Data transformations

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values.
- Each old value can be identified with one of the new values
- **Smoothing:** Remove noise from data
  - Binning, regression and clustering
- **Attribute/feature construction**
  - New attributes constructed from the given ones
- **Aggregation:** Summarization, data cube construction
  - Helps to analysis data at multiple abstraction levels.

Data Transformation

- **Normalization:** Scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Discretization:** Raw values of numeric attribute are replaced by interval labels (0-10, 11-20 etc.)or conceptual labels (youth, adult,senior)
  - Labels are organized as higher level concepts resulting in concept hierarchy for numeric data.
- **Concept hierarchy generation for nominal data:** Hierarchies for nominal data are implicit within the database.

Data Transformation -Need

- Measuring unit can effect analysis
- To avoid dependence on the choice of measurements units the data should be normalized or standardized
- Allows data to fall within a smaller common range
- Data are transformed or consolidated results in efficient mining process and the patterns are understandable.
- Normalization is used in classification algorithms
  - Speeds up the learning phase

Min-Max Normalization

- **Min-max normalization**:performs a linear transformation on the original data.
- min_A and max_A are the minimum and maximum values of an attribute A
- A be the numeric attribute with n observed values v1,v2,….vn
- Min-max normalization maps a value vi to vi' in the range of [new_minA, new_maxA]

$$v' = \frac{v - \mathbf{min}\,A}{\mathbf{max}\,A - \mathbf{min}\,A}\left(\mathbf{new}_{\mathbf{max}}\,A - \mathbf{new}_{\mathbf{min}}\,A\right) + new_{min}\,A$$

- Ex. Let income range $12,000 to $98,000 normalized to [0.0, 1.0] Then $73,600 is mapped to

$$\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$$

Normalization

- **Z score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu A}{\sigma A}$$

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

- **Normalization by decimal scaling:** Normalizes by moving the decimal point of values of attribute A.

$$v' = \frac{v}{10^j}$$

- Where j is the smallest integer such that Max(|ν'|) < 1
- A range from -986 to 917 the maximum absolute value is 986
- Divide each value by 1000 so -987 normalizes to -.987 and 917 to .917

Discretization

- **Discretization:** Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Techniques based on how the discretization is performed using class information or which direction it proceeds.
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute

Data Discretization Methods

- **Binning** : Top-down split, unsupervised
- **Histogram analysis :** Top-down split, unsupervised
- **Clustering analysis :** Unsupervised, top-down split or bottom-up merge
- **Decision-tree analysis:** Supervised, top-down split
- **Correlation** (e.g., x^2) analysis :Unsupervised, bottom-up merge
- Note: All the methods can be applied recursively

Discretization by Binning

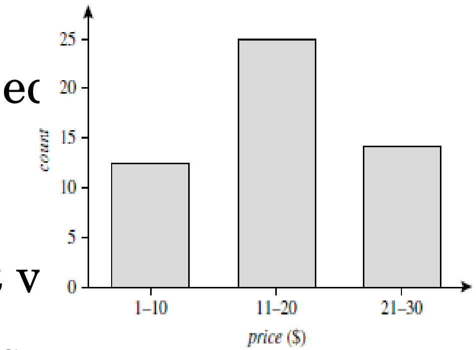
- Discretization by binning:
  - It is a top-down splitting technique based on specified number of bins.
  - Attribute values are discretized by applying equal-width or equal-frequency.
  - Replacing each bin value by the bin mean or median as in smoothening by bin means or medians.
  - Don’t use class labels so unsupervised discretization.

Simple Discretization: Histogram Analysis

- It is an unsupervised discretization technique since not using class information.
- Partitions values of an attribute A into disjoint ranges called buckets or bins.
- Various rules used to define histograms.
- It is equal-width or equal-frequency histogram

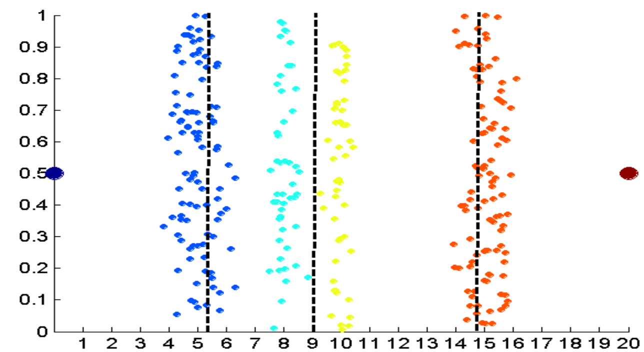
Discretization:Histogram Analysis

- Equal-width (distance) partitioning
  - Divides the range into N intervals of ec size: uniform grid
  - if A and B are the lowest and highest v of the attribute, the width of intervals W = (B –A)/N.
  - Skewed data is not handled well



An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of $10.

- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, eac containing approximately same numbe samples
  - Good data scaling and managing categorical attributes can be tricky



Clustering

- Clustering used to discretize the values of the attribute into clusters or groups
- Can discretize numeric data taking into consideration of closeness of data points.
- Generate concept hierarchy following either top-down splitting strategy or bottom-up merging strategy
- Top-down Approach: Splits clusters further forms a lower level of hierarchy
- Bottom_up: Groups up neighboring clusters in order to higher-level concepts

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Class distribution information is used in the calculation or determination of split points . Eg:Entropy determines split point (discretization point)
  - Purpose of split is resulting partition contains as many tuples as same class.
  - Top-down, recursive split

Discretization by Classification & Correlation Analysis

- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: Find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition
  - Eg: Each distinct value of the attribute considered to be one interval
    - Perform chi-squared tests on the pairs of adjacent intervals.
    - Adjacent intervals with least values are merged.
    - Since low pair indicate similar class distribution.

Concept Hierarchy Generation

- Concept hierarchy organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse.
- Mostly hierarchies are implicit within database schema and defined at schema level.
- Concept hierarchies facilitate to view data in multiple granularity
- **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 youth, adult, or senior)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data—For numeric data, use discretization methods.

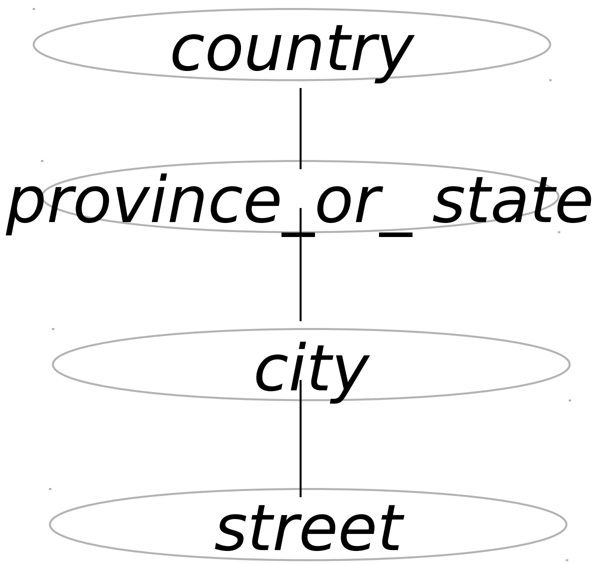
Concept Hierarchy Generation for Nominal Data

- **Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts:**
  - User or expert define concept hierarchy by specifying total or partial ordering.
  - street < city < state < country
- **Specification of a hierarchy for a set of values by explicit data grouping:**
  - Specify explicit groupings for small portion of intermediate-level data.
  - {Urbana, Champaign, Chicago} < Illinois

Concept Hierarchy Generation for Nominal Data

- **Specification of a set of attributes explicitly but not for their partial ordering.**
  - User specify set of attributes forming concept hierarchy but omit to state partial ordering.
  - System generate the attribute ordering to construct meaningful concept hierarchy
  - 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}
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - The lower the number of distinct values the higher is the generated concept.

Automatic Concept Hierarchy Generation



15 distinct values

365 distinct values

3567 distinct values

674,339 distinct values

Concept Hierarchy Generation for Nominal Data

- **Specification of only a partial set of attributes**
  - User be careless or have only vague idea in including hierarchy.
  - Eg: instead of including all hierarchical information the user may specified street and city, not others
  - To overcome embed data semantics in the database schema and pinned together with attributes.
  - The specifications one attribute may trigger a whole group of semantically tightly linked attributes to be “dragged in” to form complete hierarchy

- Use methods to normalize the data
  - 200,300,400,600,1000
  - Min-max normalization min=0 and max=1
  - Z-score
- No of transactions5000
- Transactions with hot dog=3000
- Transactions with brugers=2500
- Transaction containing both=2000
- Draw contingency table and prove the rule is strong rule or not
- Hot_dogs=>brugers