Machine Learning

Machine Learning at its heart, is concerned with algorithms that transform information into actionable intelligence. This makes Machine Learning amazingly well suited for the present-day era of Big data. Without Machine learning it would be impossible to keep up with the massive streams of information that is at our disposal.

This course aims to combine hands on case studies with the essential theory so that can understand how things work under the hood. We would use R as the chosen programming language. R programming is growing in prominence as a cross platform, zero cost statistical programming environment and offers powerful but easy to learn set of tools that can assist in finding insights in data.

Introduction to Machine Learning.

- The origins of Machine Learning – uses and abuses of Machine learning.
    - Machine learning successes.
    - The Limits of Machine learning.
    - Ethics & Machine learning.
- How machine learn.
    - Data storage.
    - Abstraction.
    - Generalization.
    - Evaluation.
- Machine learning in Practice.
    - Types of input data.
    - Types of machine learning algorithms.
    - Machine learning input data to algorithms.
- Machine learning with R.
    - Installing R packages.
    - Loading and unloading R packages.

Lazy learning – Classification using Nearest Neighbours.

- Understanding nearest neighbour classification.
    - The k-NN algorithm.
        - Measuring similarity with distance.
        - Choosing appropriate k.
        - Preparing data for use with k-NN.
    - Why is the k-NN algorithm lazy?
- **Case 1 – diagnosing breast cancer with the k-NN algorithm.**
    - Step 1: Collecting data.
    - Step 2: Exploring and preparing data.
        - Transformation – normalizing numeric data.
        - Data preparation – creating training and test datasets.

- Step 3: Training a model on the data.
- Step 4: Evaluating model performance.
- Step 5: Improving model performance.
  - Transformation – z score standardization.
  - Testing alternative values of k.

Probabilistic Learning – Classification using Naïve Bayes

- Understanding Naïve Bayes
  - Basic concepts of Bayesian methods.
  - Understanding probability.
  - Understanding joint probability.
  - Computing conditional probability with Bayes Theorem.
- The Naïve Bayes Algorithm.
  - Classification with Naïve bayes.
  - The laplace estimator.
  - Using numeric features with Naïve bayes.
- **Case 2 – filtering mobile phone spam with Naïve Bayes Algorithm.**

  Step 1: Collecting data.

  Step 2: Exploring and preparing the data.

  Data preparation- cleaning and standardizing text data.

  Data preparation – splitting text documents into words.

  Data preparation – creating training and test datasets.

  Visualizing test data – word clouds.

  Data preparation – creating indicator features for frequent words.

  Step 3: Training the model on the data.

  Step 4: Evaluating model performance.

  Step 5: Improving model performance.

Classification Using Decision Trees and Rules.

- Understanding decision trees.
  - Divide and conquer.
  - The C5.0 decision tree algorithm.
    - Choosing the best split.
    - Pruning the decision tree.
- Case 3 – Identifying risky bank loans using C5.0 decision tree.
  - Step 1: Collecting data.
  - Step 2: Exploring and preparing the data.
    - Data preparation – creating random training and test datasets.
  - Step 3: Training a model on the data.
  - Step 4: Evaluating model performance.

    - Step 5: Improving model performance.
        - Boosting the accuracy of decision trees.
        - Making some mistakes, that can be costly.
- Understanding classification rules.
    - Separate and conquer.
    - The 1R algorithm.
    - The RIPPER algorithm.
    - Rules from the decision trees.
    - What makes trees and rules greedy?
- **Case 4 – Identifying poisonous mushrooms with the rule learners**
    - Step 1: Collecting data.
    - Step 2: Exploring and preparing data.
    - Step 3: Training a model on the data.
    - Step 4: Evaluating model performance.
    - Step 5: Improving model performance.

Regression Methods

- Understanding regression.
    - Simple linear regression
    - Ordinary least squared estimation.
    - Correlations.
    - Multiple linear regression.
- Case 5: Predicting medical Expenses using Linear regression.
    - Step 1: collecting data.
    - Step 2: exploring and preparing data.
        - Exploring relationships among features- the correlation matrix.
        - Visualizing relationships among features- the scatterplot matrix.
    - Step 3: training a model on the data.
    - Step 4: evaluation model performance.
    - Step 5: improving the model performance.

Black Box Methods – Neural Networks.

- Biological to artificial neurons.
- Activation functions.
- Network topology.
    - The number of layers.
    - The direction of information travel.
    - The number of nodes in each layer
- Training the Neural Network with backpropagation.
- Case 6 – Modelling the strength of concrete with ANNs.
    - Step 1: Collecting data
    - Step 2: exploring and preparing data.
    - Step 3: training a model on the data.
    - Step 4: evaluating model performance.
    - Step 5: improving model performance.

Finding Patterns – Market basket analysis using Association Rules.

- Understanding association rules.
    - The Apriori algorithm for association rule learning.
    - Measuring rule interest – support and confidence.
    - Building a set of rules with Apriori principle.
- Case 7 – Identifying frequently purchased groceries with association rules.
    - Step 1: Collecting data
    - Step 2: Exploring and preparing data.
        - Data preparation – creating sparse matrix for transaction data.
        - Visualizing item support – item frequency plots.
        - Visualizing the transaction data – plotting the sparse matrix.
    - Step 3: training the model on the data.
    - Step 4: evaluating model performance.
    - Step 5: improving model performance.
        - Sorting the set of association rules.
        - Taking subsets of association rules
        - Saving association rules to a file or data frame.

Finding Groups of Data – Clustering with k-Means

- Understanding Clustering
    - Clustering as a machine learning task.
    - The k-means clustering algorithm.
        - Using distance to assign and update cluster.
        - Choosing the appropriate number of cluster.
- **Case 8 – finding the teen market segment using k-means clustering**
    - Step 1: collect data
    - Step 2: exploring and preparing the data.
        - Data preparation – dummy coding, missing values.
        - Data preparation – imputing the missing values.
    - Step 3: training a model on the data.
    - Step 4: evaluation model performance.
    - Step 5: improving model performance.

Evaluating Model Performance:

- Measuring Performance for classification
    - Understanding a classifier's predictions.
    - A closer look at confusion matrices.
    - Using confusion matrices to measure performance.
    - Beyond accuracy – other measures of performance.
        - The kappa statistic
        - Sensitivity and specificity
        - Precision and recall
        - The F-Measure
    - Visualizing performance trade-offs with ROC curves.

- Estimating future performance
    - The holdout method
        - Cross validation.
        - Bootstrap sampling.

Improving Model Performance:

- Tuning stock models for better performance
    - Using caret for automated parameter tuning.
        - Creating simple tuned model.
        - Customizing the tuned process.
- Improving the model performance with meta learning
    - Understanding ensembles.
    - Bagging
    - Boosting
    - Random forests
        - Training random forests.
        - Evaluating random forest performance.