Data Analytics

Objectives

This module aims at developing the following competences:

1. Improve the condition of a real large dataset for data analysis
2. Analyze data to draw meaningful insights to solve real-world problems
3. Prepare the graphical representation of information and data for the effective communication of analytical tasks

Learning Outcomes

Upon the completion of this module, the trainees will be able to:

1. Demonstrate understanding of improving the condition of the provided datasets
2. Demonstrate understanding of conducting statistical analysis on the provided dataset
3. Demonstrate understanding of basic data mining techniques (prediction, classification, clustering, association analysis)
4. Demonstrate understanding of preparing the graphical representation of information for the analysis of the provided datasets
5. Demonstrate capability to improve the condition of the practical real large datasets
6. Demonstrate capability to analyze the practical dataset by using statistical analysis and/or analytic algorithms
7. Demonstrate capability to prepare the graphical representation of information for the analysis of the practical datasets.
8. Apply data analytics to model and interpret the data systematically.

Prerequisite: None

Outline:

- **Data inspection and cleaning**
  - Data inspection via graphical and numerical analysis
  - Treatment for missing data
  - Treatment for inconsistency
  - Treatment for multi-distribution
  - Identifying outliers and their treatment
- **Statistical Inferences and Hypothesis Testing**
  - Point Estimation and Required Properties of Point Estimators
  - Interval Estimations for Mean, Proportion and Variance of Population
  - Sample Size Determination
  - Hypothesis Testing for Mean, Proportion and Variance of Population – Single Sample Test
  - Hypothesis Testing for Mean, Proportion and Variance of Population – Two Samples Test

  - Type I and Type II Errors – Power of the Test
  - Observed Significance Level

- **Regression Analysis**
  - Linear Regression and Least Square Method
  - Residual Analysis
  - Multiple Regression
  - Goodness of Fit Tests
- **Data Classification**
  - k-Nearest Neighbor Algorithm for Estimation and Prediction
  - Distance Functions: Euclidian, Manhattan, Minkowski, Min-Max Normalization, Z-Score Standardization
  - Logistics Regression
  - Bayesian Networks
  - Model Evaluation Measures for Classification Task
- **Data Clustering**
  - Hierarchical Clustering Method
  - k-Means Clustering
  - Measuring Cluster Goodness: The Silhouette Method and The Pseudo-F Statistic
- **Association Rules**
  - Affinity Analysis
  - The a Priori Algorithm – Generating Frequent Itemsets
  - The a Priori Algorithm – Generating Association Rules
  - Measure the Usefulness of Associate Rules
- **Practicing statistical analysis and data analytic algorithms on provided data sets to draw meaningful conclusions**
- **Graphical methods for describing the data analytics' models.**
  - Scatter plots
  - Overlay plots
  - 3D plots
- **Practicing on improving the condition of data using real datasets.**
- **Practicing on preparing the graphical representation of data analytics' models.**

<u>Learning Activities</u>:

- Short lectures
- Class discussion
- Group discussion
- Group work
- In-class assignment
- Project assignment
- Oral presentation

<u>Time Distribution and Study Load:</u>

- Training: 15hours
- Coaching: 30 hours
- Group project: ~~80~~ 60 hours

Assessments:

- Class discussion and participation
- In-class assignment
- Project assignment
- Presentations

Developer(s): Dr. Huynh Trung Luong (AIT, Thailand) and Dr. Danaipong Chetchotsak (KKU, Thailand)