1.4.2 | Data Preprocessing Part 2

In the previous session, you have learned data cleaning and data integration tasks of data preprocessing. In this session, you will learn data reduction, data transformation, and data discretization.

At the end of the session, you should be able to:

1. Describe the various preprocessing methods.
2. Apply a data preprocessing method that is appropriate for a given data.

Task 3. Data reduction

Data reduction is applied to data to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same or almost the same results.

Why data reduction?

A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

1. Dimensionality reduction, e.g., remove unimportant attributes
   - Wavelet transforms
   - Principal Components Analysis (PCA)
   - Feature subset selection, feature creation
2. Numerosity reduction (some simply call it: Data Reduction)
   - Regression and Log-Linear Models
   - Histograms, clustering, sampling
   - Data cube aggregation

3. Data compression

Data Reduction 1: Dimensionality Reduction

Curse of dimensionality

- When dimensionality increases, data becomes increasingly sparse
- Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- The possible combinations of subspaces will grow exponentially

Dimensionality reduction

- Avoid the curse of dimensionality
- Help eliminate irrelevant features and reduce noise
- Reduce the time and space required in data mining
- Allow easier visualization

Dimensionality reduction techniques

- Wavelet transforms
- Principal Component Analysis
- Supervised and nonlinear techniques (e.g., feature selection)

Wavelet Transform

The **discrete wavelet transform** is primarily used for image compression. It decomposes a signal into different frequency subbands. Data are transformed to preserve the relative distance between objects at different levels of resolution. It allows natural clusters to become more distinguishable.

Applications of this technique will be covered in your next Data Science courses.

Principal Component Analysis (PCA)

PCA is used to find a projection that captures the largest amount of variation in data. The original data are projected onto a much smaller space, resulting in dimensionality reduction. Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space.

Steps in Principal Component Analysis

Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data. This works for numeric data only.

- Normalize input data: Each attribute falls within the same range
- Compute k orthonormal (unit) vectors, i.e., principal components
- Each input data (vector) is a linear combination of the k principal component vectors
- The principal components are sorted in order of decreasing “significance” or strength
- Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

Attribute Subset Selection

Attribute subset selection is another way to reduce the dimensionality of data. Redundant attributes. Duplicate much or all of the information contained in one or more other attributes.

```
e.g., purchase price of a product and the amount of sales tax paid
```

Irrelevant attributes contain no information that is useful for the data mining task at hand

```
e.g., students' ID is often irrelevant to the task of predicting students'
 GPA
```

Heuristic Search in Attribute Selection

There are 2^d possible attribute combinations of d attributes. The typical heuristic attribute selection methods are:

- Best single attribute under the attribute independence assumption
  - choose by significance tests
- Best step-wise feature selection
  - The best single-attribute is picked first
  - Then next best attribute condition to the first, ...
- Step-wise attribute elimination
  - Repeatedly eliminate the worst attribute
- Best combined attribute selection and elimination
- Optimal branch and bound
  - Use attribute elimination and backtracking

Attribute Creation (Feature Generation)

This is performed to create new attributes (features) that can capture the important information in a data set more effectively than the original ones. The three general methodologies include:

- Attribute extraction or domain-specific attribute extraction
- Mapping data to new space (e.g. wavelet transformation)
- Attribute construction or combining of features or data discretization

---

Data Reduction 2: Numerosity Reduction

Numerosity reduction is used to reduce data volume by choosing an alternative, smaller forms of data representation. This can be performed using the following methods:

1. **Parametric methods** (e.g., regression).
2. **Non-parametric** methods (e.g. histograms, clustering, sampling)

Parametric Data Reduction: Regression and Log-Linear Models

1. In **linear** regression, data is modelled to fit a straight line. This often uses the least-square method to fit the line.
2. **Multiple regression** allows a response variable Y to be modelled as a linear function of the multidimensional feature vector.
3. The **log-linear model** approximates discrete multidimensional probability distributions.

Regression analysis is a collective name for techniques for the modelling and analysis of numerical data consisting of values of a **dependent variable** (also called response variable or measurement) and of one or more **independent variables** (aka. explanatory variables or predictors).

The parameters are estimated so as to give a "**best fit**" of the data. Most commonly the best fit is evaluated by using the ***least-squares method***, but other criteria have also been used. Regression is used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modelling of causal relationships.

