CS 422: Data Mining

Solution to Spring 2018: Homework 2

Saptarshi Chatterjee
`CWID: A20413922`

March 3, 2018

1 Tan, Chapter 3 , Exercises 8, 9, 10

Q.8 Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

Answer -

If the length of the whiskers and the 2 halves of the box are exactly same then the data is symmetric. If the line representing the median of the data is in the middle of the box, then at-least $\frac{3}{4}^{th}$ of the data between the first and third quartiles is symmetrically distributed.

Sepal length and width are relatively symmetrically distributed. Though Sepal length are slightly skewed toward 4th quartiles. Petal width is slightly asymmetric having more values in second quartiles .Petal length is asymmetric having way more values in 2nd quartiles than any other .

Q.9.Compare sepal length, sepal width, petal length, and petal width, using Figure 3.

Answer -

Setosa, $\rightarrow$ sepal length $>$ sepal width $>$ petal length $>$ petal width { Though a few sepal width is larger than sepal length }

Versicolour $\rightarrow$ sepal length $>$ petal length $>$ sepal width $>$ petal width, { Though a few overlap between sepal length, petal length, sepal width }

Virginiica $\rightarrow$ sepal length $>$ petal length $>$ sepal width $>$ petal width, { Though

a lot of overlap between sepal length and petal length}

Q.10.Comment on the use of a box plot to explore a data set with four attributes: age, weight, height, and income

Answer -
(I) Individual box plots about those attributes will tell us the Max and Min values of those attributes and how they are distributed.
(II) Across attribute, if we take age in the Y axis and dwaw box plots for Height and Weight on X , we will see how Weight and height is increasing with age .

2 Tan, Chapter 4. Exercise 2, 3, 5.

Q.2 . Consider the training examples shown in Table 4.7 for a binary classification problem.?

Q.2(a) Compute the Gini index for the overall collection of training examples.?

Answer - Overall Gini $= 1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = 0.5$

Q.2(b) Compute the Gini index for the Customer ID attribute?

Answer -

$$\begin{aligned} CustomerIDGini &= \sum_{eachnode} \text{Weighted Average the node} \times \text{CustomerID Gini of the node} \\ &= \sum_{eachnode} \text{Weighted Average the node} \times 0 \\ &= 0 \end{aligned}$$

Q.2(c) Compute the Gini index for the Gender attribute.?

Answer - $\frac{10}{20} \times \{1 - (\frac{6}{10})^2 - (\frac{4}{10})^2\} + \frac{10}{20} \times \{1 - (\frac{6}{10})^2 - (\frac{4}{10})^2\} = 0.48$

Q.2(d) Compute the Gini index for the Car Type attribute using multiway split.?

Answer -

$$
\begin{aligned}
GiniofCar &= Family + Sports + Luxury \\
&= \frac{4}{20}(1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) + 0 + \frac{8}{20}(1 - (\frac{1}{8})^2 - (\frac{7}{8})^2) \\
&= .075 + .0875 = .1625
\end{aligned}
$$

Q.2(e) Compute the Gini index for the Shirt Size attribute using multiway split?

Answer -

$$
\begin{aligned}
GiniofShirt &= Small + Medium + Large + ExtraLarge \\
&= \frac{5}{20}(1 - (\frac{3}{5})^2 - (\frac{2}{5})^2) + \frac{7}{20}(1 - (\frac{3}{7})^2 - (\frac{4}{7})^2) + \frac{4}{20}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) + \frac{4}{20}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) \\
&= 0.12 + 0.171 + 0.1 + 0.1 = .491
\end{aligned}
$$

Q.2(f) Which attribute is better, Gender, Car Type, or Shirt Size?

Answer - Car type has the lowest Gini of .1625 of these 3 attributes , so It's better than rest of the two.

Q.2(g) Explain why Custoner ID should not be used as the attribute test condition even though it has the lowest Gini..?

