1/7/19

BOOK: Data mining : Concepts & Techniques
J. Han, M. Kamber

PG. NO. 1
DATE / /

Data mining - uses old data to find some (Knowledge discovery from data) pattern among them.

Data — Structured (RDBMS)
— Semi - structured (Social media sites)
— Non - structured (Unstructured) (Google drive) any file - audio, video, text can be uploaded)

3/7/19

Data Pre Processing

Collected data can be:
- Inconsistent
- Noisy
- Incomplete

can occur due to the tendency of humans to keep some information private. This can lead to missing the incomplete data in most cases.

Noisy data refers to the irrelevant data punched in. Ex- Entering 'abc' in salary field. Hence, data validation is a must.

Inconsistent data refers to the values of fields which are incoherent
Ex: Age: 23 DOB: 19/7/2018

We can say that ~~it is~~ Quality of data & Quality of mining result.

Thus, to avoid erroneous results, data must be preprocessed.

Bon Ferroni Principle
Some events are very rare even then, the event has to be considered. To increase its probability, the dataset should be enhanced.
Based on this idea, Bon Ferroni principle states that more the data, higher are the chances to detect some rare patterns.

⇒ A - 10
Apple - 20
Ap - 15
B - 21
} represent same item. Hence, they need to be transformed after integration from various sources

If transformation is not applied, the answer comes to be B, which is incorrect.

Data reduction - keeping only the relevant data which is related to the analysis to be performed but without affecting the mining results

can be achieved by:

Dimensionality reduction:

Ex- Text summarization deals with extraction & abstraction

take out the important info

read & understand the entire text then prepare a

PAGE No.
DATE / /

Thus, to achieve an optimal solution, taking a subset of the dimensions (subset of data) is required. Like- not considering attributes such as age, DOB, name, address for analysis the sales trend is appropriate and in fact required.

- Wavelet Transformation
- Data cube aggregation

⇒ PCA deals with Eigen values

$AX = \lambda X$ — data compression (part of reduction)

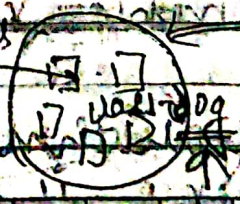
matrix ↑ ← Eigen value.

Eigen vector

Data Preprocessing
- Data cleaning / Inconsistency removal.
- Data integration
- Data Transformation
- Data reduction.

8/7/19

Data cleaning:
- Handling missing values:
  i) Ignores the tuple.
  Not ideal but suitable for problem of class label.



Here, in case of Training data, the tuple can ignored if we don't have

ii) fill data manually.
Possible for very small dataset
the missing data is filled manually
based on heuristic predictions

iii) use a global ~~value~~ constant to
fill missing value.
Ex- putting in (keeping) global constant
as NAN, unknown;
Some interesting pattern may be
lost if freq. of this constant
increases.

iv) Replace missing value by mean
(mode can also be used)

v) Use attribute mean of same class

Ex:

Good
Good
Good
Good

if some value is missing, it can be replaced by the mean of the values of same class.

(Good ⇒ from same class)

(Good → class here)

Risky
Risky
Risky

mean

Risky

vi) Replace most probable value in the tuple.
classification algorithms need to be employed and inference is to be carried out.

PAGE No.
DATE / /

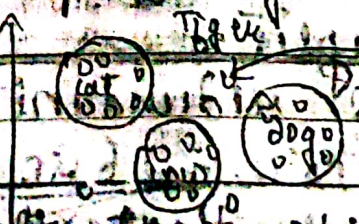
⇒ (iii) to (vi) can lead to biasing or biased data set because they involve mean

- Noisy data:
Ex - Age / Salary having -ve values
Random error in variance obtained in a variable is noise.

i) Binning
eg: 4, 8, 15, 21, 21, 24, 25, 28, 34
Sort the data.
(Already sorted)
Partition into (equi depth) bins:
- Bin 1 : 4, 8, 15
Bin 2 : 21, 21, 24
Bin 3 : 25, 28, 34
Smoothening by mean
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29
Smoothening by bin boundary

Bin 1: 4, 4, 15	If bin 1: 4, 7, 15 then, it becomes bin 1: 4, 4, 15
Bin 2: 21, 21, 24	
Bin 3: 25, 25, 34	

ii) Clustering
- grouping the similar items.
- Part of unsupervised data.
- Outlier (which mostly is considered as noise) and is simply dropped.



iii) Combine computer and human inspections.
It involves clustering as the first step and then combining the outliers. Human heuristic is used to classify them further.

iv) Regression
Selecting the best fit line which covers most of the points.
The points which are far away from this line are considered as outliers

- Data Inconsistency
Ex: DOB Age
19/7/2018 23

9/7/19 Data Integration
Problem may arise due to data inconsistency
Ex: Mr Zukerburg } refer to same
Mr Mark } person but, when combined, they may be considered as separate entity.
Hence, entity recognition req

Co-relation analysis:
If any info. is redundant, it can be easily ignored.
It may be possible that the other

value

attribute, is not derived directly / exactly But, if both the attr. have the same trend, one of them can be dropped.

