Artificial Intelligence and Machine Learning

Data Exploration

Goal of Data Exploration

- Assemble a set of clean and relevant data
- For predictive model building specifically: designing the analytics base table
  - Each row in the ABT will represent one instance of the prediction subject—the phrase one-row-per-subject is often used to describe this structure.
  - Each row is composed of a number of attributes/features that capture the basic characteristics of an instance.
  - An attribute/feature is a property or characteristic of an instance that may vary, either from one instance to another or from one time to another.
  - One of the attributes is designated as the target feature. The rest of the attributes are descriptive features.

Data Sources

- Recorded Data
  - Operational Databases
  - Data Warehouses and Data Marts
  - Flat Files
  - Sources with Non-traditional Data Formats
    - Non-text Databases
    - Time-series Data
    - External Data Feeds
- Concepts/Prior Knowledge
- Examples

Attribute Selection

- Defining attributes/features can be difficult.
- A good way to define features is to identify the key domain concepts and then to base the features on these concepts.
- Domain concepts usually come from domain knowledge understanding.
- Attribute can be derived.
- Key considerations while designing features:
  - Data availability
  - Timing
  - Longevity
- Legal Issues
  - Anti-discrimination legislation
  - Data protection legislation

Attribute Types

- Measurement Scale: map an attribute to a numerical or symbolic value.
- Type of attributes (according to type of measurement scale)
  - Nominal/Categorical
  - Ordinal
  - Interval
  - Ratio
- According to number of values
  - Discrete
  - Continuous

Data Exploration

- Goal: conduct preliminary investigation of the data in order to better understand its specific characteristics.
- Key motivations of data exploration include
  - Understanding the data
  - Helping to select the right tool for preprocessing or analysis
  - Helping with feature selection
  - Identifying data problems
- Approaches:
  - Summary statistics
  - Visualization

Measurement of Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Identify Data Quality Issues

- A data quality issue is loosely defined as anything unusual about the data in an ABT.
- The most common data quality issues are:
  - incomplete data
  - noisy data
  - inconsistent data
- Common handling approaches:
  - Detect and clean data
    - Fill in missing values
    - Identify outliers and smooth out noisy data
    - Correct inconsistent data
  - Use ML algorithms that tolerate poor data quality

Missing Values

- Some rows may miss some attributes.
- Missing data may be caused by:
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - not entered due to misunderstanding
  - not considered important at the time of entry
  - no register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- Drop the entire attribute or example that has missing values
  - OK if the missing part is really important and can not be inferred.
  - Not OK if too many attributes/tuples get disqualified.
- Estimate missing value
  - manually
  - a default value, e.g., NULL or unknown (easy but may skew the data distribution).
  - the attribute mean (better, but still may skew data distribution) .
  - the most probable value, especially the most probable values in the same class.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

Noisy data vs outliers

- Outliers are
  - Data objects that are different
  - Attribute values that are unusual compared to the typical values for that attribute
- Outliers can be legitimate objects or values.
- Outliers may be of interest.

Handling Noisy Data

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - smooth by bin means, by bin median, or by bin boundaries, etc.
- Clamp transformation: assign upper/lower threshold values to the offending outliers
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

Other data problems

- duplicate records (especially after data integration)
- incomplete data
- inconsistent data

Data Preparation

- Data Integration
- Data Reduction
- Data Transformation

Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id vs B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundant Data in Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies

Data Reduction Strategies

- Warehouse may store terabytes of data
- Data reduction strategies
  - Data cube aggregation
  - Data sampling
  - Dimensionality reduction
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Data Aggregation

- Less is more
  - Smaller data sets require less memory and processing time
  - Providing high-level view of data
  - The behavior of groups of objects or attributes is often more stable than that of individual objects or attributes
- How to do the aggregation?
  - Quantitative attributes: sum or avg
  - Qualitative attributes: omitted or as a union set or using concept hierarchy

Data Cube Aggregation

- Data as multi-dimensional array
- Data cubes created for varying levels of abstraction are often referred to as cuboids.
- Queries regarding aggregated information should be answered using the smallest available cuboids relevant to the given task.

Data Sampling

- Select a subset of the data objects to be analyzed.
- Choose a representative subset of the data
- Sampling Approaches
  - Simple random sampling
    - May have very poor performance in the presence of skew
  - Stratified sampling
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
  - Progressive (adaptive) sampling
    - Gradually increase the sample size until sufficient

Dimensionality Reduction

- The Curse of Dimensionality
- Benefits of dimensionality reduction
  - Eliminate irrelevant features
  - Reduce noise
  - Lead to a more understandable model later
  - Allow the data to be more easily visualized
  - Reduce the amount of time and memory required by the machine learning algorithms

Dimensionality Reduction (Cont.)

