Chapter 10

1. What is the essential difference between association rules and decision rules?

Decision rules are based on IF-THEN checks at each node. Which means that the tree looks at the information it is trying to classify and looks to see IF the information provided for an attribute is a certain value, THEN it will predict the resulting class. Association rules are more about popularity and frequency meaning that if the attributes for a classification are more often a certain way for a class it will decide to predict that class. For example, if someone usually buys an item from a store with another item, the classifier could try to predict that the customer will also buy that item.

1. Given a simple transactional database X:

|  |  |  |
| --- | --- | --- |
| X: | TID | Items |
|  | T01 | A, B, C, D |
|  | T02 | A, C, D, F |
|  | T03 | C, D, E, G, A |
|  | T04 | A, D, F, B |
|  | T05 | B, C, G |
|  | T06 | D, F, G |
|  | T07 | A, B, G |
|  | T08 | C, D, F, G |

Using the threshold values support = 25% and confidence = 60%, find:

A) All large item sets in database X.

|  |  |  |  |
| --- | --- | --- | --- |
| Item sets C1 | Count | S[%] | Which are large |
| A | 5 | 62.5 | 1 |
| B | 4 | 50 | 1 |
| C | 5 | 62.5 | 1 |
| D | 5 | 62.5 | 1 |
| E | 1 | 12.5 | 0 |
| F | 4 | 50 | 1 |
| G | 5 | 62.5 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Item sets C2 | Count | S[%] | Which are large |
| A, B | 3 | 37.5 | 1 |
| A, C | 3 | 37.5 | 1 |
| A, D | 4 | 50 | 1 |
| A, F | 2 | 25 | 1 |
| A, G | 2 | 25 | 1 |
| B, C | 2 | 25 | 1 |
| B, D | 2 | 25 | 1 |
| B, F | 1 | 12.5 | 0 |
| B, G | 2 | 25 | 1 |
| C, D | 4 | 50 | 1 |
| C, F | 2 | 25 | 1 |
| C, G | 3 | 37.5 | 1 |
| D, F | 4 | 50 | 1 |
| D, G | 3 | 37.5 | 1 |
| F, G | 2 | 25 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Item sets C3 | Count | S[%] | Which are large |
| A, B, C | 1 | 12.5 | 0 |
| A, B, D | 2 | 25 | 1 |
| A, B, G | 1 | 12.5 | 0 |
| A, C, D | 3 | 37.5 | 1 |
| A, C, F | 1 | 12.5 | 0 |
| A, C, G | 1 | 12.5 | 0 |
| A, D, F | 2 | 25 | 1 |
| A, D, G | 1 | 12.5 | 0 |
| B, C, D | 1 | 12.5 | 0 |
| B, C, G | 1 | 12.5 | 0 |
| C, D, F | 2 | 25 | 1 |
| C, D, G | 2 | 25 | 1 |
| D, F, G | 2 | 25 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Item sets C4 | Count | S[%] | Which are large |
| C, D, F, G | 1 | 12.5 | 0 |

All large data sets {A}, {B}, {C}, {D}, {F}, {G}, {A, B}, {A, C}, {A, D}, {A, F}, {A, G}, {B, C}, {B, D}, {B, G}, {C, D}, {C, F}, {C, G}, {D, F}, {D, G}, {F, G}, {A, B, D}, {A, C, D}, {A, D, F}, {C, D, F}, {C, D, G}, {D, F, G}

B) Strong association rules for database X.