Answer - Selecting Customer Id will be overfitting as each record has unique Cust ID. So it will fail to predict any outcome.

Q.3. Consider the training examples shown in T able 4.8 for a binary classification problem?

Q.3. (a) What is the entropy of this collection of training examples with respect to the positive class?

Answer - Entropy $= -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.99107$

Q.3. (b) What are the information gains of a_1 and a_2 relative to these training examples??

Answer -

$$\text{Entropy for } a_1 = \frac{4}{9}\Big[-(\frac{3}{4}\log_2\frac{3}{4}) - (\frac{1}{4}\log_2\frac{1}{4})\Big] + \frac{5}{9}\Big[-(\frac{1}{5}\log_2\frac{1}{5}) - (\frac{4}{5}\log_2\frac{4}{5})\Big]$$
$$= 0.7616$$

$$\Delta_{info} \quad a_1 = 0.99107 - 0.7616$$
$$= 0.22947$$

$$\text{Entropy for } a_2 = \frac{5}{9}\Big[-(\frac{2}{5}\log_2\frac{2}{5}) - (\frac{3}{5}\log_2\frac{3}{5})\Big] + \frac{4}{9}\Big[-2(\frac{2}{4}\log_2\frac{2}{4})\Big]$$
$$= 0.9839$$

$$\Delta_{info} \quad a_2 = 0.99107 - 0.9839$$
$$= 0.00717$$

Q.3. (c)For a_3, which is a continuous attribute, compute the information gain for every possible split.?

Answer -

Split Point	Entropy Equation	Entropy	Infogain
2	$\frac{1}{9}[-(1\log_2 1)] + \frac{8}{9}\Big[-(\frac{3}{8}\log_2\frac{3}{8}) - (\frac{5}{8}\log_2\frac{5}{8})\Big]$	0.8483	0.99107 - 0.8483 = 0.1427
3.5		0.9885	0.0026
4.5	$\frac{3}{9}[-(\frac{2}{3}\log_2\frac{2}{3}) - (\frac{1}{3}\log_2\frac{1}{3})] + \frac{6}{9}\Big[-(\frac{2}{6}\log_2\frac{2}{6}) - (\frac{4}{6}\log_2\frac{4}{6})\Big]$	0.9182	0.99107 - 0.9182 = 0.07287
5.5		0.9839	0.0072
6.5	$\frac{6}{9}[-2(\frac{3}{6}\log_2\frac{3}{6})] + \frac{3}{9}\Big[-(\frac{2}{3}\log_2\frac{2}{3}) - (\frac{1}{3}\log_2\frac{1}{3})\Big]$	0.9728	0.99107 - 0.9728 = 0.01827
7.5		0.8889	0.1022

Q.3. (d) What is the best split (among a1, a2, and a3) according to the information gain?

Answer - If we split on attribute a_1 we get highest info gain 0.22947 amongst the three .

Q.3. (e) What is the best split (between a1 and a2) according to the classification error rate?

Answer -

Error rate for $a_1 = 2/9$. Error rate for $a_2 = 4/9$. So according to error rate, a_1 produces the best split.

Q.3. (f) What is the best split (between a1 and a2) according to the Gini index?

Answer -

Gini Index for $a_1 = \frac{4}{9}(1 - (\frac{3}{4})^2 - (\frac{1}{4})^2) + \frac{5}{9}(1 - (\frac{1}{5})^2 - (\frac{4}{5})^2) = 0.3444$

Gini Index for $a_2 = \frac{5}{9}(1 - (\frac{2}{5})^2 - (\frac{3}{5})^2) + \frac{4}{9}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) = 0.4889$
As gini index for a_1 is smaller, it produces the better split.

Q.5.Consider the following data set for a binary class problem.

Q.5. (a)Calculate the information gain when splitting on ,A and B. Which attribute would the decision tree induction algorithm choose?

