BITS WILP Data Mining End-Sem Exam 2017-H1 (Regular)

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2016-2017
Comprehensive Examination (EC-3 Regular)

Course No. : IS ZC415
Course Title : DATA MINING
Nature of Exam : Open Book
Weightage : 50%
Duration : 3 Hours
Date of Exam : 08/04/2017 (AN)
No. of pages: 3
No. of questions: 5
Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1(a) What is the difference between lift value of 0 vs 1?
[1]
Answer 1(a):

$$\text{lift}(A \Rightarrow 0) = \frac{P(0 \mid A)}{P(0)} = \frac{P(A \wedge 0)}{P(A)P(0)}$$
$$\text{lift}(B \Rightarrow 1) = \frac{P(1 \mid B)}{P(1)} = \frac{P(B \wedge 1)}{P(B)P(1)}$$

If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
URL: https://en.wikipedia.org/wiki/Lift_(data_mining)

Example
Assume the data set being mined is:

Antecedent	Consequent
A	0
A	0
A	1
A	0
B	1
B	0
B	1

where the antecedent is the input variable that we can control, and the consequent is the variable we are trying to predict. Real mining problems would typically have more complex antecedents, but usually focus on single-value consequents.
Most mining algorithms would determine the following rules (targeting models):

- Rule 1: A implies 0
- Rule 2: B implies 1

because these are simply the most common patterns found in the data. A simple review of the above table should make these rules obvious.
The *support* for Rule 1 is 3/7 because that is the number of items in the dataset in which the antecedent is A and the consequent 0. The support for Rule 2 is 2/7

Pages

- Postings Index
- Index of BITS WILP Exam Papers and Content
- Index of Lessons in Technology
- Index of Guest Interviews
- Downloads
- Book Requests

Blog Archive

▼ 2020 (31)
▼ May (1)
Covid-19 and response of IT companies (by Divjot S...
► April (6)
► March (12)
► February (6)
► January (6)
► 2019 (48)
► 2018 (31)
► 2017 (15)
► 2016 (6)

Popular Posts

You Are a Badass. How to stop doubting your greatness and start living an awesome life (Jen Sincero, 2013)
INTRODUCTION The language used in the book extremely funny and Jen Sincero still makes sure that she m...

Covid-19 and response of IT companies (by Divjot Singh)
As the Covid-19 pandemic ravages the world, many domains like airlines, tourism and services...

Innovation to beat the Coronavirus (Covid19)
Coronovirus' Exponential growth and decline In the first phase of the pandemic, we saw a...

Download fiction books (March 2018)
Download fiction books for free: Link for Google Dr...

Life Lessons By Steve Jobs
Steve Jobs' last words will change your views on life. The billionaire passed away at the ...

Effects of news and world events on Nifty50 and stock market
Day: 10th Aug 2017 Sensex tanks 267 points. Nifty hits one-month low. 1. Market outlook: ...


Why Bill Gates would raise chickens
I'm excited about the poverty-fighting power of poultry. If you were living on $2 a day, wh...

because two of the seven records meet the antecedent of B and the consequent of 1. The supports can be written as:

$$\text{supp}(A \Rightarrow 0) = P(A \wedge 0) = P(A)P(0 \mid A) = P(0)P(A \mid 0)$$
$$\text{supp}(B \Rightarrow 1) = P(B \wedge 1) = P(B)P(1 \mid B) = P(1)P(B \mid 1)$$

The *confidence* for Rule 1 is 3/4 because three of the four records that meet the antecedent of A meet the consequent of 0. The confidence for Rule 2 is 2/3 because two of the three records that meet the antecedent of B meet the consequent of 1. The confidences can be written as:

$$\text{conf}(A \Rightarrow 0) = P(0 \mid A)$$
$$\text{conf}(B \Rightarrow 1) = P(1 \mid B)$$

Lift can be found by dividing the confidence by the unconditional probability of the consequent, or by dividing the support by the probability of the antecedent times the probability of the consequent, so:

- The lift for Rule 1 is (3/4)/(4/7) = (3*7)/(4 * 4) = 21/16 ≈ 1.31
- The lift for Rule 2 is (2/3)/(3/7) = (2*7)/(3 * 3) = 14/9 ≈ 1.56

$$\text{lift}(A \Rightarrow 0) = \frac{P(0 \mid A)}{P(0)} = \frac{P(A \wedge 0)}{P(A)P(0)}$$
$$\text{lift}(B \Rightarrow 1) = \frac{P(1 \mid B)}{P(1)} = \frac{P(B \wedge 1)}{P(B)P(1)}$$

If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
If the lift is > 1, like it is here for Rules 1 and 2, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
Observe that even though Rule 1 has higher confidence, it has lower lift. Intuitively, it would seem that Rule 1 is more valuable because of its higher confidence—it seems more accurate (better supported). But accuracy of the rule independent of the data set can be misleading. The value of lift is that it considers both the confidence of the rule and the overall data set.

