RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-40

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

- LOESS CURVE
- GRAPHIC DISPLAYS OF BASIC STATISTICAL DESCRIPTIONS
- AO INDUCTION VS. LEARNING-FROM- EXAMPLE PARADIGM
- COMPARISON OF ENTIRE VS. FACTORED VERSION SPACE
- INCREMENTAL AND PARALLEL MINING OF CONCEPT DESCRIPTION
- MCQ
- REFERENCES

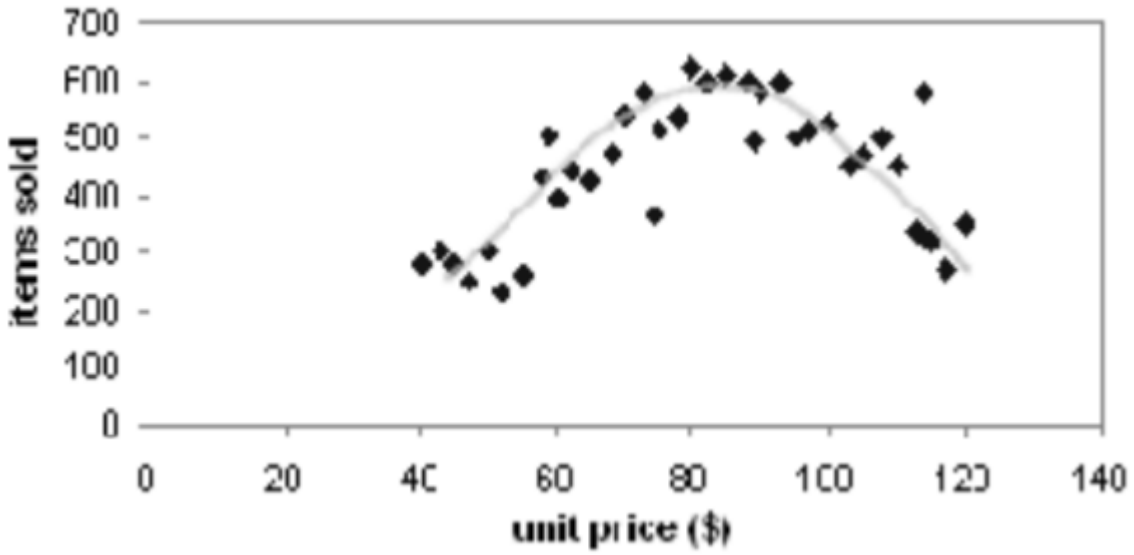
Loess Curve

Adds a smooth curve to a scatter plot in order to provide provide better perception of the pattern of dependence

• Loess curve is fitted by setting two parameters: parameters: a

smoothing parameter, and the degree of the

polynomials that are fitted by the regression



Graphic Displays of Basic Statistical Descriptions

Histogram

• Boxplot

• Quantile plot: each value xi is paired with fi indicating that approximately 100 fi % of data are □ xi

• Q il uant e-quantile (q-q) l h h) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

• Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

• Loess (local regression) curve: regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

AO Induction vs. Learning-from- example Paradigm

Difference Difference in philosophies philosophies and basic assumptions assumptions

– Positive and negative samples in learning-from-example: positive used

for generalization, negative - for specialization

– Posi it ve sampl l i d les only in data mi i n ng: hence generali i zat on-b d ase , to