c({A} -> B) = s(A, B) / s(A) = 3/5 = 0.6 y

c({A} -> C) = s(A, C) / s(A) = 3/5 = 0.6 y

c({A} -> D) = s(A, D) / s(A) = 4/5 = 0.8 y

c({A} -> F) = s(A, F) / s(A) = 2/5 = 0.4

c({A} -> G) = s(A, G) / s(A) = 2/5 = 0.4

c({B} -> C) = s(B, C) / s(B) = 2/4 = 0.5

c({B} -> D) = s(B, D) / s(B) = 2/4 = 0.5

c({B} -> G) = s(B, G) / s(B) = 2/4 = 0.5

c({C} -> D) = s(C, D) / s(C) = 4/5 = 0.8 y

c({C} -> F) = s(C, F) / s(C) = 2/5 = 0.4

c({C} -> G) = s(C, G) / s(C) = 4/5 = 0.8 y

c({D} -> F) = s(D, F) / s(D) = 4/5 = 0.8 y

c({D} -> G) = s(D, G) / s(D) = 3/5 = 0.6 y

c({F} -> G) = s(F, G) / s(F) = 2/4 = 0.5

c({A, B} -> D) = s(A, B, D) / s(A, B) = 2/3 = 0.6667 y

c({B, D} -> A) = s(A, B, D) / s(B, D) = 2/2 = 1 y

c({A, C} -> D) = s(A, C, D) / s(A, C) = 3/3 = 1 y

c({C, D} -> A) = s(A, C, D) / s(C, D) = 3/4 = 0.75 y

c({A, D} -> F) = s(A, D, F) / s(A, D) = 2/4 = 0.5

c({D, F} -> A) = s(A, D, F) / s(D, F) = 2/4 = 0.5

c({C, D} -> F) = s(C, D, F) / s(C, D) = 2/4 = 0.5

c({D, F} -> C) = s(C, D, F) / s(D, F) = 2/4 = 0.5

c({C, D} -> G) = s(C, D, G) / s(C, D) = 2/4 = 0.5

c({D, G} -> C) = s(C, D, G) / s(D, G) = 2/3 = 0.6667 y

c({D, F} -> G) = s(D, F, G) / s(D, F) = 2/4 = 0.5

c({F, G} -> D) = s(D, F, G) / s(F, G) = 2/2 = 1 y

All rules that have y next to it are strong association

A->B

A->C

A->D

C->D

C->G

D->F

D->G

AB->D

BD->A

AC->D

CD->A

DG->C

FG->D

The rules above are the strong associations, if it has FG it is likely to have D is an example of the last rule

C) Analyze misleading associations for the rule set obtained in (b)

Some misleading rules would be the rule of having FG in the data set is likely to have D and having DG is likely to have C in the data set. These are misleading because the rules don’t necessarily work if the dataset has the reverse of what the rule states. Meaning that if the data set has CD it does not likely contain CDG. This is misleading since there is a rule that states essentially the opposite but only if the data set has DG instead of CD.

1. Assume that we have a data set containing information about 200 individuals. One hundred of these individuals have purchased life insurance. A supervised data-mining session has discovered the following rule:  
   If age < 30 & credit card insurance = yes  
   THEN life insurance = yes  
   (Rule Accuracy = 70%, Rule Coverage = 63%)  
     
   How many individuals in the class life insurance = no have credit card insurance and are less than 30 years old?

63% of 200 is 126

30% of 126 is 37.8

Round down since I don’t think you have a fraction of an individual so 37 individuals individuals in class life insurance have no credit card insurance and are less then 30

Chapter 11

1. Given a table of linked web pages

|  |  |
| --- | --- |
| Page | Linked to Page |
| A | B, D, E, F |
| B | C, D, E |
| C | B, E, F |
| D | A, F, E |
| E | B, C, F |
| F | A, B |

Diagram, line chart

Description automatically generated

1. Find authorities using two iterations of the HITS algorithm

Authority matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F |
| A | 0 | 1 | 0 | 1 | 1 | 1 |
| B | 0 | 0 | 1 | 1 | 1 | 0 |
| C | 0 | 1 | 0 | 0 | 1 | 1 |
| D | 1 | 0 | 0 | 0 | 1 | 1 |
| E | 0 | 1 | 1 | 0 | 0 | 1 |
| F | 1 | 1 | 0 | 0 | 0 | 0 |

At K = 1

|  |  |
| --- | --- |
|  | Authority |
| A | 2 |
| B | 4 |
| C | 2 |
| D | 2 |
| E | 4 |
| F | 4 |

New authority at K = 2

2^2 + 4^2 + 2^2 + 2^2 + 4^2 + 4^2 = 60

At K = 2

|  |  |
| --- | --- |
|  | Authority |
| A | 2/ 0.258 |
| B | 4/0.516 |
| C | 2/ 0.258 |
| D | 2/ 0.258 |
| E | 4/0.516 |
| F | 4/0.516 |

1. Find hubs using two iterations of the HITS algorithm

At start

|  |  |
| --- | --- |
|  | Hub |
| A | 4 |
| B | 3 |
| C | 3 |
| D | 3 |
| E | 3 |
| F | 2 |

Authority matrix \* authority vector

For K = 1

|  |  |
| --- | --- |
|  | Hub |
| A | 12 |
| B | 6 |
| C | 10 |
| D | 8 |
| E | 10 |
| F | 6 |

Calc K=2

12^2 + 6^2 + 10^2 + 8^2 + 10^2 + 6^2 = 480

|  |  |
| --- | --- |
|  | Hub |
| A | 12/0.5477 |
| B | 6/0.2738 |
| C | 10/0.456 |
| D | 8/0.365 |
| E | 10/0.456 |
| F | 6/0.2738 |

1. FIND the PageRank scores for each page after one iteration using 0.1 as the dampening factor.

Pr(A) = 1/6 = 0.1667

Pr(B) = 0.1667

Pr(C) = 0.1667

Pr(D) = 0.1667

Pr(E) = 0.1667

Pr(F) = 0.1667

K=1

Pr(x) = + d(

Pr(A) = + 0.1( + = 0.15 + 0.01389 = 0.16389

Pr(B) = + 0.1( + = 0.15 + 0.0236 = 0.1736

Pr(C) = + 0.1( + = 0.15 + 0.01111 = 0.16111

Pr(D) = + 0.1( + = 0.15972

Pr(E) = + 0.1( + = 0.15 + 0.0208375 = 0.1708375

Pr(F) = + 0.1( + = 0.15 + 0.0152808 = 0.1652808

1. Explain the HITS and Page Rank authority ranking obtained in a and c.

In both a and c B, E, and F are the highest ranked or largest in terms of authority. However in part a they are all tied for their authority while in part c there was a hierarchy between the three that was very close. This shows that these three pages are likely the most important in the list of pages since both methods point to that being the case. In terms of which pages are more important between the three themselves we most likely have to look at the results of the PageRank algorithm. In that case the rank has B, E, and F in that order of rank.