Q1(b) What do you mean by stratified sampling? Give an example.
[2]
Answer 1B:
Stratification for data mining
Published on October 4, 2007 in class distribution, cross-validation, stratification by Sandro Saitta

One common issue in data mining is the size of the data set. It is often limited. When this is the case, the test of the model is an issue. Usually, 2/3 of the data are used for training and validation and 1/3 for final testing. By chance, the training or the test set may not be representative of the overall data set. Consider for example a data set of 200 samples and 10 classes. It is likely that one of these 10 classes is not represented in the validation or test set.

To avoid this problem, you should take care of the fact that each class should be correctly represented in both the training and testing sets. This process is called stratification. One way to avoid doing stratification, regarding the training phase is to use k-fold cross-validation. Instead of having only one given validation set with a given class distribution, k different validation sets are used. However, this process doesn't guarantee a correct class distribution among the training and validation sets.

And what about the test set? The test set can only be used once, on the final model. Therefore, no method such as cross-validation can be used. There is no guarantee that the test contains all the classes that are present in the data sets. However, this situation is more likely to happen when the number of samples is small and the number of class is high. In this situation, the stratification process may be crucial. I'm wondering if people usually apply stratification or not and why. Feel free to comment on this issue regarding your personal experience.

More details about stratification can be found in the book Data Mining: Practical Machine Learning tools and techniques, by Witten and Frank (2005).

Q1(c) What do you mean by bootstrap aggregating and how is it helpful?
[2]
Answer 1C:

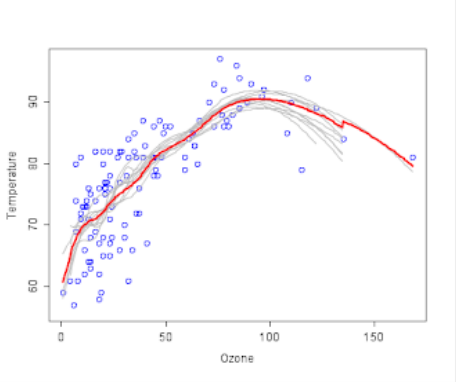
Bootstrap aggregating, also called **bagging**, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

Description of the technique

Given a standard training set D of size *n*, bagging generates *m* new training sets Di, each of size *n′*, by sampling from *D* uniformly and with replacement. By sampling with replacement, some observations may be repeated in each Di. If *n′*=*n*, then for large *n* the set Di is expected to have the fraction (1 - 1/*e*) (≈63.2%) of the unique examples of *D*, the rest being duplicates.[1] This kind of sample is known as a bootstrap sample. The *m* models are fitted using the above *m* bootstrap samples and combined by averaging the output (for regression) or voting (for classification). Bagging leads to "improvements for unstable procedures" (Breiman, 1996), which include, for example, artificial neural networks, classification and regression trees, and subset selection in linear regression (Breiman, 1994). An interesting application of bagging showing improvement in preimage learning is provided here.[2][3] On the other hand, it can mildly degrade the performance of stable methods such as K-nearest neighbors (Breiman, 1996).

Example: Ozone data

To illustrate the basic principles of bagging, below is an analysis on the relationship between ozone and temperature (data from Rousseeuw and Leroy (1986), analysis done in R).
The relationship between temperature and ozone in this data set is apparently non-linear, based on the scatter plot. To mathematically describe this relationship, LOESS smoothers (with span 0.5) are used. Instead of building a single smoother from the complete data set, 100 bootstrap samples of the data were drawn. Each sample is different from the original data set, yet resembles it in distribution and variability. For each bootstrap sample, a LOESS smoother was fit. Predictions from these 100 smoothers were then made across the range of the data. The first 10 predicted smooth fits appear as grey lines in the figure below. The lines are clearly very *wiggly* and they overfit the data - a result of the span being too low.
By taking the average of 100 smoothers, each fitted to a subset of the original data set, we arrive at one bagged predictor (red line). Clearly, the mean is more stable and there is less overfit.