Mean (Average) : Biased towards the smallest or largest data values. Might not lead to correct prediction for a individual data.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median : Gives the middle value for a data set.

Mode : Gives the data value having the highest frequency.

Mean, median, mode are all measures of central tendency.

⇒ Weighted average : $\mu = \frac{\sum_{i=1}^{n} w_i x_i}{n}$

Variance : Deviation from mean value. (rough measure) can be given by :

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n} \quad \text{or} \quad \frac{\Sigma (x - \bar{x})^2}{n}$$

Standard deviation : $\sigma_x = \sqrt{\text{variance}} = \sqrt{\sigma_x^2}$
(more exact measure)

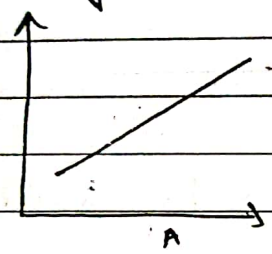
→ Co-variance :

$$\gamma_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A\,\sigma_B}$$

$\gamma_{A,B}$ → Pearson's correlation coefficient

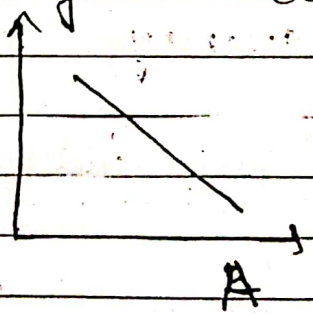
If $\gamma_{A,B} > 0$ +ve co-relation
$= 0$ Independent of each other
< 0 -ve co-relation

+ve-ly co-related : A ↑ B ↑



$\gamma_{A,B} = 1$, then they are strongly co-related

-ve-ly co-related : A ↑ B ↓



Chi-square : (Given by Pearson)

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right)$$

O : Observed value
e : Expected value

PAGE No.
DATE / /

Data Transformation

- Smoothening
- Aggregation
- Generalization (Based on granularity level the data is put into generalized labels)
- Normalization (in case of numeric data
  - MIN-MAX Normalization

Ex: Age → normalize in range [0, 1]

56
24
26

find min_A, max_A

new_max_A

new_min_A

∴ for every value, v:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

(normalized value)

⇒ min = $ 12,000
max = $ 98,000
N.R = [0.0, 1.0]

Normalize value : $ 73,600

$$v' = \frac{73600 - 12000}{98000 - 12000}(1 - 0) + 0$$

$$= \frac{61600}{86000} = 0.716$$

- Z-score normalization (0-mean)

$$v' = \frac{v - \bar{A}}{\Delta_A}$$

Helpful when min & max value is not known explicitly

Scanned by CamScanner

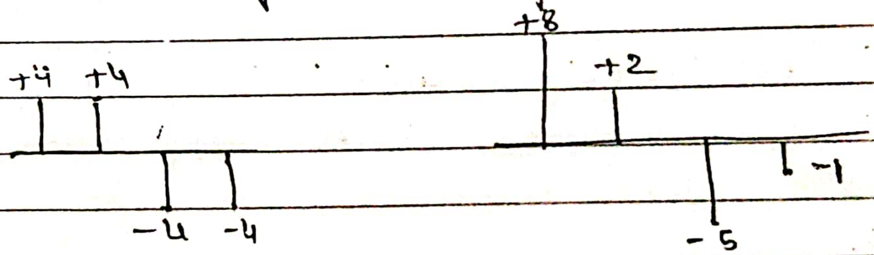
11/7/19

variance = $\frac{\Sigma(x-\bar{x})^2}{n}$ ← Entire population

S.D = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$

n ____ when we are considering a subset of the population

If we have uniform data, variance works perfectly correct. But, when the data is skewed, standard deviation gives a more exact answer, whereas variance is quite fluctuating.



variance = +4   variance = +4 [which not very accurate]

- Decimal scaling

$v' = \frac{v}{10^j}$; j is smallest Int such that ~~(max|v'|) < 1~~ (max|v'|) < 1

Ex- Range (A) : -986, 917
max. absolute value of A = 986
To normalize, $j = 3$
∴ -986 becomes -0.986

1) Data reduction
To get a (reduced) version of the original

dataset without hampering the integrity of the data.