1. For the traversal log: {X, Y, Z, W, Y, A, B, C, D, Y, C, D, E, F, D, E, X, Y, A, B, M, N}

A picture containing sky, line, colorful

Description automatically generated

1. Find maximal forward references.

MFR = {XYZW, XYABCDEF, XYABMN}

1. Find large references sequences if the threshold value is 0.3

Lengths 2 and 3

LRS = {XY, YA, AB, BC, CD, XYA, YAB, YCD, CDE}

1. Find maximal reference sequences.

MRS = {YAB, XY, CDE}

1. Given the table of linked web pages and a dampening factor of 0.15

|  |  |
| --- | --- |
| Page | Linked to Page |
| A | F |
| B | F |
| C | F |
| D | F |
| E | A, F |
| F | E |

1. Fine the Page Rank scores for each page after one iteration.

Since there are 6 pages each has a score of 1/6 initially and distributes these scores the pages they are linked to after the iteration.

Pr(A) = + 0.15( = 0.141666 + 0.01333 = 0.154996

Pr(B) = + 0.15(0 = 0.141666

Pr(C) = + 0.15(0 = 0.141666

Pr(D) = + 0.15(00.141666

Pr(E) = + 0.15( = 0.141666 + 0.025 = 0.166666

Pr(F) = + 0.15(= 0.141666 + 0.1125 = 0.254166

A gets half of 1/6

E gets 1/6 from F

F gets (1/6 \* 4) + 1/12 = 9/12

|  |  |
| --- | --- |
| Page | Score |
| A | 0.154996 |
| B | 0.141666 |
| C | 0.141666 |
| D | 0.141666 |
| E | 0.166666 |
| F | 0.254166 |

1. Search the Web to find the basic characteristics of publicly available or commercial software tools for association rule discovery. Document the result of your search.

There a lot software applications that can be found that have different methods to find association rules. I was able to find many of these applications on nuggets.com which has links to many different applications as well as blogs and tutorials for some of their more popular applications that they have links for. Here is a link that I found for nuggets searching for association.

<https://www.kdnuggets.com/software/associations.html>

From the link I found things like IBM's modeler as well as many others

<https://www.ibm.com/products/spss-statistics>

<https://www.lpa.co.uk/dtm.htm>

Chapter 15

1. The following is the data set X:

|  |  |  |  |
| --- | --- | --- | --- |
| X: | Year | A | B |
|  | 1996 | 7 | 100 |
|  | 1997 | 5 | 150 |
|  | 1998 | 7 | 120 |
|  | 1999 | 9 | 150 |
|  | 2000 | 5 | 130 |
|  | 2001 | 7 | 150 |

Although the following visualization techniques are not explained with enough details in the book, use your knowledge from earlier studies of statistics and other sources to create 2D Presentations:

I used R  
Show a bar chart for the variable A.

Chart, bar chart

Description automatically generated

Show a histogram for the variable B

Chart, histogram

Description automatically generated

Show a line chart for the variable B

Chart, line chart

Description automatically generated

Show a pie chart for the variable A

Chart, pie chart

Description automatically generated

Show a scatter plot for A and B variables.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Given seven-dimensional samples:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X1 | X2 | X3 | X4 | X5 | X6 | X7 |
| A | 1 | 25 | 7 | T | 1 | 5 |
| B | 3 | 27 | 3 | T | 2 | 9 |
| A | 5 | 29 | 5 | T | 1 | 7 |
| A | 2 | 21 | 9 | F | 3 | 2 |
| B | 5 | 30 | 7 | F | 1 | 7 |

A) Make a graphical representation of samples using the parallel-coordinates technique.

I used the GGally library in R

Chart

Description automatically generated

B) Are there any outliers in the given data set?

It looks like there are outliers in x3 x4, x6 and x7 based on the shape of the graph in part a. They have peaks above and below 1 and -1 which indicate that it deviates from the norm of the other data found in that column.

1. Search the we to find the basic characteristics of publicly available or commercial software tools for visualization of n-dimensional samples. Document the results of your search.

Through some research I found many useful visualizations tools such as Tableau, chartblocks, and datawraper. These are applications or adorns that can be used to visualize data easily.

<https://www.tableau.com/>

<https://www.chartblocks.com/en>

<https://www.datawrapper.de/>

There are also libraries and extensions for things like pythond, Java, and R that can be used for visualization as well. As show before I used R studio to make my charts and other libraries like D3.js for java and matplotlib for python can be used for visualization as well.

<https://d3js.org/>

<https://www.rstudio.com/>

<https://matplotlib.org/>