Q1(d) Give an example of microaveraged vs macroaveraged Precision.
[2]
Answer 1D:
URL: http://rushdishams.blogspot.in/2011/08/micro-and-macro-average-of-precision.html

Micro- and Macro-average of Precision, Recall and F-Score

1. Micro-average Method
In Micro-average method, you sum up the individual true positives, false positives, and false negatives of the system for different sets and the apply them to get the statistics. For example, for a set of data, the system's
True positive (TP1)= 12
False positive (FP1)=9
False negative (FN1)=3
Then precision (P1) and recall (R1) will be 57.14 and 80
And for a different set of data, the system's
True positive (TP2)= 50
False positive (FP2)=23
False negative (FN2)=9

Then precision (P2) and recall (R2) will be 68.49 and 84.75
Now, the average precision and recall of the system using the Micro-average method is
Micro-average of precision = (TP1+TP2)/(TP1+TP2+FP1+FP2) = (12+50)/(12+50+9+23) = 65.96
Micro-average of recall = (TP1+TP2)/(TP1+TP2+FN1+FN2) = (12+50)/(12+50+3+9) = 83.78
The Micro-average F-Score will be simply the harmonic mean of these two figures.

2. Macro-average Method
The method is straight forward. Just take the average of the precision and recall of the system on different sets. For example, the macro-average precision and recall of the system for the given example is
Macro-average precision = (P1+P2)/2 = (57.14+68.49)/2 = 62.82
Macro-average recall = (R1+R2)/2 = (80+84.75)/2 = 82.25
The Macro-average F-Score will be simply the harmonic mean of these two figures.

Suitability
Macro-average method can be used when you want to know how the system performs overall across the sets of data. You should not come up with any specific decision with this average.
On the other hand, micro-average can be a useful measure when your dataset varies in size.

Q1(e) We generally will be more interested in association rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)? [2]
Answer 1E:
Confidence is defined as:
conf(X ⇒ Y) = supp(X∪Y) / supp(X).
For example, the rule {butter, bread} ⇒ {milk} has a confidence of 0.2 / 0.2 = 1.0 in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

URL: https://en.wikipedia.org/wiki/Association_rule_learning

Example database with 5 transactions and 5 items

transaction ID	milk	bread	Butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Let X be an itemset, X ⇒ Y an association rule and T a set of transactions of a given database.
Support
Support is an indication of how frequently the itemset appears in the dataset.
The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.
supp(X) = |{t∈T; X ⊆ t}| / |T|
In the example dataset, the itemset X = { beer, diapers } has a support of 1 / 5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions). The argument of supp( ) is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive).[3]
Confidence
Confidence is an indication of how often the rule has been found to be true.
The confidence value of a rule, X ⇒ Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.
Confidence is defined as:
conf(X ⇒ Y) = supp(X∪Y) / supp(X).
For example, the rule {butter, bread} ⇒ {milk} has a confidence of 0.2 / 0.2 = 1.0 in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).
Note that supp(X∪Y) means the support of the union of the items in X and Y. This is somewhat confusing since we normally think in terms of probabilities of events and not sets of items. We can rewrite supp(X∪Y) as the probability P(EX ∧ EY), where EX and EY are the events that a transaction contains itemset X and Y, respectively.[4]
Thus confidence can be interpreted as an estimate of the conditional probability P(EY | EX), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.[3][5]
Lift
The lift of a rule is defined as:
lift(X⇒Y) = supp(X∪Y) / supp(X) × supp(Y)
or the ratio of the observed support to that expected if X and Y were independent.
For example, the rule {milk, bread} ⇒ {butter} has a lift of 0.2 / 0.4 × 0.4 = 1.25.

If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
The value of lift is that it considers both the confidence of the rule and the overall data set.

Q1(f) Give an example where two different initial selection of centroids result in different clusters in k-means clustering. [3]
Answer 1F:
K-means Clustering

- Given *k*, the *k-means* algorithm consists of four steps:
  - — Select initial centroids at random.
  - — Assign each object to the cluster with the nearest centroid.
  - — Compute each centroid as the mean of the objects assigned to it.
  - — Repeat previous 2 steps until there is no change.

Q2(a): You are given the following data of marks obtained by 12 students in midterm exam and final exam. Given that the mean(x) = 72.167 and mean(y) = 74, derive the linear regression line equation to predict the final exam marks, given the midterm marks. [5]

Midterm exam(x)	Final exam(y)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

Anwser 2A:

Linear Regression

- Linear regression: involves a response variable y and a single predictor variable x

  $y = w_0 + w_1 x$

  where w_0 (y-intercept) and w_1 (slope) are regression coefficients
- Method of least squares: estimates the best-fitting straight line

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

- Multiple linear regression: involves more than one predictor variable
  - Training data is of the form $(\mathbf{X_1}, y_1), (\mathbf{X_2}, y_2), \ldots, (\mathbf{X_{|D|}}, y_{|D|})$
  - Ex. For 2-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
  - Solvable by extension of least square method or using SAS, S-Plus
  - Many nonlinear functions can be transformed into the above

