Data Science I

Description:

This course introduces applied machine learning, with a greater focus on practical techniques. The main goal of machine learning is to develop computer models that can learn from historical data to predict the new unseen data. In the past few decades, machine learning has become a powerful tool in artificial intelligence and data mining, and it has made major impacts in many real-world applications (Transportation and logistics, Renewable energy sources, Health care, etc.).

This course covers a variety of topics including supervised learning (Regression and Classification) and validation. To develop the design and programming skills that will assist students to apply ML techniques in different research areas, we will have Python coding sessions to carry out some analysis.

Course Objectives:

At the end of this course, students should be able to (1) understand fundamental concepts of several major topics in Machine Learning, (2) perform a proper technique on real data, and (3) interpret the analysis. More advanced ML topics will be covered in Data Science II.

Prerequisite:

Students are expected to have pre-existing knowledge of probability, linear algebra, statistics, and basic programming.

Course schedule:

The approximate number of lectures is tentative. Class time will also be used to cover programming sessions, conduct exam questions, and review sessions. Chapter sections are listed below.

0. ML fundamentals: Probability, Hypothesis testing
1. Introduction: Machine Learning, Supervised learning
2. The Multiple Linear Regression Model: Loss Function, Gradient descent, Stochastic Gradient, Vectorization and Analysis of variance
3. Logistic Regression: Binary Classification and K-Nearest Neighbor (KNN)
4. Regularization: Lasso, Ridge Regression
5. Feature selection: Principle Component Analysis
6. Validation: Bias-Variance tradeoff, Overfitting, Cross-validation, and model selection
7. Neural Networks

Course Materials:

There is one required text book for this course.

The Elements of Statistical Learning: Data Mining, Inference and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2001.

The following books are recommended as optional reading:

Tom Mitchell, Machine Learning. McGraw-Hill, 1997.

Richard Sutton and Andrew Barto, Reinforcement Learning: An introduction. MIT Press, 1998.

Applied Linear Statistical Models, Fifth Edition by Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li, 2004.