Linear regression : $\hat{y} = wx + b$

w = weight, b = bias (original)

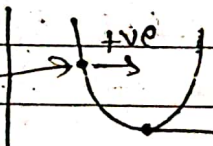
w & b are simple parameters here.
To find hyper-parameters, w & b we need to find:

$$\min\left[\frac{\sum(\hat{y}_i - y_i)^2}{n}\right]$$

so as to (almost) correctly predict the trend in x & y.
Hence, to find one parameter, we need another parameter.

$$w' = w - \alpha \frac{\partial}{\partial w} \text{(adjusted error)}$$

↑ hyper parameter / learning rate



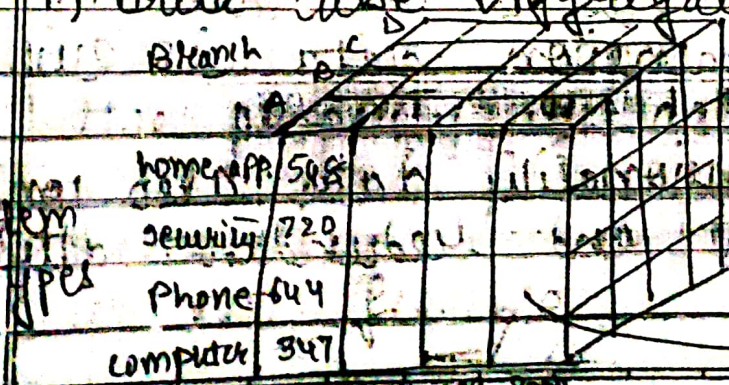
min. error.

Thus, we need to add some value to adjust the error.

15/7/19

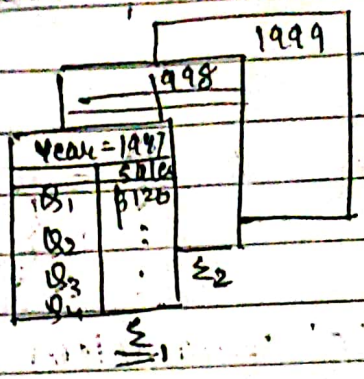
Techniques for data reduction :

i) Data Cube Aggregation :



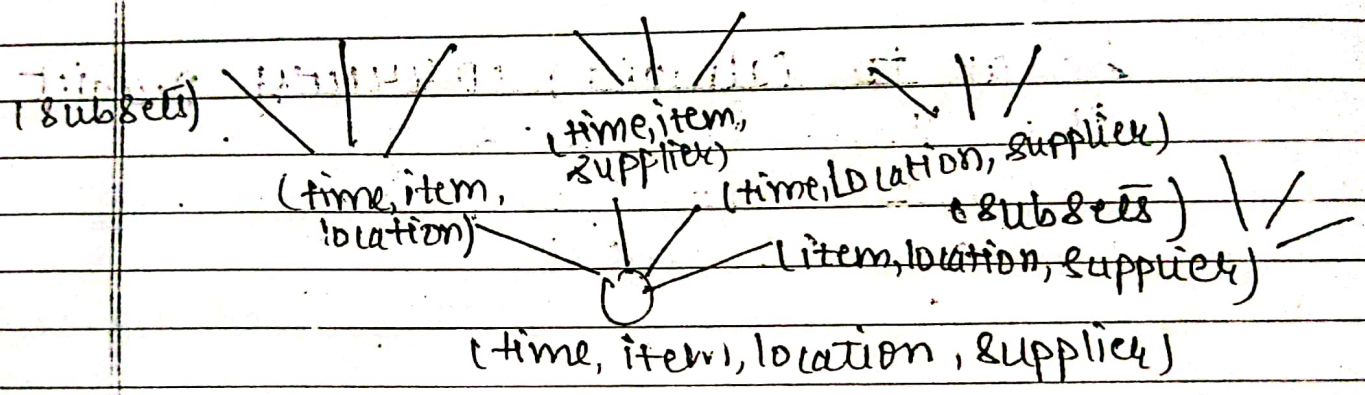
This slice represents data for branch A, ... name & year)

Compress dimensionality:



Year	Sales
1997	Σ_1
1998	Σ_2

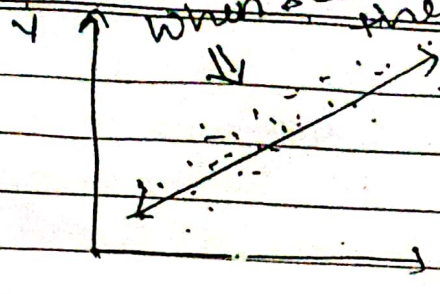
Concept hierarchy



From these various levels of abstraction, we can abstract out the info relevant to us and thus assist in ~~mining~~ mining the req. data.