October 10, 2017

Q2(b) Given the following pairs of credit ranking and fraud outcome in a data set, calculate the information gain (using entropy) when split by Ranking= L. Note: You only need to write the formulation; no need to solve it completely.
[4]

	1	2	3	4	5	6	7	8	9	10
Fraud	No	No	No	Yes	No	No	No	Yes	No	Yes
Ranking	L	L	M	H	H	L	L	H	M	M

Answer 2B:

Attribute Selection Measures

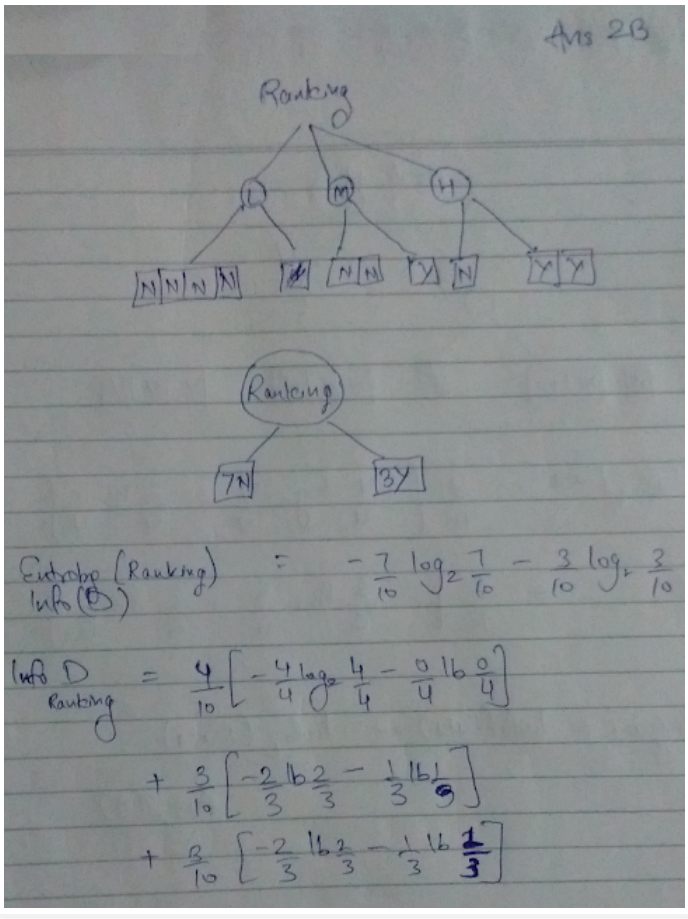
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excelle	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits}$$

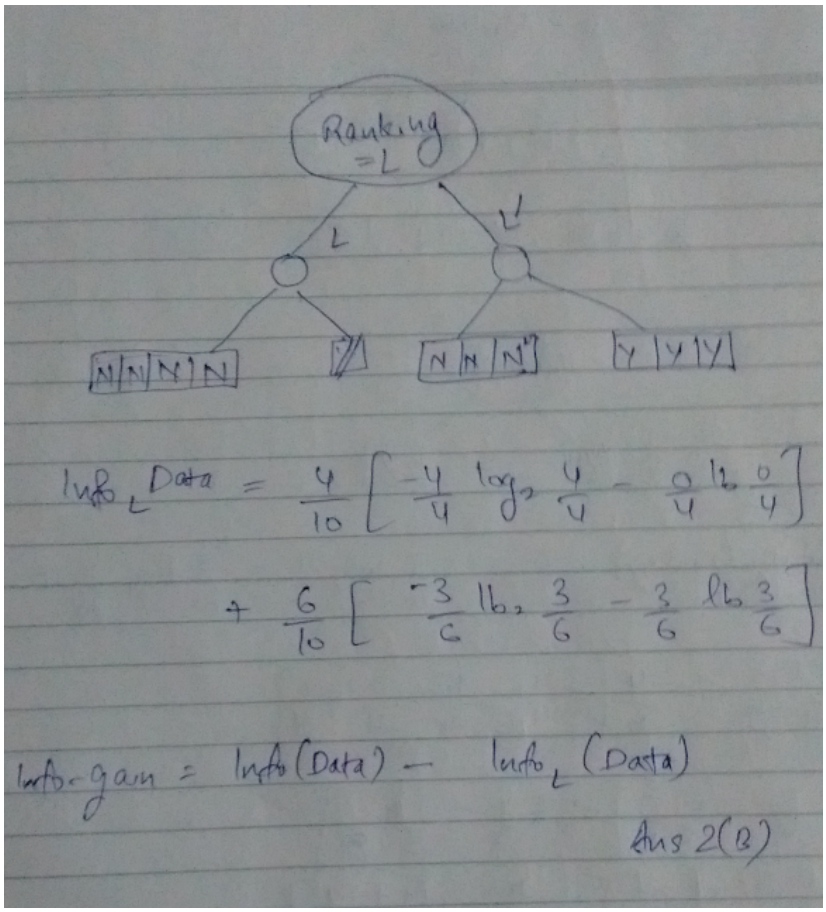
$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits}$$

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) = 0.694 \text{ bits.}$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$



...