- Approaches
  - Feature subset selection
  - Feature creation
    - Linear algebra techniques (especially for continuous data)
      - Principal components analysis
      - Fourier transforms
      - Wavelet transforms
    - Feature extraction
    - Feature construction

Feature Subset Selection

- Goal: select the best subset of features
- Easy task: eliminate redundant features and irrelevant features
- Brute force: try all combination
  - Advantage: reflecting the objective and bias of the learning algorithm that will eventually be used.
  - Problem: There are 2^d possible set of sub-features of d features.
- Heuristic feature selection methods

Heuristic Feature Selection Methods

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

Feature Creation

- It is sometimes possible to create, from the original attributes, a new set of attributes that captures the important information in a data set much more effectively.
- The number of new attributes can be smaller than the number of original attributes.
- Approaches:
  - Mapping the data to a new space
  - Feature extraction
  - Feature construction

Numerosity Reduction

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - 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 and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Regress Analysis and Log-Linear Models

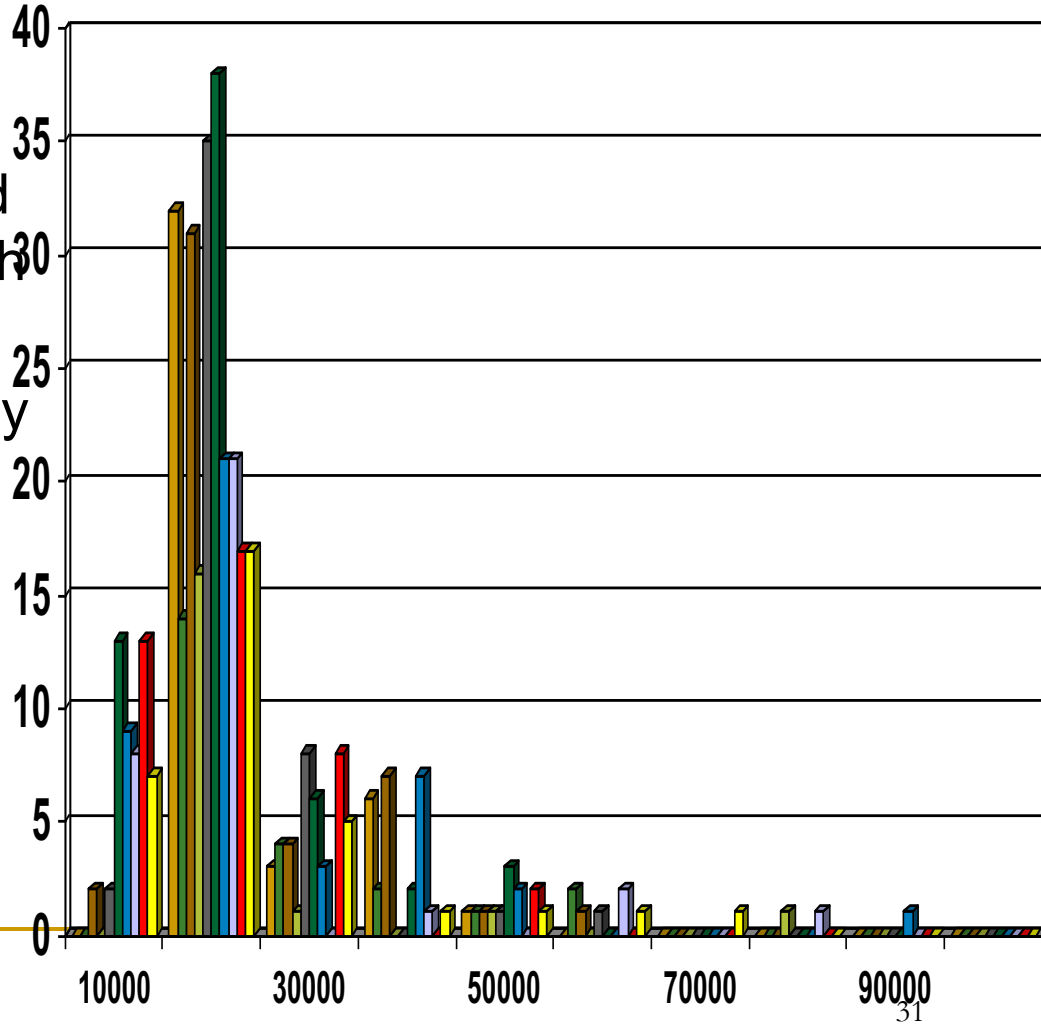
- Linear regression: $Y = \alpha + \beta X$
  - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of *Y1, Y2, …, X1, X2, ….*
- Multiple regression: $Y = b0 + b1\,X1 + b2\,X2.$
  - Many nonlinear functions can be transformed into the above.

Binning Methods

- Equal-width (distance) partitioning:
  - It 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:
  - It divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



Clustering

- Partition data set into clusters, and one can store cluster representation 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
- More clustering definitions and clustering algorithms later if time permitting

Data Transformation

- Variable (attribute) transformation
  - Nominal: one-to one mapping: v' = f(v)
  - Ordinal: order-preserving mapping, v' = f(v)
  - Interval: v' = a*v + b
  - Ratio: v' = a*v
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

Data Transformation: Normalization

- min-max normalization

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

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

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

Discretization

- 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
- Binarization

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

Discretization and concept hierarchy generation for numeric data

- Binning
- Histogram analysis
- Clustering analysis
- Entropy-based discretization
- Segmentation by natural partitioning

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- 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, e.g.,

$$Ent(S) - E(T,S) > \delta$$

- Experiments show that it may reduce data size and improve classification accuracy

Concept hierarchy generation for categorical data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
- Specification of a portion of a hierarchy by explicit data grouping
- Specification of a set of attributes, but not of their partial ordering
- Specification of only a partial set of attributes