Announcements

- Assignment 2 has been assigned
  - details on Course Website
  - Due Wed. Sept. 12th/Thurs. Sept. 13th at beginning of class
Major Tasks in Data Preprocessing

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical 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
  - Dimensionality reduction, e.g., remove unimportant attributes
  - Numerosity reduction (some simply call it: Data Reduction)
  - 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 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)

Mapping Data to a New Space

- **Fourier transform**
- **Wavelet transform**

Two Sine Waves

Two Sine Waves + Noise

Frequency
What Is Wavelet Transform?

- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression

Principal Component Analysis (PCA)

- 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.
- We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space
Attribute Subset Selection

• Another way to reduce 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
• Typical heuristic attribute selection methods:
  – 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)

• Create new attributes (features) that can capture the important information in a data set more effectively than the original ones

• Three general methodologies
  – Attribute extraction
    • Domain-specific
  – Mapping data to new space (see: data reduction)
    • E.g., Fourier transformation, wavelet transformation
    • Attribute construction
    • Combining features (see: discriminative frequent patterns in Chapter on “Advanced Classification”)
    • Data discretization

Data Reduction 2: Numerosity Reduction

• Reduce data volume by choosing alternative, smaller forms of data representation

• **Parametric methods** (e.g., regression)
  – Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  – Ex.: Log-linear models—obtain value at a point in \( m \)-D space as the product on appropriate marginal subspaces

• **Non-parametric** methods
  – Do not assume models
  – Major families: histograms, clustering, sampling, ...
Regression Analysis

• Regression analysis: A collective name for techniques for the modeling 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

• Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regression Analysis and Log-Linear Models

• Linear regression: \( Y = w X + b \)
  – Two regression coefficients, \( w \) and \( b \), specify the line and are to be estimated by using the data at hand
  – Using the least squares criterion to the known values of \( Y_1, Y_2, ..., X_1, X_2, \ldots \)

• Multiple regression: \( Y = b_0 + b_1 X_1 + b_2 X_2 \)
  – Many nonlinear functions can be transformed into the above

• 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

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)

Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
Sampling

• Sampling: obtaining a small sample $s$ to represent the whole data set $N$

• Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data

• Key principle: Choose a representative subset of the data
  – Simple random sampling may have very poor performance in the presence of skew
  – Develop adaptive sampling methods, e.g., stratified sampling

Types of Sampling

• Simple random sampling
  – There is an equal probability of selecting any particular item

• Sampling without replacement
  – Once an object is selected, it is removed from the population

• Sampling with replacement
  – A selected object is not removed from the population

• Stratified sampling:
  – Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  – Used in conjunction with skewed data
Sampling: With or without Replacement

Sampling: Cluster or Stratified Sampling
Data Reduction 3: Data Compression

Data Transformation

- 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
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction (New attributes constructed from the given ones)
  - Aggregation: Summarization
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing
Normalization

• **Min-max normalization**: to \([new_{\min}, new_{\max}]\)

\[
v' = \frac{v - min_v}{max_v - min_v} \cdot (new_{\max} - new_{\min}) + new_{\min}
\]

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

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

• **Z-score normalization** (\(\mu\): mean, \(\sigma\): standard deviation):

\[
v' = \frac{v - \mu_v}{\sigma_v}
\]

- Ex. Let \(\mu = 54,000\), \(\sigma = 16,000\). 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 \(\text{Max}(|v'|) < 1\)

Data Discretization Methods

• **Data discretization** – transforms numeric data by mapping values to interval or concept labels.

• Typical methods: All the methods can be applied recursively

  - **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., \(\chi^2\))** analysis** (unsupervised, bottom-up merge)
Simple Discretization: Binning

- **Equal-width** (distance) partitioning
  - Divides the range into \( N \) intervals of equal size: uniform grid
  - if \( A \) and \( B \) are the lowest and highest values of the attribute, the width of intervals will be: \( W = (B - A)/N \).
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

- **Equal-depth** (frequency) partitioning
  - Divides the range into \( N \) intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

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
Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse.
- 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 shown.

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
  - {Memphis, Nashville, Knoxville} < Tennessee
- 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

![Diagram of a concept hierarchy with nodes labeled as follows:
- country (15 distinct values)
- province_or_state (365 distinct values)
- city (3567 distinct values)
- street (674,339 distinct values)]

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem; Remove redundancies; Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction; Numerosity reduction; Data compression
- **Data transformation and data discretization**
  - Normalization; Concept hierarchy generation
In-Class Activity

Due by the start of next class – hand in 1 copy with all names on it.