Day 1 – Introduction to Machine Learning for Data Science: Part 1

Module 1 – Supervised Learning

- You Are the Teacher
  - What Is Machine Learning?
  - Supervised Learning
  - Classification Versus Regression
- Why Decision Trees?
  - Ease of Use
  - Optimal for Many Business Problems

Module 2 – Classification Trees

- Basic Intuition
  - Trees Are Rules
  - Sample Decision Tree
- Overfitting Intuition
  - The Bugbear of Machine Learning
  - The Model Is Good! Or Is It?

Module 3 – Classification Tree Math

- Gini Impurity
- Gini Change
- Many Categories Impurity
- Numeric Feature Impurity

Module 4 – Using Classification Trees

- Building Classification Trees
  - Model Specifications
  - Workflows
  - Model Fitting
- Hands-On Lab #1

Module 5 – Introducing the Bias-Variance Tradeoff

- Under/Overfitting
    - The Goldilocks Zone
    - Controlling Complexity
- The Bias-Variance Tradeoff
    - Intuitive Example
    - Model Example

Module 6 – Model Tuning

- Supervising the Data
    - Splitting the Data
    - Cross-Validation
- Model Tuning Intuition
    - Making an Intuitive Example Real
    - Estimating Generalization Error
    - What About the Test Set?
- Pruning Classification Trees
    - Pruning Intuition
    - Pre-Pruning
    - Post-Pruning

Module 7 – Model Tuning

- Measuring Model Accuracy
    - Accuracy
    - Confusion Matrices
    - Sensitivity
    - Specificity
- Performing Model Tuning
    - Setting Up Cross-Validation
    - Cross-Validation Results
    - Tuning the Tree
    - Tuning Results
- Hands-On Lab #2

Day 2 – Introduction to Machine Learning for Data Science: Part 2

Module 8 – Feature Engineering

- Intuition
    - What Is Feature Engineering?
    - An Example
    - Extracting Features
    - Row Versus Column Features
- Data Leakage
    - What Is It?
    - An Example
    - Avoiding Data Leakage
- Engineering Features for Decision Trees
    - Decision Boundaries
    - Visualizing Decision Boundaries
    - Concepts to Remember
- Missing Data
    - Why Is Data Missing?
    - Dealing with Missing Data
    - What Is Imputation?
    - Performing Imputation
- Hands-On Lab #3

Module 9 – Regression Trees

- The Basics
    - Regression Trees Minimize SSE
    - Calculating SSE
- Numeric Feature SSE
- Many Categories SSE
- Building Regression Trees
    - Measuring Accuracy
    - Model Specification
    - Regression Trees in Practice

Module 10 – The Mighty Random Forest

- Bad, Tree! Bad!
    - Decision Tree Variance
    - High Variance Leads to Overfitting
    - Real-World Decision Trees
- Ensembles
    - Wisdom of the Crowd
    - Manufacturing Independence
- Bagging
    - Randomizing Rows
    - Bagging in Action
    - The Power of Bagging
- Feature Randomization
    - Intuition
    - Randomizing Columns
    - Feature Randomization in Action

Module 11 – Using the Random Forest

- Tuning Random Forests
    - The Bias-Variance Tradeoff
    - Random Forest Hyperparameters
- Feature Importance
    - Out of Bag (OOB) Data
    - Permutation Importance
    - An Example
- Building Random Forests
- Hands-On Lab #4

Module 12 – Workshop Wrap-Up

- Want to Kaggle?
- Additional Resources

Day 3 – Cluster Analysis for Data Science

Module 1: Introduction

- What is Cluster Analysis?
- Cluster Analysis Use Cases
- The Challenge of Clustering Data

Module 2 – Data Sets Used in the Course

- The Iris Data Set
- The Hand-Written Digits Data Set
- The Heart Data Set

Module 3 – Types of Clusterings and Clusters

- Hierarchical, Partitional, and Overlapping Clustering
- Prototype Clusters
- Density-Based Clusters

Module 4 – K-Means Clustering

- Introducing K-Means
- The K-Means Algorithm
- Euclidian Distance
- The Problem with Outliers
- Data Standardization
- K-Means Caveats
- Hands-On Lab #1

Module 5 – Optimizing K-Means

- Evaluating Clusters
- Cluster Cohesion
- Evaluating Cohesion with the Elbow Method
- The Silhouette Coefficient
- Evaluating a Clustering Using the Silhouette Score
- Hands-On Lab #2

Module 6 – DBSCAN Clustering

- Introducing DBSCAN

- The DBSCAN Algorithm
- DBSCAN Caveats

Module 7 – Optimizing DBSCAN

- Considerations for Optimizing DBSCAN
- Calculating min_samples
- Choosing the eps Value
- Introducing Nearest Neighbors
- Evaluating eps Using the Elbow Method
- K-Means vs DBSCAN
- Hands-On Lab #3

Module 8 – Dimensionality Reduction

- Introducing Dimensionality Reduction
- Introducing Principal Component Analysis (PCA)
- PCA Concepts
- Hands-On Lab #4

Module 9 – Categorical Data

- The Problem with Categories
- One-Hot Encoding
- Factor Analysis of Mixed Data (FAMD)

Module 10 – Additional Resources

Hands-On Lab #5