

# Data Preprocessing

---

- Lecture 3: Overview of data preprocessing/  
Descriptive data summarization
- Lecture 4: Data cleaning / Data  
integration/transformation
- Lecture 5: Data reduction
- **Lecture 6: Data discretization and concept  
hierarchy generation and summary**

# Data Discretization

---

- **Three types of attributes:**

- **Nominal** — values from an unordered set, e.g., color, profession
- **Ordinal** — values from an ordered set, e.g., military or academic rank
- **Continuous** — real numbers, e.g., integers or real numbers

- **Data discretization:**

- Divide the range of a continuous attribute into intervals
- Some classification algorithms only accept categorical attributes.
- Reduce data size by discretization
- Prepare for further analysis

# Discretization and Concept Hierarchy

---

## ■ **Discretization**

- Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
- Interval labels can then be used to replace actual data values
- Supervised (use class information) vs. unsupervised
- Split (top-down) vs. merge (bottom-up)
- Discretization can be performed recursively on an attribute

## ■ **Concept hierarchy formation**

- Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

# Discretization and Concept Hierarchy Generation for Numeric Data

---

- **Typical methods: All the methods can be applied recursively**
  - **Binning**
  - **Top-down split, unsupervised**
  - **Histogram analysis**
    - **Top-down split, unsupervised**
  - **Clustering analysis**
    - **Either top-down split or bottom-up merge, unsupervised**
  - **Entropy-based discretization**: supervised, top-down split
  - **Interval merging by  $\chi^2$  Analysis**: unsupervised, bottom-up merge
  - **Segmentation by natural partitioning**: top-down split,

# Entropy-Based Discretization

---

- Given a set of samples  $S$ , if  $S$  is partitioned into two intervals  $S_1$  and  $S_2$  using boundary  $T$ , the information gain after partitioning is

$$I(S, T) = \frac{|S_1|}{|S|} \text{Entropy}(S_1) + \frac{|S_2|}{|S|} \text{Entropy}(S_2)$$

- Entropy is calculated based on class distribution of the samples in the set. Given  $m$  classes, the entropy of  $S_1$  is

$$\text{Entropy}(S_1) = - \sum_{i=1}^m p_i \log_2(p_i)$$

where  $p_i$  is the probability of class  $i$  in  $S_1$

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

# Interval Merge by $\chi^2$ Analysis

---

- Merging-based (bottom-up) vs. splitting-based methods
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- **ChiMerge** [Kerber AAI 1992, See also Liu et al. DMKD 2002]
  - Initially, each distinct value of a numerical attr. A is considered to be one interval
  - $\chi^2$  tests are performed for every pair of adjacent intervals
  - Adjacent intervals with the least  $\chi^2$  values are merged together, since low  $\chi^2$  values for a pair indicate similar class distributions
  - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

# Segmentation by Natural Partitioning

---

- A simply 3-4-5 rule can be used to segment numerical data into relatively uniform, “natural” intervals.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals ( 3 equal-width intervals for 3, 6, and 9)
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals



# Concept Hierarchy Generation for Categorical Data

---

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Acton, Canberra, ACT} < Australia
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

# Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



# Summary

---

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lots methods have been developed but data preprocessing is still an active area of research

# References

---

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Communications of ACM*, 42:73-78, 1999
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. [Mining Database Structure; Or, How to Build a Data Quality Browser](#). SIGMOD'02.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), December 1997
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, 1992
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. *Communications of ACM*, 39:86-95, 1996
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995