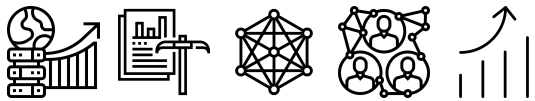


CSC 5741
Lecture 7: Linear Regression, Classification and Clustering

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia

Announcements—May 14, 2019

- Assessments
  - Class Theory Test: May 21, 2019
  - Mini Project Deliverables: May 20, 2019
    - (i)Technical Report; (ii) Code Repository for Fully Functional Implementation (including interactive Jupyter Notebook) + Labelled Dataset; (iii) Presentation Slides
  - Mini Project Presentations: Mary 28, 2019
    - Presentations [10 minutes]; Demonstrations [2 minutes]; Q&A [3 minutes]
  - Epilogue Lecture: May 28, 2019
    - Theory of Estimators
    - Academic Talk + Beyond CSC 5741

Lecture Series Outline

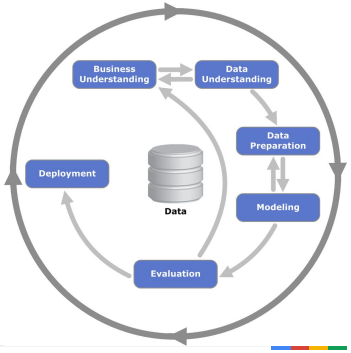
- Part I: Linear Regression, Classification and Clustering
- Part II: Jupyter Notebook Walkthrough

Lecture Series Outline

- Part I: Linear Regression
  - Introduction
  - Regression
  - Linear Regression
  - Classification
  - Clustering
- Part II: Jupyter Notebook Walkthrough

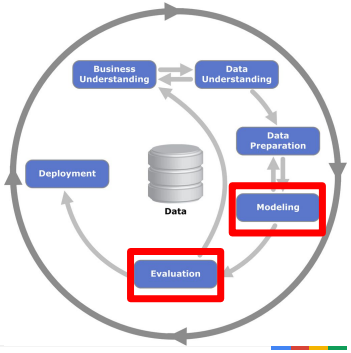
Introduction (1/3)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
  - CRISP-DM segments a data mining project into six phases with no strict order of execution
  - Surveys conducted suggest CRISP-DM is the most widely used methodology



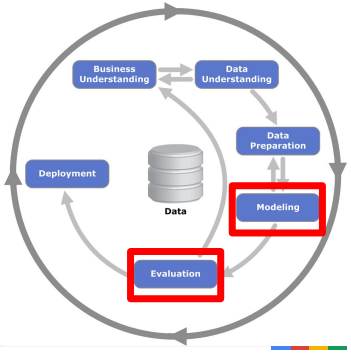
Introduction (2/3)

- Define the model components, features, how it behaves and how to interpret it
- Evaluate the various alternative techniques that can be integrated with the model
  - e.g. Evaluate different classification algorithms



Introduction (3/3)

- Finding patterns in data that provide insight or enable fast and accurate decision making
  - Prediction
  - Pattern recognition



Regression (1/2)

- Regression generally involves predicting one variable from another
- It is a statistical modeling technique that evaluates the relationship between one variable (dependent variable) and one or more other variables (independent variables)
- Uses a single equation for determining the relationship between the dependent variable and the independent variables

Regression (2/2)

- **Variable**
  - Any factor that can take on a value
    - Definition of value is aligned with data attributes—numeric, categorical, ordinal
- **Dependent variable**
  - The observed or measured variable
- **Independent variable**
  - Variable that is manipulated in order to observe desired outcome

Linear Regression (1/3)

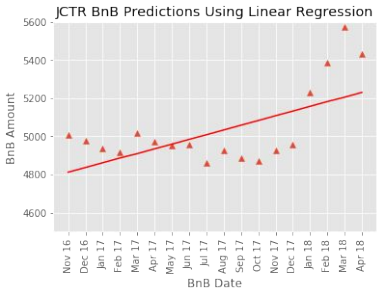
- **Linear Regression is used to fit a linear model to data where the dependent variable is continuous/numeric variable**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

- **Given a set of points (Xi,f(xi)), we wish to find a linear function (or line in 2 dimensions) that "goes through" these points.**

Linear Regression (2/3)

- **The associated error is computed by finding the distance between the data point and the straight line**
  - Observed value - Predicted value
  - Yi - f(xi)



Linear Regression (3/3)