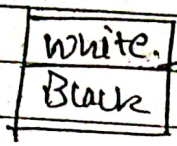
ii) Dimensionality Reduction: speed
The computational power and storage required to maintain high dimension data is quite large.

High dimensionality data also contains irrelevant and redundant attributes too.

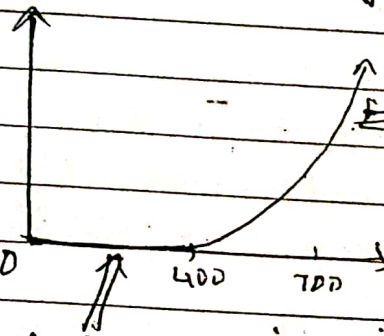


When seen from this angle the 2D-dimensional data can be viewed as one-dimensional.

⇒ The dimension having the maximum variance is the most important.



This line of separation represents the highest frequency of change in intensity (for images). Thus, displaying max. variance.



After the removal of irrelevant data, the dataset is drastically reduced and the computations speed as well as storage is enhanced.

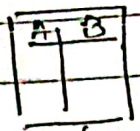
This portion has almost negligible variance and thus, removing it from the data set does not affect the quality of the image. The mining result stays unaffected.

⇒ The data distribution (in the probability space) of the original dataset and the data obtained by taking a subset of attributes should remain same. Then we can say that the required property of the data is preserved.

⇒ Exhaustive search: go on creating subsets until it is most feasible.

A, B, C.

Take subset of 2



Find co-variance:

Combine all the co-variance values to form a co-variance matrix. Finding this matrix's determinant, we can know how related two attributes are.

Ex:

$$\begin{vmatrix} 1 & 1 & 3 \\ 2 & 2 & 4 \\ 2 & 2 & 5 \end{vmatrix} = 0$$

(pair)

Same column. Thus, these two attributes are said to be strongly dependent.

iii) Step wise forward selection:

- $\{\}$ // Start with empty set
- $\{A_1\}$ // Add the attribute having nigest variance
- $\{A_1, A_4\}$ // repeat (2)

Results are checked after every addition and when the desired results are obtained, the process is stopped and we get the subset of attributes which are most relevant for our need.

iv) Step wise backward selection.

- $\{A_1, A_2 \ldots A_n\}$

- $\{A_1, A_3, A_4 \ldots A_n\}$ A_2 ✗
- $\{A_1, A_3 \ldots A_n\}$ A_4 ✗

Reverse of forward selection.

v) Combination of iii) & iv)

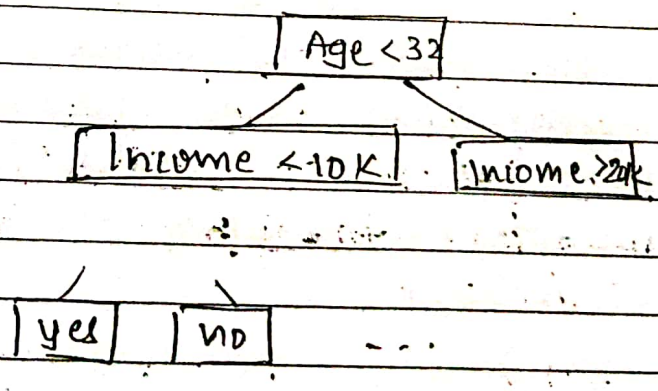
$\{A_1, A_2, A_3 \ldots A_n\}$

A_2 worst ✗ — keep removing the irrelevant attributes

best $\{A_3\}$ — keep adding the best one to another set (at every step)

vi) Decision Tree Induction.

[Wrapper Approach & Filter Approach]

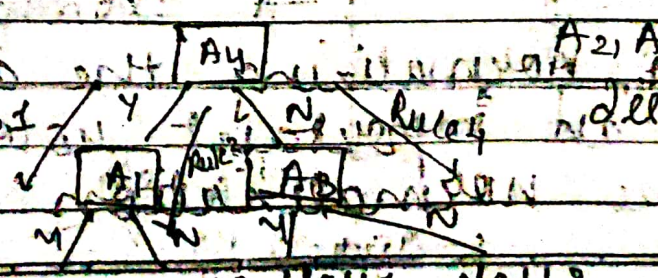


Tree is built based on the existing data and the new data is run through the tree to predict the probability of a certain event

Ex: $\{A_1, A_2, A_3 \ldots A_6\}$

↓

Populate a tree



A_2, A_3, A_6 can be discarded.

