Data management and manipulation

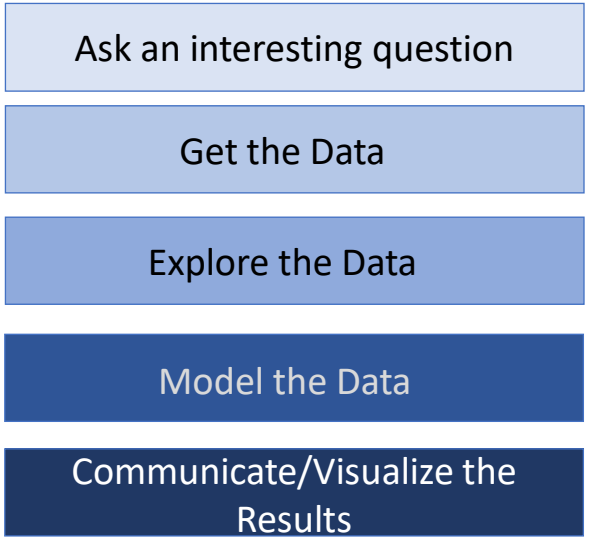
Overview

- **Data Overview**
- **Data Types**
- **Data Preprocessing**
- **Missingness in Details**
- **Imputation Methods**

Data Overview

The Data Science Process

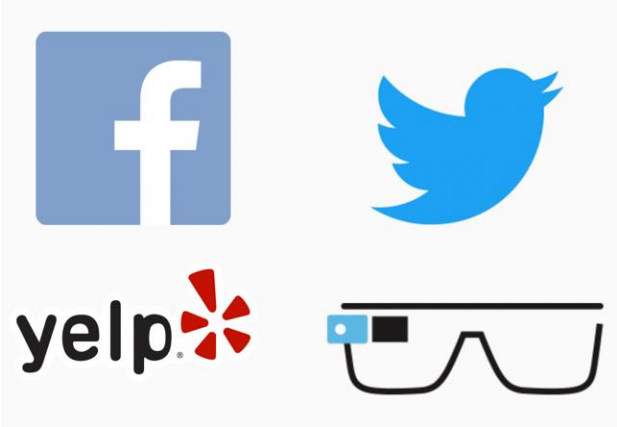
- Recall the data science process.



- Today we will begin introducing the data collection and data exploration steps.

What are data?

- “A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements.”
- Claim: everything is (can be) data!



Where do data come from?

- **Internal sources:** already collected by or is part of the overall data collection of you organization.
  For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources:** available in ready to read format from an outside source for free or for a fee.
  For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference.
- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing.
  For example: data appearing only in print form, or data on websites.

An example data set

- In many examples of codes, we use iris data.
- X has four variables.

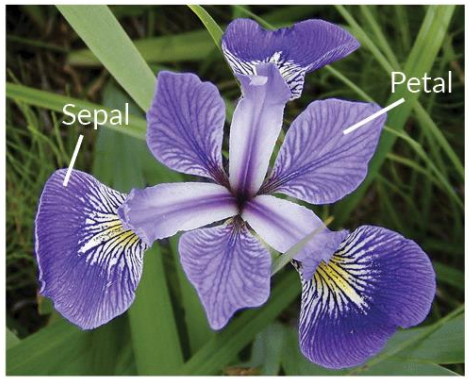
1.sepal length in cm

2.sepal width in cm

3.petal length in cm

4.petal width in cm

- Y has three outputs.
- 0 -- Iris Setosa
- 1 -- Iris Versicolour
- 2 -- Iris Virginica



Iris Versicolor


Iris Setosa


Iris Virginica

Data Types

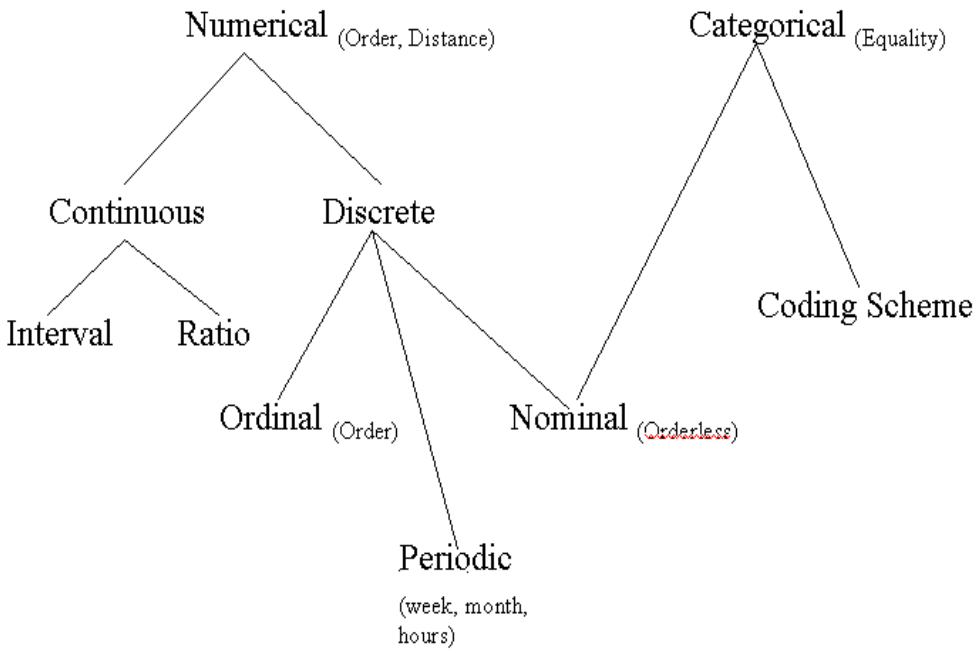
Types of data

- What kind of values are in your data (data types)?