Q3(a) The following rules identify different video files that customers watched on Youtube in the same session. Along with the rules, certain measures are also given. As a consultant to Youtube, make a recommendation on what video thumbnails to suggest/display to the user based on the association rule.

video1 → video2 with low support, high confidence, and lift = n where n is large. [2]

Answer 3A:

The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

supp(X) = |{t∈T; X ⊆ t}| / |T|

The rule has low support meaning there are few transactions in which {video1, video2} occur together.

The confidence value of a rule, X ⇒ Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

Confidence is defined as: conf(X ⇒ Y) = supp(X∪Y) / supp(X).

For example, the rule {butter, bread} ⇒ {milk} has a confidence of 0.2 / 0.2 = 1.0 in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

Confidence is high, meaning, for most transactions in which video1 was on LHS, video2 was on RHS.

The lift of a rule is defined as:

lift(X⇒Y) = supp(X∪Y) / supp(X) × supp(Y)

or the ratio of the observed support to that expected if X and Y were independent.

For example, the rule {milk, bread} ⇒ {butter} has a lift of 0.2 / 0.4 × 0.4 = 1.25.

If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

The value of lift is that it considers both the confidence of the rule and the overall data set.

Lift is high, large number 'n', so this means video1 and video2 are dependent on one another closely.

Recommendation:

If a user is watching video1, recommend video2 to him/her.

Q3(b) The following frequent itemset identifies different video files that customers watched on Youtube over two consecutive sessions. Make a recommendation on what video thumbnails to suggest/display to the user based on this.

<{video1, video2}, {video3}> with high support

[2]

Answer 3B:

The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

supp(X) = |{t∈T; X ⊆ t}| / |T|

The rule has high support meaning there are large number of transactions in which {video1, video2, video3} occur together. This means if user is watching any of these videos, recommend the other two to him.

Q3(c) The following contingency table summarizes supermarket transaction data,

	diaper	Not(diaper)	**Sum over row**
air freshener	2,000	500	2,500
Not(air freshener)	1,000	1,500	2,500
Sum over column	3,000	2,000	5,000

3C(I) Suppose that association rule "diapers => air freshener" is mined. Given that the minimum support threshold is 30% and the minimum confidence threshold is 60%, is this association rule strong? [1]

Answe 3C(I)

Given a set of transactions T, the goal of association rule mining is to find all rules having

- support ≥ *minsup* threshold
- confidence ≥ *minconf* threshold

These rules are called "strong rules". *minsup* threshold and *minconf* threshold are set by domain expert.

#(transactions with both air freshener and diapers) = 2000
#(transactions) = 5000
Support(XUY) = 2000/5000 = 0.4

The confidence value of a rule, X ⇒ Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

Confidence is defined as: conf(X ⇒ Y) = supp(X∪Y) / supp(X).

Support(air freshener) = 2500 /5000 = 0.5
Confidence (X => Y) = 0.4 / 0.5 = 0.8

3C(II) Is the purchase of diapers independent of the purchase of air freshener? Calculate the correlation that exists between the two. [1]

Answer **3C(II)**

URL: https://en.wikipedia.org/wiki/Contingency_table

The degree of association between the two variables can be assessed by a number of coefficients. The simplest, applicable only to the case of 2×2 contingency tables, is the phi coefficient defined by

$$\phi = \pm\sqrt{\frac{\chi^2}{N}},$$

where χ^2 is computed as in Pearson's chi-squared test, and N is the grand total of observations. φ varies from 0 (corresponding to no association between the variables) to 1 or −1 (complete association or complete inverse association), provided it is based on frequency data represented in 2 × 2 tables. Then its sign equals the sign of the product of the main diagonal elements of the table minus the product of the off–diagonal elements. φ takes on the minimum value −1.00 or the maximum value of 1.00 if and only if every marginal proportion is equal to .50 (and two diagonal cells are empty).[2]

URL: https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. If there were 44 men in the sample and 56 women, then

$$\chi^2 = \frac{(44-50)^2}{50} + \frac{(56-50)^2}{50} = 1.44.$$

Q4. Given the following transactions and minimum support = 50% and minimum confidence = 80%:

