2\!/_4)$$

(c) infoGain would be determined as follows:

$$\begin{aligned} \text{entropyAfterSplit} &= {}^3\!/_6\,\text{entropyShort} + {}^2\!/_6\,\text{entropyMed} + {}^1\!/_6\,\text{entropyLong} \\ \text{infoGain} &= \text{entropyBeforeSplit} - \text{entropyAfterSplit} \end{aligned}$$

16 Rule Induction

(a) The partitions would be as follows:

$$\begin{aligned} \{d\}^* &= \{\{x_1\}, \{x_2, x_3\}, \{x_5\}, \{x_5\}\} \\ \{e\}^* &= \{\{x_1, x_2, x_5\}, \{x_3, x_4\}\} \\ \{d,\, e\}^* &= \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}\} \end{aligned}$$

(b) The coverings are as follows:

- $\{d\}^*$ would not work, because every block in the partition is not a subset of a block in $\{f\}^*$.
- $\{d, e\}^*$ would work, because every block in the partition is a subset of a block in $\{f\}^*$.
- $\{a, d, e\}^*$ would not work, because although every block in the partition is a subset of a block in $\{f\}^*$, it is not minimal.

(c) The rules would be as follows:

$$\begin{aligned}
d = \text{X and } e = 4 &\implies f = T \\
d = \text{S and } e = 4 &\implies f = T \\
d = \text{S and } e = 3 &\implies f = F \\
d = \text{H and } e = 3 &\implies f = F \\
d = \text{M and } e = 4 &\implies f = F
\end{aligned}$$

17 KD-Tree

Sorting, we get the following: [(2, 10), (4, 20), (6, 10), (8, 20), (10, 30)].
With a median of 6...

- $x < 6$ group: [(2, 10), (4, 20)]
- $x \geq 6$ group: [(6, 10), (8, 20), (10, 30)]

Sorting, we get the following: [(2, 10), (4, 20)] [(6, 10), (8, 20), (10, 30)]
With a median of 15 for the first group:

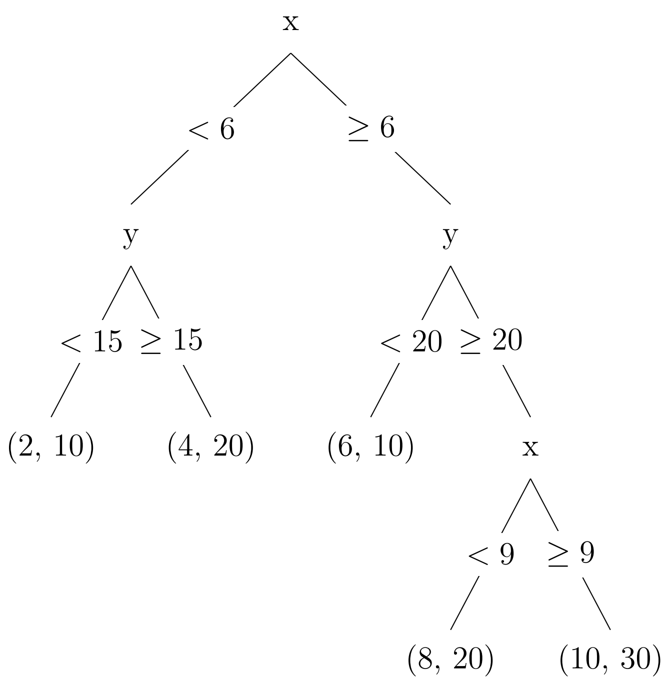
- $y < 15$ group: [(2, 10)]
- $y \geq 15$ group: [(4, 20)]

With a median of 20 for the second group:

- $y < 20$ group: (6, 10)
- $y \geq 20$ group: [(8, 20), (10, 30)]

(Using a shortcut for the final block), Sorting, and using a median of 9, our last block looks like as follows:

- $x < 9$ group: [(8, 20)]
- $y \geq 9$ group: [(10, 30)]



18 Clustering

x	y	distance to (2, 4)	distance to (5, 6)	distance to (8, 1)
2	4	0	5	9
5	6	5	0	8
8	1	9	8	0
7	3	6	5	3
4	10	8	5	13
3	0	5	8	6
9	8	11	6	8

Our clusters would be as follows:

Cluster Center (2, 4) (2, 4), (3, 0)

Cluster Center (5, 6) (5, 6), (4, 10), (9, 8)

Cluster Center (8, 1) (8, 1), (7, 3)

With means as follows:

Cluster Mean of (2, 4), (3, 0) (2.5, 2) $\approx$ (3, 2)

Cluster Mean of (5, 6), (4, 10), (9, 8) (6, 8)

Cluster Center of (8, 1), (7, 3) $(7.5, 2) \approx (8, 2)$

x	y	distance to (3, 2)	distance to (6, 8)	distance to (8, 2)
2	4	3	8	8
5	6	6	3	7
8	1	6	9	1
7	3	5	6	2
4	10	9	4	12
3	0	2	11	7
9	8	12	3	7

Cluster Center (3, 2) (2, 4), (3, 0)

Cluster Center (6, 8) (5, 6), (4, 10), (9, 8)

Cluster Center (8, 2) (8, 1), (7, 3)

Clusters haven't changed! Final cluster centers and instances are as follows:

Cluster Center (3, 2) (2, 4, 11, yes), (3, 0, 3, yes)

Cluster Center (6, 8) (5, 6, 5, no), (4, 10, 8, yes), (9, 8, 1, no)

Cluster Center (8, 2) (8, 1, 7, no), (7, 3, 4, yes)

19 Confusion Table

(a) For a randomly produced results, there were 8 values that we predicted to be B, when they were actually G.

(b) For a classifier produced results, there were 30 values that we predicted to be B, and were actually B.

(c) The non-random classifier, 90 were predicted correctly. For the random classifier, 39 were predicted correctly. Therefore, 51 more were predicted correctly.

(d) Kappa Statistic would be

$$\frac{\text{Non-Random Correct - Random Correct}}{\text{Total}}$$

Which would be as follows:

$$\frac{90 - 39}{100}$$