Introduction to Data Science

CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

Practicums

https://usu.instructure.com/courses/431417/discussion_topics/1385921

Practicum Presenters and Dates

- Sept 15:
  - Tableau public (Sirisha Rani)
  - AWS (Anuj Khasgiwala)
  - Scikit-learn (Vishal)
- Sept 22:
- ?:

Project Overview

https://usu.instructure.com/courses/431417/discussion_topics/1387815

Project Details

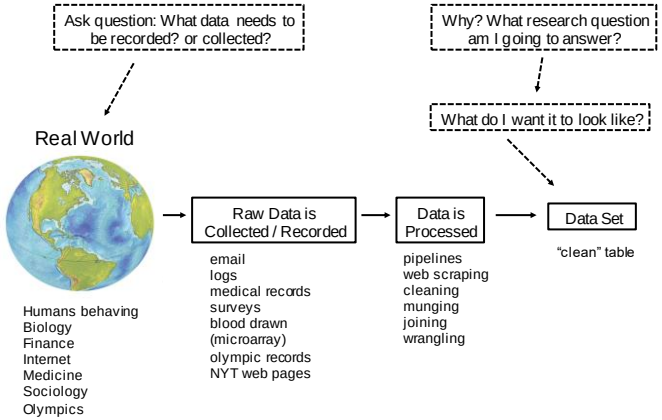
- 2 or 3-person team
- Code must live on github
- Must use non-trivial data
- Must apply some mining or analytics algorithm
- Must use the "data science loop"
- Final output:
  - Data project or data visualization
  - Ideally, a "rich" or "interactive" final product; a single figure is insufficient

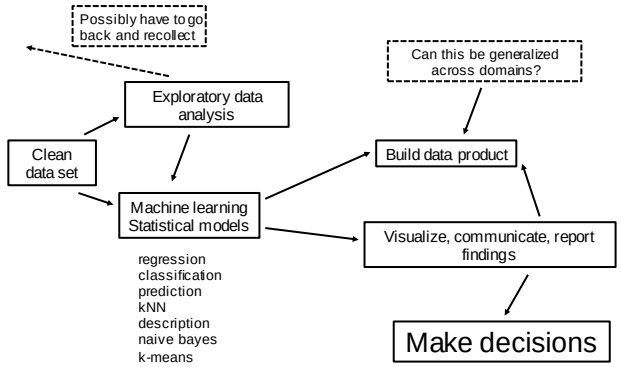
Project Phases

- [25%] Proposal: October 20
- [25%] Check Point: November 17
- [50%] Project Workshop: Dec 6 and 8 in-class

Previous Class…

Data Science Process (Loop)





Back to Basics:

Getting to Know Our Data

Data models vs. Conceptual models

- Data models are low-level descriptions of the data
  - Math: Sets with operations on them
  - Example: integers with + and x operators
- Conceptual models are mental constructions
  - Include semantics and support reasoning
- Examples (data vs. conceptual)
  - (1D floats) vs. temperatures
  - (3D vector of floats) vs. space

Data Type Taxonomy

- 1D (sets and sequences)
- Temporal
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Trees (hierarchies)
- Networks (graphs)

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

Types of Datasets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Example: Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example: Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Example: Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

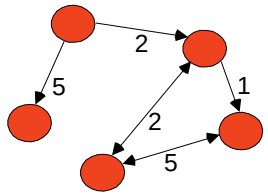
Example: Transaction Data

- A special type of record data, where each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

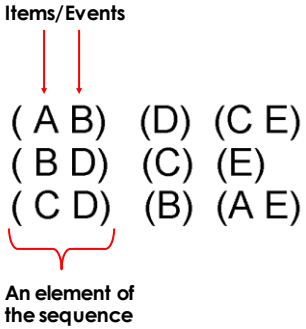
Example: Graph Data

- Examples: Generic graph and HTML Links



Example: Ordered Data

- Sequences of transactions



Example: Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
  - Examples:
    - sales database: customers, store items, sales
    - medical database: patients, treatments
    - university database: students, professors, courses
  - Also called samples, examples, instances, data points, objects, tuples.
- Data objects are described by **attributes**.
- In database… rows → data objects; columns→ attributes.

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attributes

- **Attribute** (or dimension, feature, variable): a data field, representing a characteristic or feature of a data object.
  - E.g., customer_ID, name, address
- Types:
  - Nominal / Binary (a part of Nominal)
  - Ordinal
  - Quantitative (Numeric)
    - Interval-scaled
    - Ratio-scaled

Attributes: Nominal, Ordinal, and Quantitative

- **Nominal** (categories, states, labels :: "names of things")
  - Ex. Fruits: Apples, oranges, …
  - Special case of Nominal: **Binary**
- **Ordinal** (Ordered)
  - Values have a meaningful order (rank), but magnitude between successive values is unknown
  - Quality of meat: Grade A, AA, AAA
- **(Q) Interval** (No true zero-point)
  - Calendar dates: Jan 24, 2012; Location (Lat/Long)
  - Only differences (intervals) may be compared
- **(Q) Ratio** (Inherent zero-point)
  - Physical measurements: Length, Mass, …
  - Counts and amounts

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countable set of values
    - E.g., zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as floating-point variables.