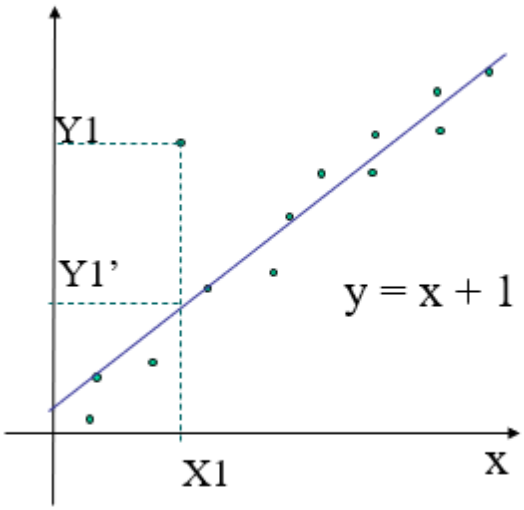

Figure 1.4.2.1 Regression

Regress Analysis and Log-Linear Models

The following are Log-linear models.

1. Linear regression: ***Y = w X + b***
   - Two regression coefficients, *w* and *b,* specify the line and are to be estimated
   - Using the least-squares criterion to the known values of *Y1, Y2, …, X1, X2, …*
2. Multiple regression: ***Y = b0 + b1 X1 + b2 X2***
   - Many nonlinear functions can be transformed into the above
3. Log-linear models
   - Approximate discrete multidimensional probability distributions
   - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
   - Useful for dimensionality reduction and data smoothing

Histogram Analysis

Histograms use binning to approximate data distributions and are a popular form of data reduction. It divides data into buckets and store average (sum) for each bucket.

Partitioning rules:

Equal-width: equal bucket range

Equal-frequency (or equal-depth)

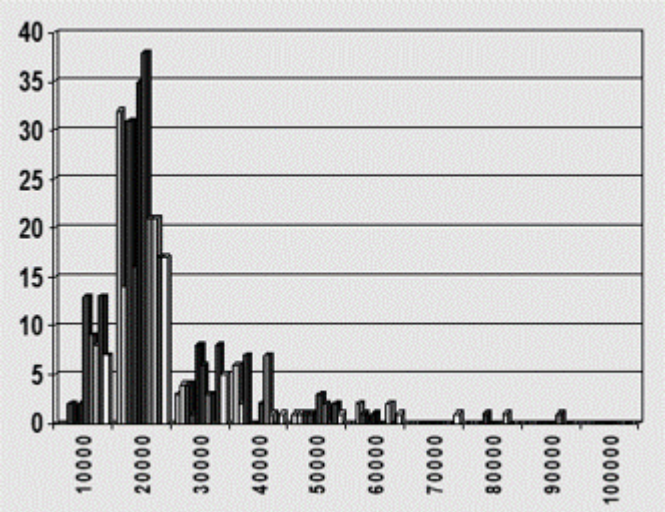

Figure 1.4.2.2 Histogram

Task 5. 1 Data Transformation

Data transformation is a function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values.

The data transformation methods are the following:

1. **Smoothing**: Remove noise from data
2. **Attribute/feature construction**
   - New attributes constructed from the given ones
3. **Aggregation**: Summarization, data cube construction
4. **Normalization**: Scaled to fall within a smaller, specified range
   - nmin-max normalization
   - nz-score normalization
   - nnormalization by decimal scaling
5. **Discretization**: Concept hierarchy climbing

Normalization

Min-max normalization: to [new_min_A, new_max_A]

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

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

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

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Figure 1.4.2.3 Normalization

Discretization

There are three general types of attributes that you will use in data mining.

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

Discretization divides the range of a continuous attribute into intervals.

- Interval labels can then be used to replace actual data values
- Reduce data size by discretization
- Supervised vs. unsupervised
- Split (top-down) vs. merge (bottom-up)
- Discretization can be performed recursively on an attribute
- Prepare for further analysis, e.g., classification

Data Discretization Methods

The following are the typical data discretization methods. All the methods can be applied recursively.

- **Binning** uses top-down split which is used for unsupervised learning.
- **Histogram analysis** uses top-down split which also used for unsupervised learning.
- **Clustering analysis** is used for unsupervised learning, uses top-down split or bottom-up merge method.

- A **decision-tree analysis** is used for supervised and uses top-down split
- **Correlation** (e.g., c2) analysis is used for unsupervised learning and uses a bottom-up method.

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
Partition into equal-frequency (**equi-depth**) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

Smoothing by **bin means**:
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Smoothing by **bin boundaries**:
- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
- Correlation analysis (e.g., Chi-merge: χ2-based discretization)

Concept Hierarchy Generation

Concept hierarchy organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse. Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularities.

Concept hierarchy formation recursively reduces 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

street < city < state < country

Specification of a hierarchy for a set of values by explicit data grouping

{Urbana, Champaign, Chicago} < Illinois

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}

Tap the button to assess your understanding of data preprocessing.

Assignment

(https://tip.instructure.com/courses/9953/assignments/156977?module_item_id=481247)