18/7/19

Data Compression:

depends on whether the uncompressed data after uncompression is the orig. data

- Lossy — Ex: PCA, Wavelet
- Loss-less — Ex: Huffman

PCA (Principle component analysis)

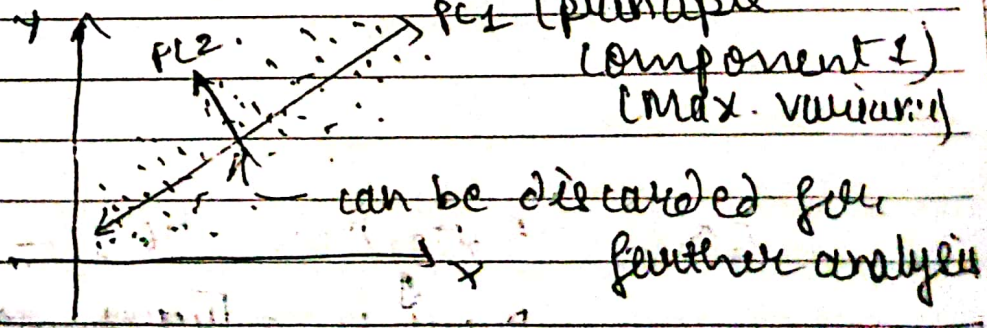
To reduce multi-dimensional data into a set with only relevant attributes

can be carried out by considering:

i) max. variance
ii) Least square approximation
iii) Matrix factorization (spectral decomposition)

- Select principle components

PCA transforms the entire data set in some other co-ordinate system. of dimensions/ attributes rather than taking a subset of the attributes



Steps:

i) For PCA, normalizing the attributes plays an important role. Mostly, Z-score normalization is applied.

ii) Construct co-variance matrix
Ex: For three attributes.

A	COV(A,A)	COV(A,B)	COV(A,C)
B	COV(B,A)	COV(B,B)	COV(B,C)
C	COV(C,A)	COV(C,B)	COV(C,C)

Where, COV(A,A) = V(A) [variance of A]
and COV(A,B) = COV(B,A)

iii) Find Eigen value & corresponding Eigen vector.

iv) Sort the Eigen values in descending order.

v) The Eigen vector corresponding to the highest Eigen value becomes the first principle component (PC1).

$\frac{\lambda_i}{\sum_{i=1} \lambda_i} \times 100$ represents the percentage that λ_i occupies as a part of whole. (importance of λ_i)

Orthonormal : Sum of all vectors should sum up to 1.

Orthogonal: A^T

V_1	V_2	V_n

If $A \cdot A^T = I$ then vectors are orthogonal (linearly independent)

PCA is used when the data is sparse (better than wavelet in this case)

new_Data = [PCA-matrix]$_{3\times5}$ * [orig-data]$_{5\times n}$

$\begin{bmatrix} -v_1- \end{bmatrix}_{1\times5}$

n → no. of observations

⇒ In this entire process, the data is modified only once [during normalization]

22/1/19

Numerosity reduction

$y = wx + b$

for any input x, we need to train w & b such that the line represented by it gives minimum outliers.

- Log-linear mode
  - useful discrete multi-dimensional probability distribution
  - Makes use of logit function [exp(wx+b)]
  - Sigmoid func. (converts values into probability)

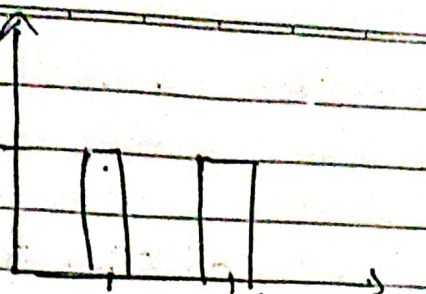
$$\Delta(x) = \frac{1}{1+e^{(-wx+b)}}$$

  - useful for sparse data
  - works correctly upto 10 dimensions

- Histogram

Values are plotted against their frequency or Range of values are plotted against their freq.

{ 2, 2, 3, 5, 2, 4, 5 }



Equi-width histogram as the bucket size uniform (range).

Equi-depth histogram (uniform freq.) the range is divided such that each range has the same freq.

* V-optimal histogram (v stands for variance) Divided based on data having least variance.

Max. difference histogram represents data ~~such~~ based on the difference between two adjacent values/range

* Clustering
  - Quality of cluster is measured in terms of diameter.
    Diameter is the max. distance between two entities in a cluster.
  - Centroid distance is the avg. distance from each obj. in a cluster to it centroid.

