Course Topics

- Preliminaries
- Data Understanding
- Data Preprocessing
- Clustering & Association
- Classification & Regression
- Validation & Interpretation
- Advanced Topics

Preprocessing the Data

Data Preprocessing

The process of making the data more suitable for data mining.

Data Preprocessing

The process of making the data more suitable for data mining.

The tasks employed in this process are informed by the process of data understanding.

Data Preprocessing Tasks

1 Data Cleaning

2 Data Transformation

3 Data Reduction

4 Data Discretization

Data Preprocessing Tasks

1. Data Cleaning

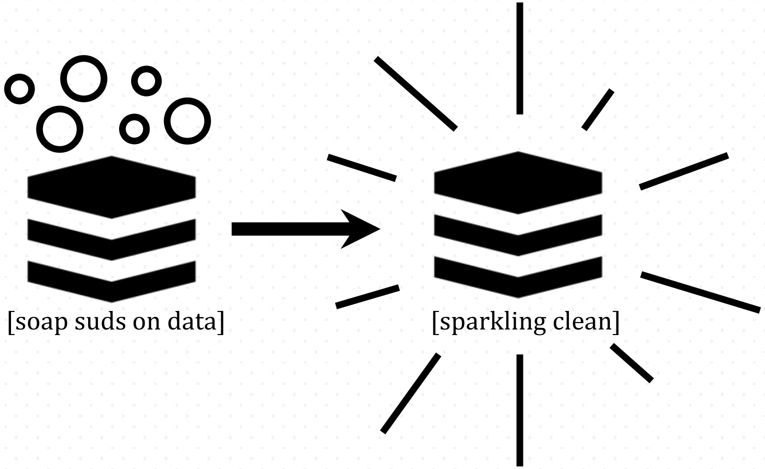
   Let’s start by looking at this task.

2. Data Transformation
3. Data Reduction
4. Data Discretization

Data Cleaning

Data cleaning involves the correction of data quality problems. These tasks include:

- Filling in missing data
- Smoothing-out noisy data
- Removing outliers and artifacts
- Correcting inconsistent data
- Removing duplicate data



Filling-in Missing Data

Ignore the instance: often not very effective, especially when few features are missing.

Fill in the missing value manually: tedious and typically infeasible.

Use a global constant to fill in the missing value: e.g., “unknown”. May be mistaken for concept.

Imputation: fill in the missing value using the feature mean or the most probable value.

Imputing Missing Data

- Delete missing observations
  - Can lead to serious biases.
  - If missing data is relatively small, may be okay.
- Cold-deck imputation
- Hot-deck imputation
- Distribution-based imputation
- Statistical imputation
- Predictive imputation

Cold-Deck Imputation

- Fill in the data using means or other analysis of the variable to fill in the value.
- Measure of central tendency (mean, median, mode)

Hot-Deck Imputation

- Identify the most similar case to the case with a missing value and substitute the most similar case’s value for the missing case’s value.
- Advantages: simplicity, maintains level of measurement, complete data at the end.
- Disadvantage: can identify more than one similar case and randomly select or use average.

Distribution-based Imputation

- Assign value based on the probability distribution of the non-missing data.
- Tries to capture the "observed" empirical distribution of data.

EM Imputation

- Expectation—Maximization (EM)
- Iterative, 2 steps:
  - E step: estimate distributions of all missing variables using a guessed parameter estimate
  - M step: using those distributions, compute a new ML estimator of the parameter
  - Repeat until convergence obtained
- Does not need to estimate actual data points.

Statistical Imputation

- Build a regressor to classify the input value
  - Consider the “missing” value as the “output” and the rest of the features as input
- Imputes the value based on other features

Predictive Imputation

- Let a classifier model the underpinnings of the missingness mechanism.

Smoothing-out Noisy Data

- Noise: Random error or variance in a measured variable.
- **Binning**: Smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of *bins*.
- **Clustering:** Detect and remove outliers.
- **Regression:** Smooth by fitting the data into regression functions.