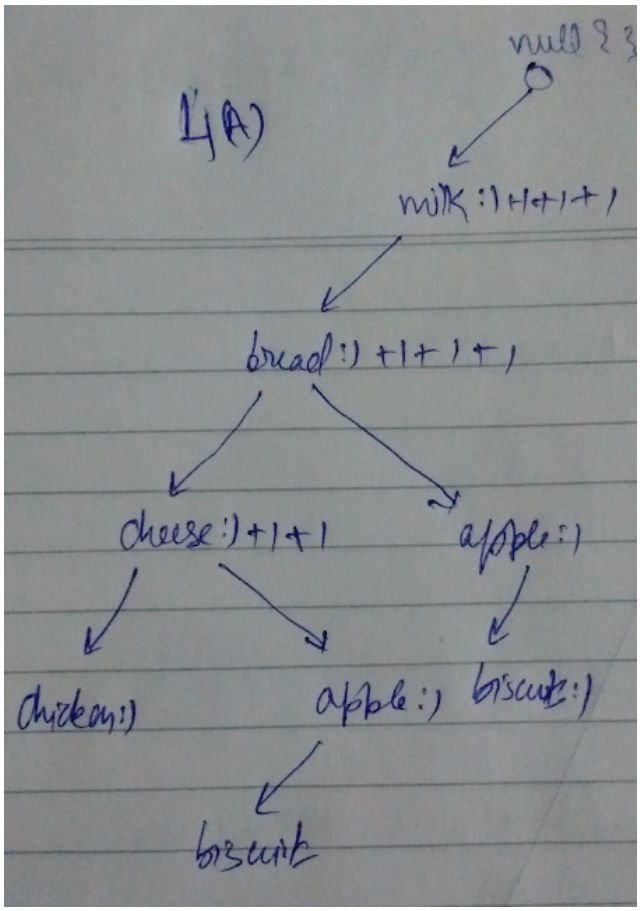
Cust Id	TID	Item_bought (in the form of brand-item_category)
01	T100	Venkys-Chicken, Amul-Milk, Nestle-Cheese, Britannia-Bread
02	T200	Britannia-Cheese, Nestle-Milk, Himalaya-Apple, Parle-Biscuit,
01	T300	Modern-Bread
03	T400	Fuji-Apple, Nestle-Milk, Modern-Bread, Parle-Biscuit
		Modern-Bread, Amul-Milk, Nestle-Cheese

a) At the granularity of *item category* (e.g., $item_i$ could be *"Milk"*), create a FP-tree of the dataset. [2]
Answer 4A:

Cust Id	TID	**Item_bought** (in the form of item_category)
01	T100	Chicken, Milk, Cheese, Bread
02	T200	Cheese, Milk, Apple, Biscuit, Bread
01	T300	Apple, Milk, Bread, Biscuit
03	T400	Bread, Milk, Cheese

Support counts => chicken: 1, milk: 4, cheese: 3, bread: 4, apple: 2, biscuit: 2
Ordered: Milk, bread, cheese, apple, biscuit, chicken
L-Order: Order the items in the transaction-table according to support-counts.
For items with equal support-counts, put them as they appear in support count order.

TID	Items
T100	Milk, bread, cheese, chicken
T200	Milk, bread, cheese, apple, biscuit
T300	Milk, bread, apple, biscuit
T400	Milk, bread, cheese



…
b) At the granularity of *brand-item category* (e.g., $item_i$ could be *"Amul-Milk"*), list all frequent itemsets using FP-growth. [5]
Support counts => Venkys-chicken: 1, amul-milk: 2, nestle-cheese: 2, Britannia-bread: 1, Britannia-cheese: 1, Nestle-milk: 2, Himalaya-apple: 1, Parle-biscuit: 2, Modern-bread: 3, Fuji-apple: 1