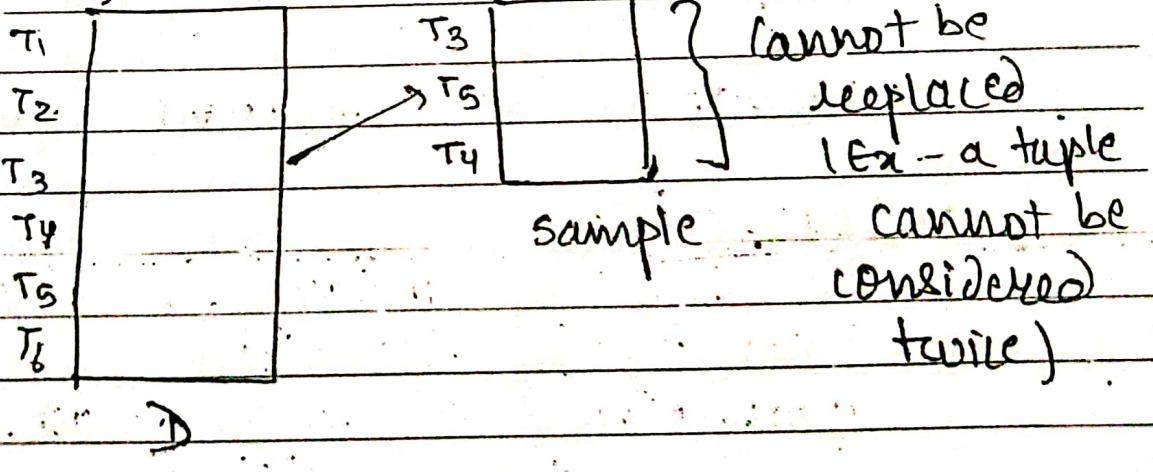
Data can be reduced by discarding values from a cluster. Ex- taking only ID-15 nearest/farthest points

from the centroid based on the requirements.

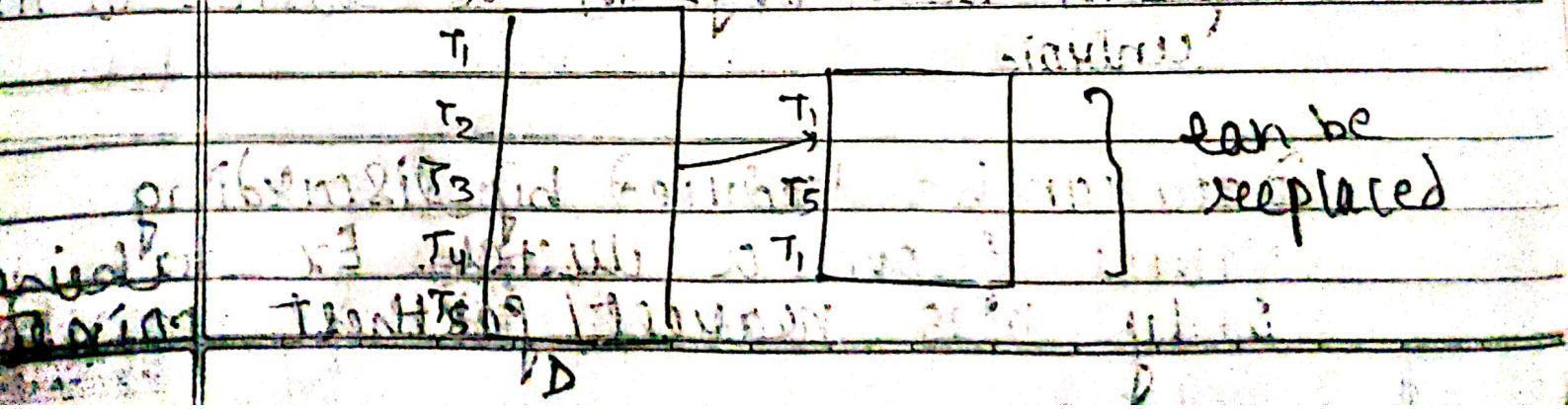
B+ tree provides abstraction from bottom to top (complete abstraction - topmost level). The data values are present in leaf nodes only. Thus, we can reduce the dataset by taking only certain keys based on the level of abstraction (A key leads only to a fixed range of data).

Sampling

- Simple Random sample without replacement (SRSWOR)



- Simple Random sample with replacement (SRSWR)



PAGE No.
DATE / /

- Cluster sample
Divide the entire dataset, D having N tuples into mutually exclusive equal size clusters.
↑ depends on requirement (not a compulsion)

- Stratified sample
It will divide the dataset into mutually exclusive strata.

Young
Young
Middle-Age
M-A
M-A
M-A
Aged
Aged

Based on probability distribution, pick a representative (sub-set of) data from every stratum.

Young
M-A
M-A
Aged

Discretization & concept hierarchy:
- Segmentation by natural partition (3-4-5 Rule)
- Concept hierarchy generation for categorical data.

