AUBURN
UNIVERSITY

Data Preprocessing

Dr. Pei Xu
Auburn University
Thursday, August 22, 2019

Data Quality

- Missing data
- Noise and artifacts
- Outliers
- Inconsistent data
- Duplicate data

Missing Data

- Various reasons: changes in experiment, measurement not possible, human error, combining various datasets.
- Key is to know how and why data is missing.
- Missing values can have a meaning.
  - Incorporate this in the model development.
  - Example: absence of a medical test indicates a particular prognosis (value of the missing feature).

Missing Data

Three types of missing data (or "missingness"):

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

Missing Completely at Random

- Missingness does not depend on any values of any variables in the dataset.
- Missingness depends on neither the values of the observed variables, nor on those of unobserved variables.

MCAR Example

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	100,000
Customer 3	25	*Missing*
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	*Missing*

$$P(Balance\ Missing\ |Age = 25) = P(Balance\ Missing\ |Age = 60\ )$$

Missing at Random

- Missingness does not depend on the values of any of the missing or unobserved variables, but might depend on values of the observed variables.
- The pattern of missing values is identifiable.

MAR Example

Customer	Age	Account Balance
Customer 1	25	*Missing*
Customer 2	25	100,000
Customer 3	25	*Missing*
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	150,000

The account balance is primarily observed only for $Age = 60$, thus the missingness can be modeled on age.

Missing Not at Random

- Missingness depends on the values of the missing or unobserved variables.
- Pattern is non-random, non-ignorable, and typically arises due to the variable on which the data is missing.

MNAR Example

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	*Missing*
Customer 3	25	15,000
Customer 4	60	50,000
Customer 5	60	*Missing*
Customer 6	60	*Missing*

$$P(Balance\ Missing | Balance < 10{,}000\ ) = 0$$
$$P(Balance\ Missing | Balance > 10{,}000\ ) = 1$$

Noise and Artifacts

- Noise is the random component of a measurement error.
- The elimination of noise is frequently difficult.
- An important property of an algorithm is its "robustness to noise."
  - This is the stability of the algorithm on noisy data.
- Robust algorithms are often key to producing acceptable results even when noise is present.

Outliers

- Outliers are either:
  1. Data that have characteristics that are different from most of the other data.
  2. Values of a feature that are unusual with respect to the typical values for that feature.
- Unlike noise, outliers can be legitimate data or values.

Inconsistent Data

- Data can contain inconsistent values.
  - i.e., An address field with both ZIP code and city, but where the specified ZIP code area is not in the specified city.
- Some inconsistencies are easy to detect.
- Some inconsistencies may require consulting an external source.
- The correction of an inconsistency requires additional or redundant information.

Duplicate Data

- A dataset may include completely or partially duplicated data.
- Occasionally, two or more objects are identical with respect to the features measured by the database, but still represent different objects.

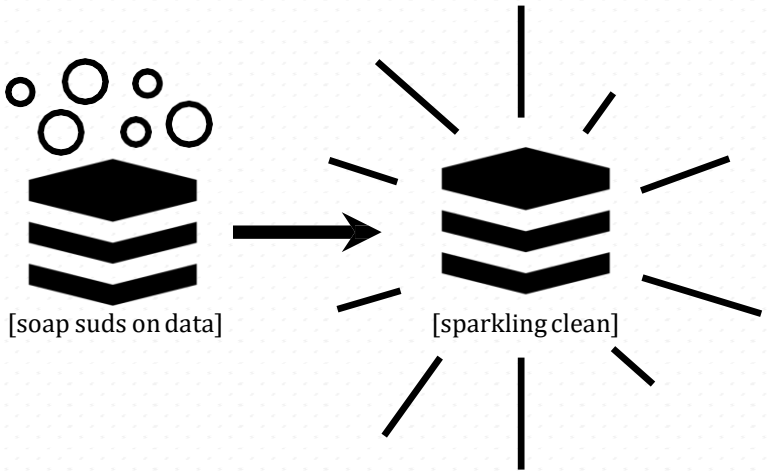
Data Preprocessing

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

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
- 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.

Predictive Imputation

- Build a regressor/classifier to estimate 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

Smoothing-out NoisyData

- 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 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:

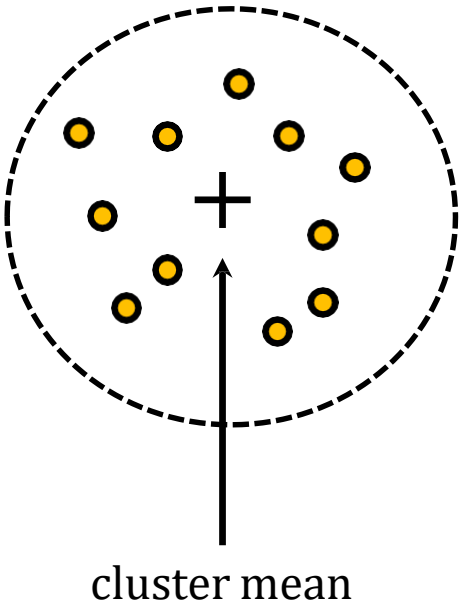
Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

Bin 3: 29, 29, 29, 29

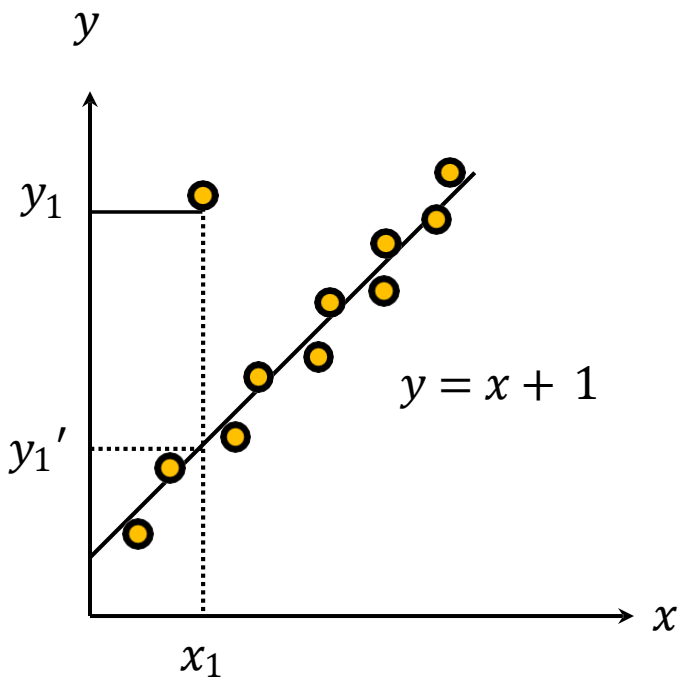
Clustering for Data Smoothing

- 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.

Reference

- James, Witten, Hastie, and Tibshirani. "*An Introduction to Statistical Learning*"
- Everaldo Aguiar, Reid Johnson, *Data Mining Slides*, University of Notre Dame
- Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin, Karpatne, Anuj. "*Introduction to Data Mining*", (Pearson, 2nd edition, 2018)
- Sanjay Ranka, *Data Mining Slides*, University of Florida.