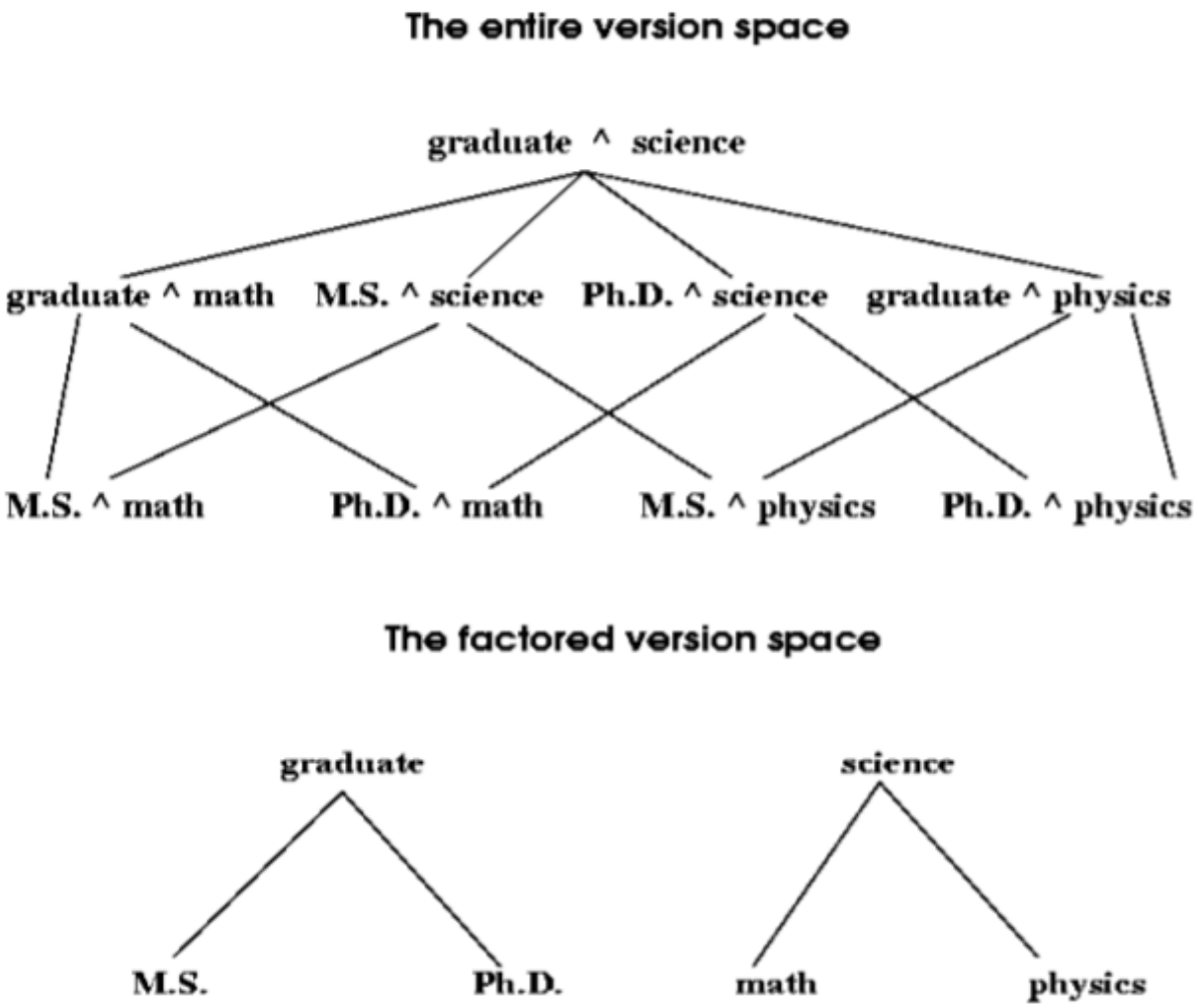
drill-down backtrack the generalization to a previous state

• Difference in methods of generalizations

– Machine learning generalizes on a tuple by tuple basis

– Data mining generalizes on an attribute by attribute basis

Comparison of Entire vs. Factored Version Space



Incremental and Parallel Mining of Concept Description

Incremental mining: revision based on newly added data □DB

– Generalize □DB to the same level of abstraction in the generalized relation R to derive □R

– Union R U □R, i.e., merge counts and other statistical information to produce a new relation R'

• Similar philosophy can be applied to data sampling, parallel and/or distributed mining, etc.

Multiple Choice Question

1. Various visualization techniques are used in _____ step of KDD.
a) selection
b) transformaion
c) data mining.
d) interpretation.

2. Extreme values that occur infrequently are called as _____.
a) outliers
b) rare values.
c) dimensionality reduction.
d) All of the above.

3. Box plot and scatter diagram techniques are _____.
a) Graphical
b) Geometric
c) Icon-based.
d) Pixel-based.

4. _____ is used to proceed from very specific knowledge to more general information.
a) Induction
b) Compression.
c) Approximation.
d) Substitution.

5. Describing some characteristics of a set of data by a general model is viewed as _____
a) Induction
b) Compression
c) Approximation
d) Summarization

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm
- https://www.geeksforgeeks.org/
- http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf DATA MINING BOOK WRITTEN BY Micheline Kamber
- https://www.javatpoint.com/three-tier-data-warehouse-architecture
- M.H. Dunham, " Data Mining: Introductory & Advanced Topics" Pearson Education
- Jiawei Han, Micheline Kamber, " Data Mining Concepts & Techniques" Elsevier
- Sam Anahory, Denniss Murray," data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, " Pearson Education
- Mallach," Data Warehousing System", TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.