- **Sum of Squared Errors (SSE) typically used to determine the accuracy of the linear equation**

$$SSE = \sum_{y} (y_{observed} - y_{predicted})^2$$

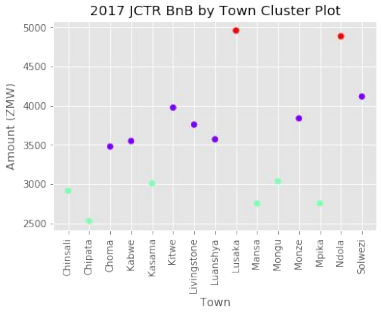
- **A small SSE value implies a better fit and is thus desirable**
- **The goal of Linear regression is to minimize SSE**

Classification (1/)

- **Classification involve the prediction of a categorical variable**
  - Binary classification involves two categorical variables
  - Multilabel classification involves more than two categorical variables
  - Multiclass classification associates multiple labels to one outcome

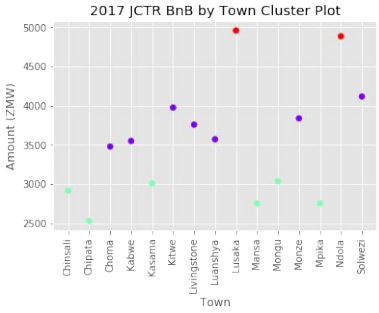
Clustering (1/)

- **Clustering is a pattern recognition technique that groups observations into groups that have meaning in the context of a particular problem.**



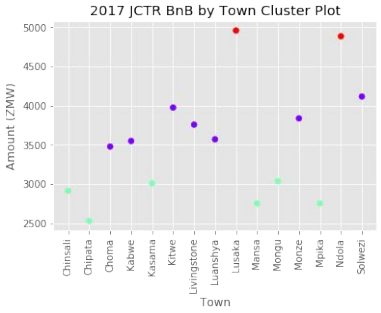
Clustering (2/)

- **Clustering is an unsupervised learning techniques**
  - Inputs are organized into an efficient representation that characterizes them.
  - Unlike linear regression and classification, does not rely on predefined classes.
  - It can uncover previously undetected relationships in a complex data set.



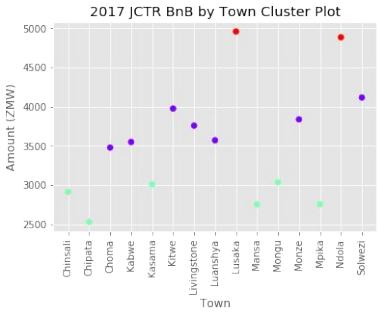
Clustering (3/)

- **Two main clustering approaches: non-hierarchical and hierarchical**
  - In nonhierarchical clustering, the relationship between clusters is undetermined.
  - In hierarchical clustering repeatedly links pairs of clusters until every data object is included in the hierarchy



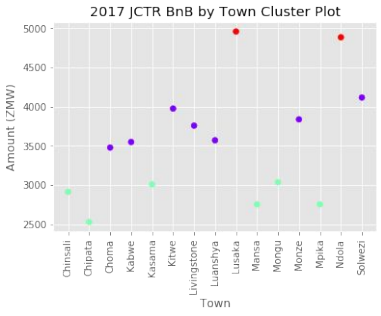
Clustering (3/)

- **Two main clustering approaches: non-hierarchical and hierarchical**
  - In nonhierarchical clustering, the relationship between clusters is undetermined. Opposite is true for hierarchical clustering



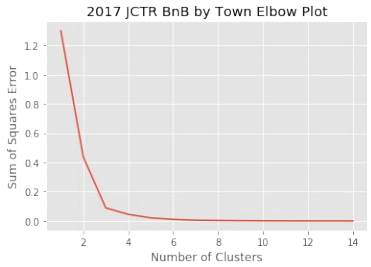
Clustering (4/)

- **Example in Jupyter Notebook uses K Mean clustering—a non-hierarchical clustering approach**
  - Select k clusters
  - Set random centroids: centers with those k clusters
    - reassigning all data objects to their closest cluster
    - Compute new cluster centers as mean value



Clustering (5/)

- **Elbow plot can be used to evaluate optimal number of clusters**



Q & A Session

- **Comments, concerns and complaints?**

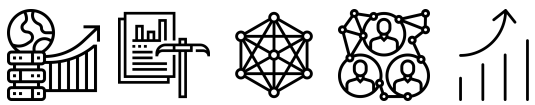
Lecture Series Outline

- Part I: Linear Regression
- Part II: Jupyter Notebook Walkthrough
  - Univariate Linear Regression
  - Multivariate Linear Regression
  - Binary Classification
  - Multilabel Classification
  - K Means Clustering

Bibliography

[1] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 2
https://www.cs.waikato.ac.nz/ml/weka/book.html

[2] An introduction to machine learning with scikit-learn
https://scikit-learn.org/stable/tutorial/basic/tutorial.html



CSC 5741
Lecture 7: Linear Regression, Classification and Clustering

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia