Test #2

CS5402 — Intro To Data Mining

Illya Starikov

Due Date: July 29th, 2018

1 C4.5

Split = 120

$$\text{EntropyLT} = -3/4 \log_2 (3/4) - 1/4 \log_2 (1/4)$$
$$\text{EntropyGt} = -2/5 \log_2 (2/5) - 3/5 \log_2 (3/5)$$
$$\text{Info Gain} = X - (4/9 \times \text{EntropyLT} + 5/9 \times \text{EntropyGT})$$

Split = 140

$$\text{EntropyLT} = -4/6 \log_2 (4/6) - 2/6 \log_2 (2/6)$$
$$\text{EntropyGt} = -2/3 \log_2 (2/3) - 1/3 \log_2 (1/3)$$
$$\text{Info Gain} = X - (6/9 \times \text{EntropyLT} + 3/9 \times \text{EntropyGT})$$

We choose the split that we calculated previously based on it's associated **highest information gain**. Values before it should be assigned *Less Than Split* and values after it should be assigned *Greater Than Split*.

2 Grouping Or Splitting

{{SciFiction, Mystery}, {NonFiction}}

$$\text{SciMystery} = -3/6 \log_2 (3/6) - 2/6 \log_2 (2/6) - 1/6 \log_2 (1/6)$$
$$\text{NonFiction} = -0 - 1/6 \log_2 (1/6) - 1/6 \log_2 (1/6)$$
$$\text{Info Gain} = 1.5575 - (6/8 \times \text{SciMystery} + 2/8 \times \text{NonFiction})$$

{{SciFiction, NonFiction}, {Mystery}}

$$\text{SciNonFiction} = -2/4 \log_2 (2/4) - 1/4 \log_2 (1/4) - 1/4 \log_2 (1/4)$$
$$\text{NonFiction} = -1/4 \log_2 (1/4) - 2/4 \log_2 (2/4) - 1/4 \log_2 (1/4)$$
$$\text{Info Gain} = 1.5575 - (4/8 \times \text{SciNonFiction} + 4/8 \times \text{Mystery})$$

{{Mystery, NonFiction}, {NonFiction}}

$$\begin{aligned}\text{SciNonFiction} &= -1/6 \log_2(1/6) - 3/6 \log_2(3/6) - 2/6 \log_2(2/6)\\ \text{NonFiction} &= -2/2 \log_2(2/2) - 0 - 0\\ \text{Info Gain} &= 1.5575 - (6/8 \times \text{SciNonFiction} + 2/8 \times \text{NonFiction})\end{aligned}$$

To determine if any attributes are to be grouped, take the **highest information gain**, and compare it to the Entropy Before Split for music preference. If information gain is higher, then the grouping was better; if not, the grouping was worse.

3 Support Vectors

The equations would reduce down to:

$$\begin{aligned}10\,\alpha_1 + 10\,\alpha_2 &= -1\\ 10\,\alpha_1 + 11\,\alpha_2 &= 1\end{aligned}$$

For which we get the following solutions:

$$\alpha_1 = -\frac{21}{10} \qquad \alpha_2 = 2$$

For the discriminating 2D hyperplane, we get as follows:

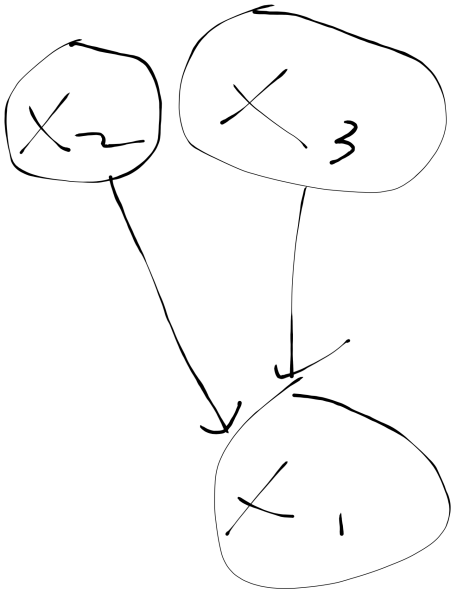
$$\begin{aligned}w &= -\frac{21}{10} \times s_1 + 2 \times s_2\\ &=< -\frac{3}{10}, 2, -\frac{41}{10} >\end{aligned}$$