- Simple categories:
  - **Numeric:** integers, floats
  - **Boolean:** binary or true false values
  - **Strings:** sequence of symbols
- More categories

Data Types and Forms

- Data types
  - numeric, categorical
- static, dynamic (temporal)
- Other kinds of data
  - text, Web
  - images, audio, video



Types of Data

- We'll see later that it's important to distinguish between classes of variables or attributes based on the type of values they can take on.
- **Quantitative variable:** is numerical and can be either:
  - **discrete -** a finite number of values are possible in any bounded interval. For example: "Number of siblings" is a discrete variable
  - **continuous -** an infinite number of values are possible in any bounded interval. For example: "Height" is a continuous variable
- **Categorical variable:** no inherent order among the values, for example, "What kind of pet you have" is a categorical variable

Tabular Data

- In tabular data, we expect each record or observation to represent a set of measurements of a single object or event.

Tabular Data

- In tabular data, we expect each record or observation to represent a set of measurements of a single object or event.
- Each type of measurement is called a **variable** or an **attribute** of the data. The number of attributes is called the **dimension**. These are often called **features**.
- We expect each table to contain a set of **records** or **observations** of the same kind of object or event.

Data Preprocessing

The Need of Data Preprocessing

- Data in the real world is usually dirty
  - incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

- No quality data, no quality analytics results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data analytics application (90%).

Multi-Dimensional Measure of Data Quality

- A well-accepted multi-dimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization (for numerical data)

Data Cleaning

- Importance
  - “Data cleaning is the number one problem in data warehousing”
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

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
  - etc
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

How to Handle Noisy Data?

- Remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Outlier

- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO's salary,
  - Noisy: One's age = 200, widely deviated points

Data Integration

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

Data Transformation

- Smoothing: remove noise from data
- Standardization: scaled to fall within a small, specified range
- Attribute/feature construction
  - New attributes constructed from the given ones
- Aggregation: summarization
- Generalization: concept hierarchy climbing

Data Reduction Strategies

- Data is too big to work with
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction — remove unimportant attributes
  - Aggregation and clustering – remove some data records
  - Sampling – remove some data records

Dimensionality Reduction

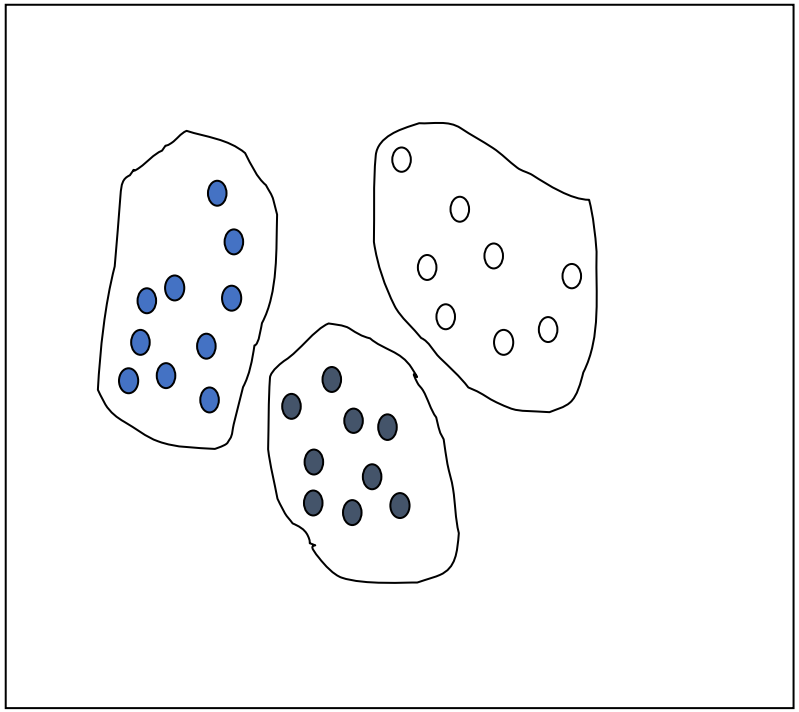
- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of attributes (features) that is sufficient for the data analytics task.
  - More contents about Feature selection in the following lectures.
- Heuristic methods:
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination etc

Clustering – reduce the number of records

- 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"
- There are many choices of clustering definitions and clustering algorithms. We will discuss them later.

