Introduction to Data Science

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

Practicum

- Today
  - Tableau (Sirisha)
  - Scikit-learn (Vishal)
- Next Thursday
  - AWS (Anuj)
  - Tableau (Astha)
  - Pandas (Prakhar & Hans)
  - OpenRefine (Ashwani)

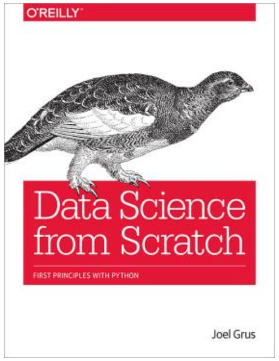
Project Teams

- Sahit Katragadda, Sree Vidya Susarla, Shivakesh Reddy Annepally
- Arshdeep Singh, Venkatesh Kadali, Rohit Gopalan
- Sidhant Chatterjee, Sirisha Rani Deekonda
- Jake Felzien, Mike Larsen, and Yancy Knight
- Bhagyashree, Meiling, Anuj
- Mounika, Akhil, Pravallika
- Prakhar Amlathe, Amitesh Mahajan, Vaibhav Sahu
- Michael (Troy) Harris, Hans Gunther, Karun Joseph
- Astha Tiwari, Kamna Yadav, Ashwani Chahal
- Ruchi Chauhan, Vishal Sharma
- Nick, Seyed, Jacob

So far 31 students
3 students missing?

New Book

- Data Science from Scratch



New Book

Chapter 1. Introduction
Chapter 2. A Crash Course in Python
Chapter 3. Visualizing Data
Chapter 4. Linear Algebra
Chapter 5. Statistics
Chapter 6. Probability
Chapter 7. Hypothesis and Inference
Chapter 8. Gradient Descent
Chapter 9. Getting Data
Chapter 10. Working with Data
Chapter 11. Machine Learning
Chapter 12. k-Nearest Neighbors
Chapter 13. Naive Bayes
Chapter 14. Simple Linear Regression
Chapter 15. Multiple Regression
Chapter 16. Logistic Regression
Chapter 17. Decision Trees
Chapter 18. Neural Networks
Chapter 19. Clustering
Chapter 20. Natural Language Processing
Chapter 21. Network Analysis
Chapter 22. Recommender Systems
Chapter 23. Databases and SQL
Chapter 24. MapReduce
Chapter 25. Go Forth and Do Data Science

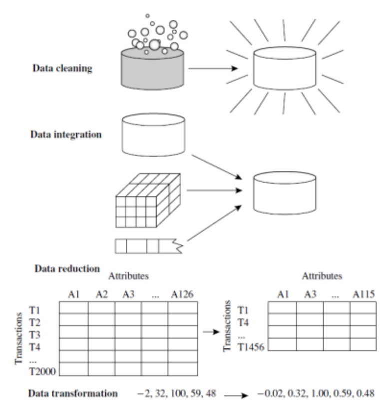
HW1

- https://usu.instructure.com/courses/431417/assignments/2117107
- Due date: September 22

Previous Class…

Measuring Similarity of documents (texts)
→ Cosine Similarity

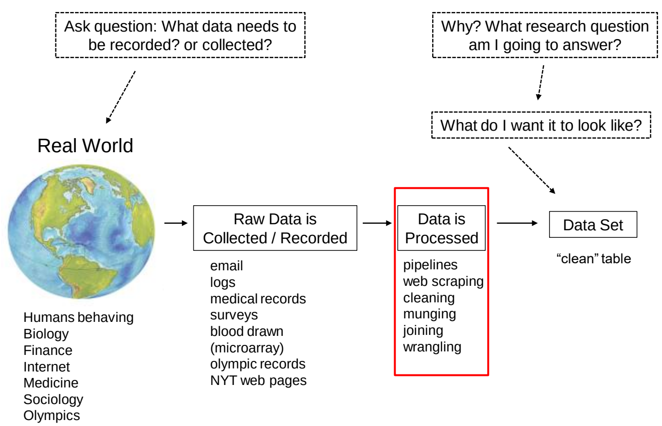
Previous Class…

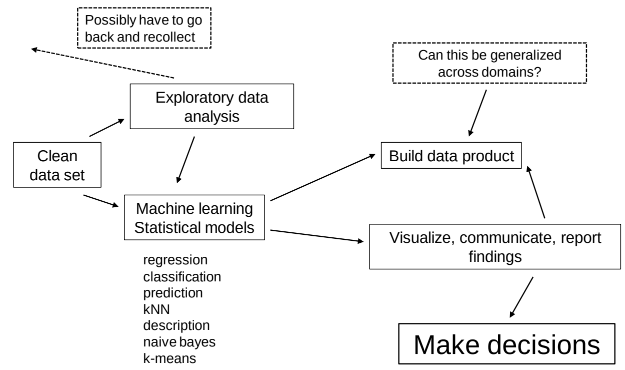


Previous Class…

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases/data sources, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization

Data Science: The Context





Data Integration

Correlation Analysis (Numeric Data)

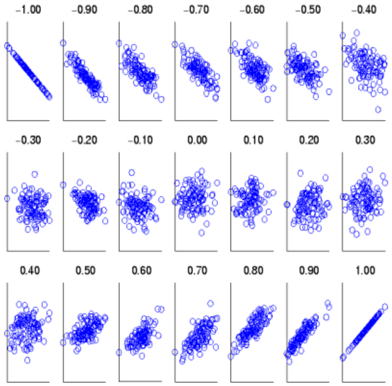
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher the value, the stronger the correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then if A is larger than its expected value, B is also likely to be larger than its expected value.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Covariance

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - $\overline{A}$ =E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - $\overline{B}$ =E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

A comparison of correlation and covariance

- Although both the correlation coefficient and the covariance are measures of linear association, they differ in the following ways:
  - Correlations coefficients are standardized. Thus, a perfect linear relationship results in a coefficient of 1.
  - Covariance values are not standardized. Thus, the value for a perfect linear relationship depends on the data.
- The correlation coefficient is a function of the covariance. The correlation coefficient is equal to the covariance divided by the product of the standard deviations of the variables. Therefore, a positive covariance always results in a positive correlation and a negative covariance always results in a negative correlation.

http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/correlation-and-covariance/basics-of-correlation-and-covariance/

Data Reduction

Data Reduction

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store petabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

Data Reduction

- Dimensionality reduction, e.g., remove unimportant attributes
  - Principal Components Analysis (PCA)
  - Feature selection (i.e., Attribute subset selection), attribute creation
- Numerosity reduction (some simply call it: Data Reduction)
  - Parametric methods: assume the data fits some model, estimate model parameters, store only the parameters, and discard the data
    - Regression and Log-Linear Models
  - Non-parametric methods: do not assume models
    - Histograms, clustering, sampling

Data Reduction Method: Non-parametric methods

- Histograms
  - Divide data into buckets and store average (sum) for each bucket
- Clustering
  - Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Sampling
  - Choose a representative subset of the data
  - Stratified sampling: approximate the percentage of each class