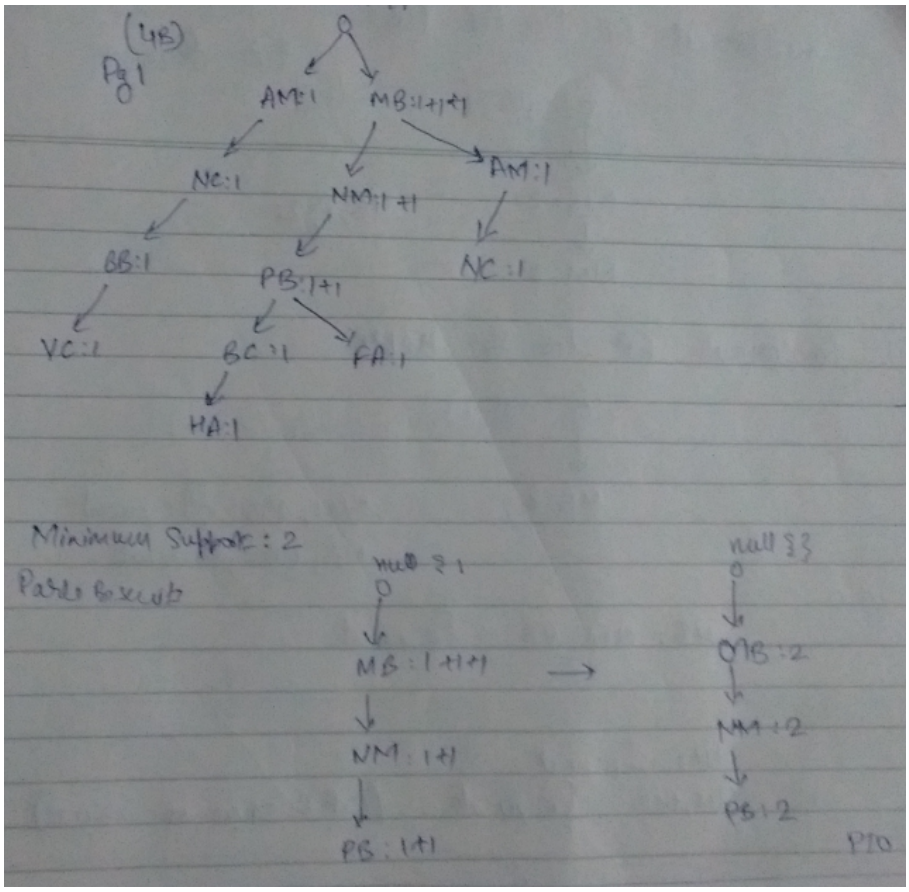
L-Order: Modern-bread: 3, Amul-milk: 2, Nestle-cheese: 2, Nestle-milk: 2, Parle-biscuit: 2, Britannia-bread: 1, Britannia-cheese: 1, Fuji-apple: 1, Himalaya-apple: 1,

Venkys-chicken: 1,

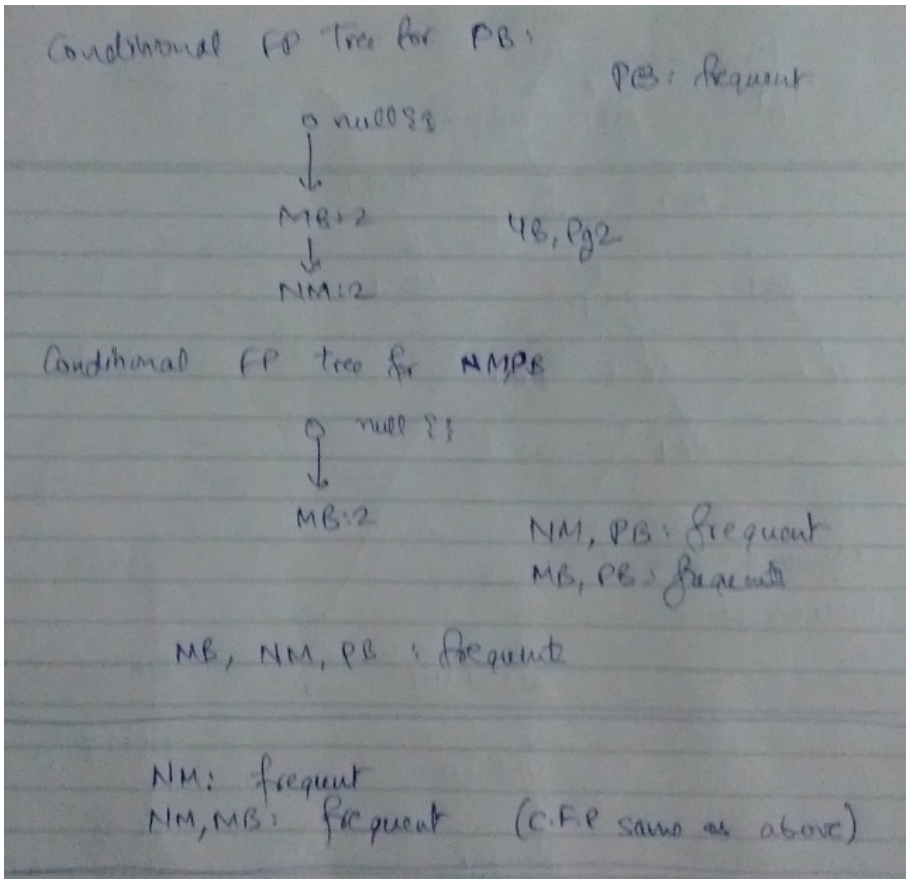
L-Order: Order the items in the transaction-table according to support-counts.
For items with equal support-counts, put them as they appear in support count order.