From data model to N, O, Q data type

- Data Model
  - 32.5, 54.0, -17.3, …
  - floats
- Conceptual Model
  - Temperature (°C)
- Data/Attribute Type
  - Burned vs. Not-Burned (N)
  - Hot, Warm, Cold (O)
  - Temperature Value (Q)

Quiz! Census Data

- **People**: # of people in group
- **Year**: 1850 – 2000 (every decade)
- **Age**: 0 – 90+
- **Sex**: Male, Female
- **Marital Status**: Single, Married, Divorced

Quiz! Census Data

- **People**
- **Year**
- **Age**
- **Sex**
- **Marital Status**
- 2,348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236

Census: N, O, Q?

- **People** — **Q-Ratio**
- **Year** — **Q-Interval (O)**
- **Age** — **Q-Ratio (O)**
- **Sex (M/F)** — **N**
- **Marital Status** — **N**

Relational data model

- Represent data as a **table** (*relation)*
- Each **row** (*tuple) represents a record*
  - Each record is a fixed-length tuple
- Each **column** (*attribute) represents a variable*
  - Each attribute has a *name and a data type*
- A table's **schema** is the set of names and types
- A **database** is a collection of tables (relations)

Relational Algebra

- Data Transformations (SQL)
- Projection (SELECT) - selects columns
- Selection (WHERE) - filters rows
- Sorting (ORDER BY)
- Aggregation (GROUP BY, SUM, MIN, MAX, …)
- Combine relations (UNION, JOIN, …)

Basic Statistical Descriptions of Data*

*(These are mainly for understanding individual attributes)

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- **Mean** (algebraic measure) (sample vs. population) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  Note: *n* is sample size and *N* is population size. $\mu = \frac{\sum x}{N}$

  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: chopping extreme values

Measuring the Central Tendency

- **Median**:
  - Middle value if odd number of values, or average of the middle two values otherwise

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

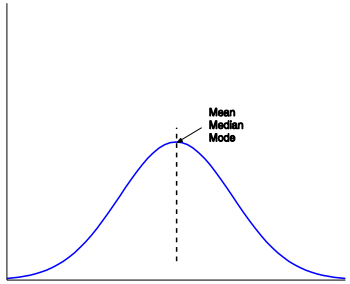
Measuring the Central Tendency

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mode \approx mean - 3 \times (mean - median)$$

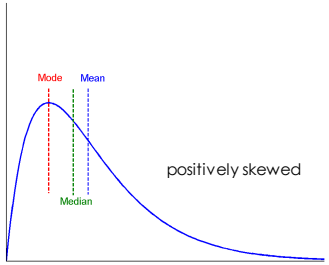
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



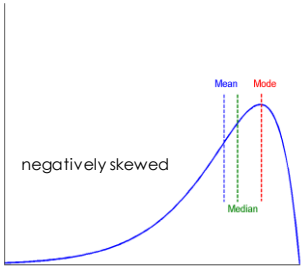
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

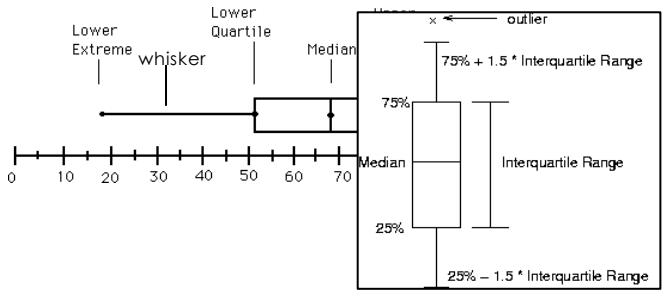


Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Five number summary: min, Q_1, median, Q_3, max
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
    - Below $Q_1 - 1.5 \times IQR$, or Above $Q_3 + 1.5 \times IQR$

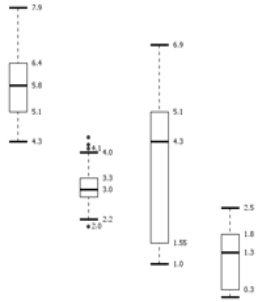
Boxplot

- Five-number summary of a distribution
  - Minimum, Q1, Median, Q3, Maximum



Boxplot Analysis

- Data is represented with a box
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extended to Minimum and Maximum
- Outliers: points beyond a specified outlier threshold, plotted individually



Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

What's Next

- Read section 1 and 2 in Data Mining Concepts and Techniques
- Email me at least two Practicum topics by tonight!
- Let me know names of your team members by Sept 13