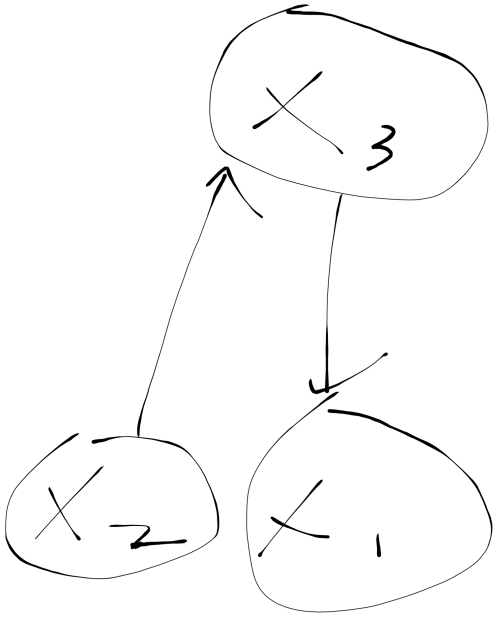
Meaning our equation would be as follows:

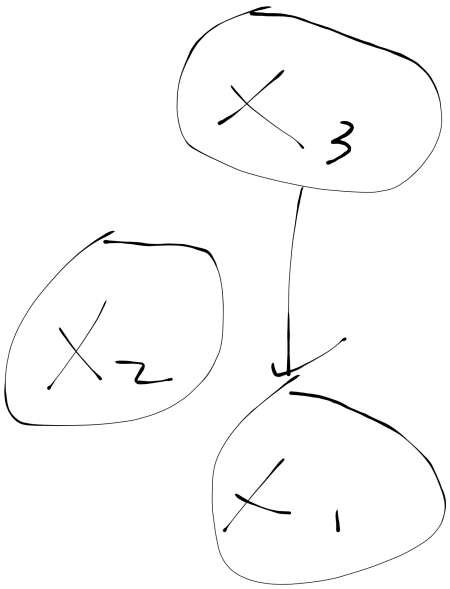
$$-3/10\,x - 2\,y - 41/10 = 0$$

4 Bayesian Network

The algorithm will enumerate the possible options (in the original ordering) for a parent of X2. It does so by calculating the probability of the particular attribute being a parent of X2, and the one with the greatest probability is the node that becomes the parents.

1. 

2. 



3.

5 DBScan

	density	designation
A1	2	Border
A2	2	Border
A3	3	Core
A4	1	Noise
A5	2	Border
A6	2	Border
A7	2	Noise
A8	1	Noise
A9	2	Noise
A10	3	Core

The clusters would be as follows: $\{A1, A2, A3\}$ and $\{A5, A6, A10\}$. Points with have are dense (density > number of points) are designated as core points. Points that are around a core points (within an ϵ of a core point) are designated as border point. Points that fall in neither of these categories are designated as noise.

6 Linkage

The single linkage would be as follows:

$$|(1,\, 4) - (3,\, 5)| = 3$$

The complete linkage would be as follows:

$$|(1,\, 2) - (4,\, 5)| = 6$$

The centroid linkage would be as follows:

$$|(1,\, {}^{11}\!/_{4}) - ({}^{14}\!/_{4},\, 5)| = {}^{19}\!/_{4}$$

The average linkage would be as follows:

$$\text{Average of all distances between points} = 5$$

7 Ensemble Classifier

Sum	-2	4	-6	2
Class	-1	1	-1	1

8 Bayes Network

For **worker = T**, the likelihood would be as follows:

$$0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.01$$

For **worker = F**, the likelihood would be as follows:

$$0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.99$$

The probability for **worker = T** would be:

$$\frac{0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.01}{0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.01 + 0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.99}$$

And the probability for **worker = F** would be:

$$\frac{0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.99}{0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.01 + 0.8 \times 0.02 \times 0.1 \times 0.1 \times 0.99}$$

Multiple Choice

9. d. repeatedly sampling from the original dataset according to a uniform probability distribution

10. c. increased, decreased

11. a. bias, variance

12. b. false

13. c. 180 instances reached this point in the decision tree, but 22 of those were not classified as tested_negative in the training dataset

14. d. age $>$ 34

15. b. compute the predicted error rate of the rule with one condition (C1, C2, C3) deleted and no conditions deleted, and, from those, use the version of the rule with the lowest rate

16. c. each instance in the dataset will have a probability of being in a particular cluster

17. b. supervised