Binning: Simple Discretization Methods

Equal-width (distance) partitioning:

- It divides the range into N intervals of equal size
- If A and B are the lowest and highest values of the attribute, the width of intervals with be: $W = (B - A)/N$.
- The most straight-forward
- But outliers may dominate presentation
- Skewed data is not handled well.

Binning: Simple Discretization Methods

Equal-depth (frequency) partitioning:

- It divides the range into *N* intervals, each containing approximately the same number of samples
- Good data scaling.
- Managing categorical features 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-depth) bins:

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

Binning Methods for Data Smoothing

Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 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

Binning Methods for Data Smoothing

Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Smoothing by bin boundaries:

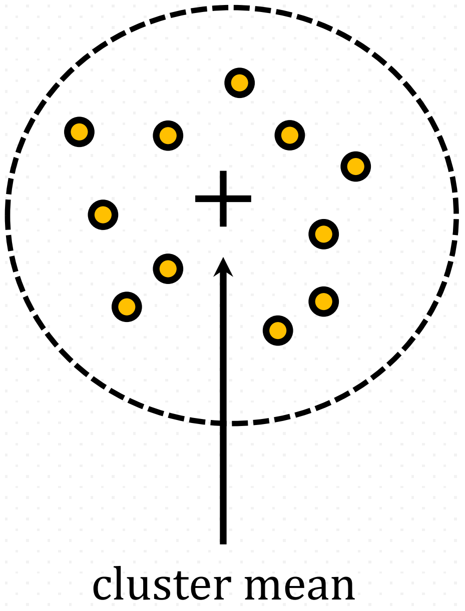
Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

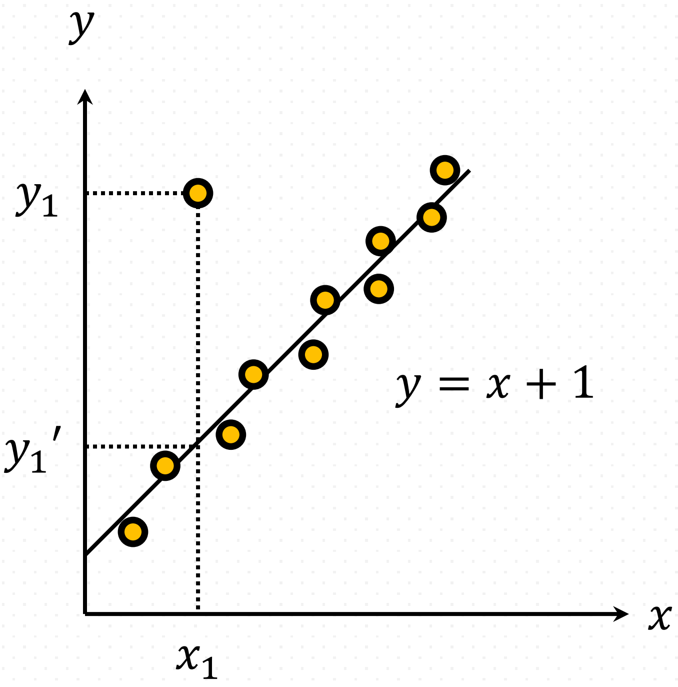
Clustering for Data Smoothing

- Cluster the data and use properties of the clusters to represent the instances constituting those clusters.



Regression for Data Smoothing

- Data can be smoothed by fitting the data to a function, such as with regression.



Removing Outliers and Artifacts

- **Proximity-based Techniques:** It is often possible to define a proximity measure between objects, with outliers being distant from most of the other data.
- **Density-based Techniques:** An outlier has a local density significantly less than that of most of its neighbors.

Correcting Inconsistent Data

- Some types of inconsistences are easy to detect.
  - e.g., a person’s height should not be negative
- In other cases, it can be necessary to consult an external source of information

Removing Duplicate Data

- Removing duplicate data raises two issues:
  1. If there are two objects that actually represent a single object, then the values of corresponding features may differ, and these inconsistent values must be resolved.
  2. Care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.
- The term deduplication is often used to refer to the process of dealing with these issues.

And Now...

Let’s clean some data!