⇒ Why values should be converted into a probability distribution graph?
→ When we need to take some decision based on a threshold value, we can mile the point by a few decimal

places. But, when the entire set is represented as a probability, we can predict the chances based on the overall set. (Ex- determining whether a cell is cancerous or not)

25/7/19 Discretization & concept hierarchy generation ~~by natural partition~~ for numerical data

- Binning
- Histogram Analysis
- Clustering.

- Segmentation by natural partition (3-4-5 rule)

i) If interval covers 3, 6, 7, 9 values at MSB, then divide your range into 3 partitions.

ii) If interval covers 2, 4, 8 values at MSB, then divide your range into 4 partitions.

iii) If interval covers 1, 5, 10 values at MSB, then divide your range into 5 partitions.

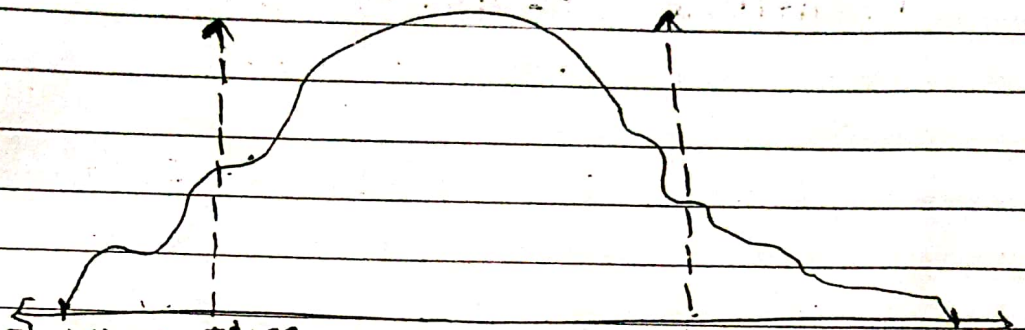
All partitions are of equal width

Ex: Starting - $351,976.00
to + $ 4,700,896.50
of the profit range

⇒ Partitions are expressed as (l, u] to avoid overlapping.

Discard 5 percentile and over 95 percentile data (clipping)

Suppose : −$159,876 to +$1838,761
(low) (high)



STEP-1 −$51,976.00 (MIN) −$159,876 (LOW) $1,838,761 (HIGH) 4,700,896.5 (MAX)

STEP-2 −169,876 —round down→ nearest 7 digit no. (−1,000,000) 'low' — (b)

1,838,761 —round up→ +2,000,000 'HIGH' (a)

$$\Rightarrow \frac{2,000,000 - (-1,000,000)}{1,000,000}$$

$$\Rightarrow 3$$

∴

(−1,000,000 , +2,000,000)

(−1,000,000, 0] (0, +1,000,000] (+1,000,000, +2,000,0

If we round up −159,876 to +1,000,00 then almost 1/3rd of the data is lost.

Since, LOW' < MIN

Partitions: Range: [-$400000, +$5,000,000]

~~[-400000, 0]~~

~~[0, +1,000,000]~~

~~[+1,000,000~~

Min. discretization at corner values.

Partitions: [-400,000, 0]

[0, +1,000,000]

[+1,000,000, +2,000,000]

[+2,000,000, +5,000,000]

5,000,000 - 2000000

1000,000

= 3

[3 partitions]

0 - (-400,000) / 100,000 = 4 [4 partitions]

1,000,000 - 0 / 1000000 = 1 [1 partitions]

2,000,000 - 1,000,000 / 1000000 = 1 [1 partitions]

Here, we see that there are more partitions in the range where max data resides.

[-400,000, -300,000]

[-300,000, -200,000]

[-200,000, -100,000]

[-100,000, 0]

Data mining primitives:

Task relevant data

Some of the tuples in the dataset can be discarded based on the task

and accordingly query is fired. This leads to correct pattern analysation

2) The kind of knowledge to be mined
We need to determine whether we need to classify, cluster the data or find an association rule

3) Background knowledge
Required for preprocessing the data and concept hierarchy generation.

4) Interestingness measure
How unique and informative a mined pattern is.
Terms: (support, confidence, simplicity, novel
Support - (Utility)

age(X, "30..39") ∧
income(X, "40K..49K")
⇒ buys(X, "VCR")
[2.2%, 60%]
↑ support ↑ confidence

} Association Rule

and

occupation(X, "student") ∧
age(X, "20..29")
⇒ buys(X, "computer")
[1.4%, 70%]

formula:
Support(A → B) = tuples containing A & B / total no. of tuples.