Cust Id	TID	Item_bought (in the form of brand-item_category)
01	T100	Amul-Milk, Nestle-Cheese, Britannia-Bread, Venkys-Chicken
02	T200	Modern-Bread, Nestle-Milk, Parle-Biscuit, Britannia-Cheese,
01	T300	Himalaya-Apple
03	T400	Modern-Bread, Nestle-Milk, Parle-Biscuit, Fuji-Apple
		Modern-Bread, Amul-Milk, Nestle-Cheese

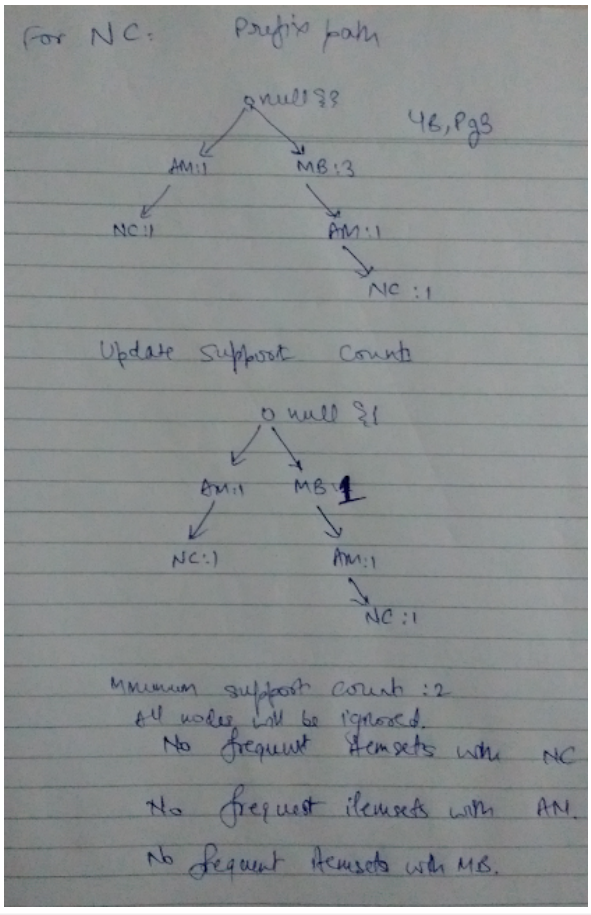
INSERT THE FP-TREE HERE



...



...



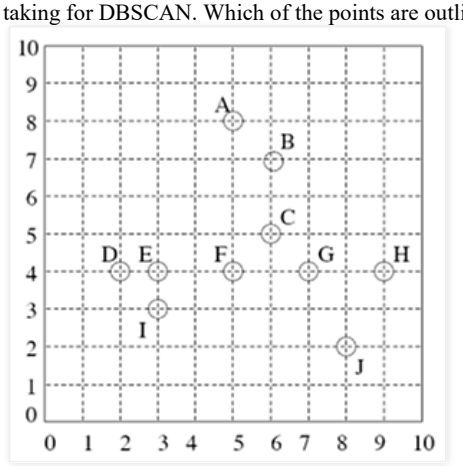
Minimum-support: 50% = 2
Possible candidates of frequent itemsets are: { Modern-bread: 3, Amul-milk: 2, Nestle-cheese: 2, Nestle-milk: 2, Parle-biscuit: 2 }
From the “Prefix path tree”, if the support-count threshold condition is met, take the item-set to be frequent.

c) List all strong association rules derived from the frequent 3-itemsets you derived in part (b) and calculate their supports, confidences and lifts. [4]
d) Give one recommendation (e.g., store layout or promotion) to store management based on the association rules and item sets you discovered. [1]

Q5(a) Given the eight points below, you want to cluster these using k-means (k=3) and Manhattan distance measure. The initial clusters are C1= {x1, x2, x3}, C2 = {x4, x5, x6}, C3 = {x7, x8}. Use k-means to continue clustering the eight points until convergence.
[5]

	A1	A2
x1	2	10
x2	2	5
x3	8	4
x4	5	8
x5	7	5
x6	6	4
x7	1	2
x8	4	9

Q5 (b) Given the 2D points in the figure below, use DBSCAN with Eps= , and MinPoints=3 (including the central point) to cluster these points using Euclidean distance measurement. List the points in each cluster. Clearly show all the steps you are taking for DBSCAN. Which of the points are outliers? [6]



No comments:

Post a Comment

Home

Subscribe to: Posts (Atom)

Followers