Answer -
The overall entropy = -0.4log0.4-0.6log0.6=0.9710

Info gain if Splitting on A
$\Delta = 0.9710 - \{7/10(-4/7log4/7 - 3/7log3/7) + 3/10(-3/3log3/3)\}$
$\Delta = 0.9710 - (7/10) * 0.9852$
$\Delta = 0.2813$

Info gain if Splitting on B
$\Delta = 0.9710 - \{4/10(-3/4log3/4 - 1/4log1/4) + 6/10(-1/6log1/6 - 5/6log5/6)\}$
$\Delta = 0.9710 - (4/10) * 0.8113 - (6/10) * 0.65$
$\Delta = 0.2565$

Info gain if split on A is high therefore A is chosen for split.

Q.5. (b) Calculate the gain in the Gini index when splitting on .A and B. Which attribute would the decision tree induction algorithm choose?

Answer - The overall gini before splitting = $1 - 0.4^2 - 0.6^2 = 0.48$

The gain in gini after splitting on A
Δ_A = 0.48 - (7/10)*0.48 - 0 = 0.1371

The gain in gini after splitting on B
Δ_A = 0.48 - (4/10)*0.3750 - (6/10)*0.2778 = 0.1633
Therefore, attribute B will be chosen to split the node as it has higher Gini

Q.5. (c)Figure 4.13 showsthat entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain?

Answer -

Yes though Entropy and Gini both monotonously increase in the range [0, 0.5] and [0.5, 1] from above two calculations of 5(a) and 5(b) we see that it is indeed possible that information gain and the gain in the Gini index favor different attributes.

3 Tan, Chapter 5 , Exercises 20

Q.20 Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes,"+" and "-." Half of the data set is used for training while the remaining half is used for testing

Q.20. (a)Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

Answer -

Expected error rate will be 0.5 or 50%

Q.20.(b)Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

Answer -

	Actual +	Actual -
Predicted +	$.5 \times .8 = .4$	$.8 \times .5 = .4$
Predicted -	$.2 \times .5 = .1$	$.2 \times .5 = .1$

Error rate $= .4 + .1/1 = .5 = 50\%$

Q.20.(c)Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

Answer -

	Actual +	Actual -
Predicted +	$1 \times 2/3 = 2/3$	$1/3$
Predicted -	0	0

Error rate $= (1/3)/1 = .333 = 33.3\%$

Q.20.(c)Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3

Answer -

	Actual +	Actual -
Predicted +	$4/9$	$2/9$
Predicted -	$2/9$	$1/9$

Error rate $= (2 * 2)/9 = .444 = 44.4\%$

4 Zaki, Chapter 8 , Exercises 1, 4, 6(a), 6(b).

Q.1.a (a) Using minsup = 3/8, show how the Apriori algorithm enumerates all frequent patterns from this dataset.

Answer - Frequent pattern rough set

Frequent pattern final

Q.1.b. With minsup = 2/8, show how FPGrowth enumerates the frequent itemsets.

Answer -

Q4. Given the database in Table 8.4. Show all rules that one can generate from the set ABE.

Answer - Total 8 rules can be generated.

Q6. (a) What is the size of the item set search space if one restricts oneself to only itemsets composed of simple items?.

Answer - size of the item set search space $= 2^{11} - 1$

Total Simple item count = 11.
Entire item space = All 1 item set + all 2 item set + ... + all 11 item set
$= {}^{11}C_1 + {}^{11}C_2 + \cdots + {}^{11}C_{11} = 2^{11} - 1$

Q6.(b) Let X = x1,x2,...,xk be a frequent item set. Let us replace some xi $\in$ X with its parent in the taxonomy (provided it exists), to obtain X?, then the support of the new item set X'? is

Answer -

If x_i is part of frequent item set then the frequency of It's parent $freq(parent(x_i)) \geq Freq(x_i)$. So support of X' would definitely $\geq$ X. So right answer is (iv) more than or equal to support of X