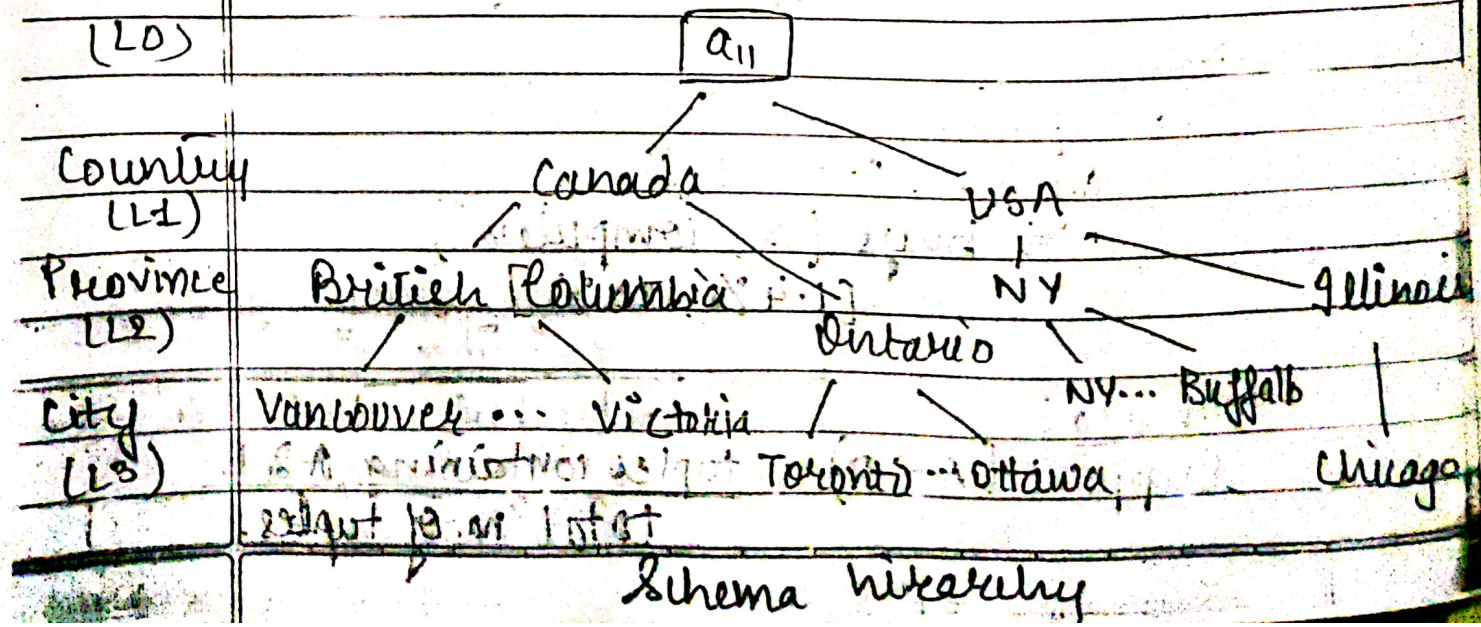
confidence $(A \rightarrow B)$ = tuples containing A & B / total A tuples
(certainty)

gives a measure of trustworthiness of a rule.

Simplicity - The rule shouldn't be too complex.

A user defines the support and confidence. A threshold is set and if the support / confidence comes out to be higher than that, the rule is considered to be an important one or the pattern is considered interesting.

Novelty - If a certain rule does not yield any new info. (or yields info. which can be mined by some other, already established rule) then it is not said to be novel.



Schema hierarchy

i) Schema hierarchy is based on total or partial ordering of the dataset.

Higher the level in the schema hierarchy lesser is its frequency (generally)

location(X, "Canada") ⇒ buys(X, "SONY-TV") → [8%, 70%]

location(X, "Montreal") ⇒ buys(X, "SONY-TV") [2%, 71%]

The second rule is redundant (in terms of confidence) as Montreal is a part of Canada and becomes a subset of the first rule. (Provides no new info. → not novel).

ii) Set-grouping

{young, middle-aged, senior} ⊂ all(age)

{20..39} ⊂ young

{40..59} ⊂ middle-aged

{60..89} ⊂ senior

iii) Operation derived hierarchy.

dmbook@cs.stu.ca

login-name < department < university < country

iv) Rule based hierarchy

low-profit margin(x) ⇐ price(x, P_1) ^ cost(x, P_2) ^ ($(P_1 - P_2 < \$50)$)

medium-profit margin(x) ⇐ price(x, P_1) ^ cost(x, P_2) ^ ($(P_1 - P_2 \geq \$50)$ ^ $(P_1 - P_2 < \$250)$)

high_profit_margin(X) ⇐ price(X, P1) ^ cost(X, P2)
^ [(P1 − P2) ≥ $250].

Representations:
- Rules
- Cross Table
- Bar charts
- Data cube
- Tables
- Pie charts
- Decision tree

[coupling 4.3, 4.4.] (Ex. 1)