Sampling – reduce the number of records

- Choose a representative subset of the data
- Random sampling
  - Simple random sampling may have poor performance in the presence of skew.



Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - divide the range of a continuous attribute into intervals because some data analytics algorithms only accept categorical attributes.
- Some techniques:
  - Binning methods – equal-width, equal-frequence

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)

Missingness in Details

Sources of Missingness

- Missing data can arise from various places in data:
- A survey was conducted and values were just randomly missed when being entered in the computer.
- A respondent chooses not to respond to a question like `Have you ever done cocaine?'.
- You decide to start collecting a new variable (like Mac vs. PC) partway through the data collection of a study.
- You want to measure the speed of meteors, and some observations are just 'too quick' to be measured properly.
- The source of missing values in data can lead to the major types of missingness:

Types of Missingness

- There are 3 major types of missingness to be concerned about:

1. **Missing Completely at Random (MCAR)** - the probability of missingness in a variable is the same for all units. Like randomly poking holes in a data set.
2. **Missing at Random (MAR)** - the probability of missingness in a variable depends only on available information (in other predictors).
3. **Missing Not at Random (MNAR)** - the probability of missingness depends on information that has not been recorded and this information also predicts the missing values.

- What are examples of each these 3 types?

Missing completely at random (MCAR)

- Missing Completely at Random is the best case scenario, and the easiest to handle:
- Examples: a coin is flipped to determine whether an entry is removed. Or when values were just randomly missed when being entered in the computer.
- Effect if you ignore: there is no effect on inferences (estimates of beta).
- How to handle: lots of options, but best to impute (more on next slide).

Missing at random (MAR)

- Missing at random is still a case that can be handled.
- Example(s): men and women respond to the question "have you ever felt harassed at work?" at different rates (and may be harassed at different rates).
- Effect if you ignore: inferences are biased (estimates of β's) and predictions are usually worsened.
- How to handle: use the information in the other predictors to build a model and **impute** a value for the missing entry.

- Key: we can fix any biases by modeling and imputing the missing values based on what is observed!

Missing Not at Random (MNAR)

- Missing Not at Random is the worst case scenario, and impossible to handle properly:
- Example(s): patients drop out of a study because they experience some really bad side effect that was not measured. Or cheaters are less likely to respond when asked if you've ever cheated.
- Effect if you ignore: there is no effect on inferences (estimates of beta) or predictions.
- How to handle: you can 'improve' things by dealing with it like it is MAR, but you [likely] may never completely fix the bias. And incorporating a **missingness indicator variable** may actually be the best approach (if it is in a predictor).

What type of missingness is present?

- Can you ever tell based on your data what type of missingness is actually present?

- Since we asked the question, the answer must be no.
- It generally cannot be determined whether data really are missing at random, or whether the missingness depends on unobserved predictors or the missing data themselves. The problem is that these potential “lurking variables” are unobserved (by definition) and so can never be completely ruled out.

- In practice, a model with as many predictors as possible is used so that the `missing at random' assumption is reasonable.

Imputation Methods

Handling Missing Data

- When encountering missing data, the approach to handling it depends on:

1. whether the missing values are in the response or in the predictors. Generally speaking, it is much easier to handle missingness in predictors.
2. whether the variable is quantitative or categorical.
3. how much missingness is present in the variable. If there is too much missingness, you may be doing more damage than good.

- Generally speaking, it is a good idea to attempt to **impute** (or 'fill in') entries for missing values in a variable (assuming your method of imputation is a good one).

Imputation Methods

- There are several different approaches to imputing missing values:

1. **Impute the mean or median** (quantitative) or most common class (categorical) for all missing values in a variable.
2. Create a new variable that is an **indicator of missingness**, and include it in any model to predict the response (also plug in zero or the mean in the actual variable).
3. **Hot deck imputation**: for each missing entry, randomly select an observed entry in the variable and plug it in.
4. **Model the imputation**: plug in predicted values ($\hat{y}$) from a model based on the other observed predictors.
5. **Model the imputation with uncertainty**: plug in predicted values plus randomness ($\hat{y} + \epsilon$) from a model based on the other observed predictors.

Imputation through modeling with uncertainty

- The schematic in the last few slides ignores the fact of imputing with uncertainty. What happens if you ignore this fact and just use the 'best' model to impute values solely on $\hat{y}$?

- The distribution of the imputed values will be too narrow and not represent real data (see next slide for illustration). The goal is to impute values that include the uncertainty of the model.

Imputation through modeling with uncertainty: an illustration

- Recall the probabilistic model in linear regression:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$$

- where $\epsilon \sim N(0, \sigma^2)$. How can we take advantage of this model
- to impute with uncertainty?
- It's a 3 step process:

1. Fit a model to predict the predictor variable with missingness from all the other predictors.
2. Predict the missing values from the model in the previous part.
3. Add in a measure of uncertainty to this prediction by randomly sampling from a $N(0, \sigma^2)$ distribution, where σ^2 is the mean square error (MSE) from the model.

Summary

- Data preparation is a big issue for data analytics
  - It is also very time consuming.
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- Many methods have been proposed but still an active area of research.
- In applications, you can also invent your own methods.