Data Mining Concepts & Techniques

by Deanne Larson

Module 0. About the Course (3 min)

Module 1. Introduction to Data Mining (25min)

- *Overview*
- *Module Overview*
- *What is Data Mining*
  - Where Did Data Mining Originate?
  - Why Data Mining?
  - What is Data Mining?
  - The Process of Data Science
  - Examples of Data Mining
  - Data Mining Methods
- *Statistics in Data Mining*
  - Statistics in Data Mining
  - Statistics Basics−Attributes
  - Summary Statistics: Categorical
  - Summary Statistics: Percentiles
  - Summary Statistics: Measures of Location
  - Summary Statistics: Measures of Spread
  - Multivariate Summary Statistics
- *Machine Learning*
- *Supervised Learning*
- *Unsupervised Learning*
- *Summary*

Module 2. The Data Mining Process (24min)

- *Module Overview*
- *Data Mining Framework*
- *Data Mining Approaches*
- *Data Mining Techniques*
  - Classification
  - Association
  - Sequencing
  - Forecasting and Prediction
  - Data Mining Algorithm
- *Data Mining Process*
  - Define the Scope
  - Collect the Data
  - Explore the Data
  - Data Reduction and Cleansing
  - Build the Model
  - Evaluation and Interpretation
  - Model Deployment
  - Life of the Model
  - CRISP-DM Parts 1 & 2
- *Summary*

Module 3. Exploratory Data Analysis (29min)

- *Overview*

- *Exploratory Data Analysis*
- *Data Profiling: Uncovering Structure*
- *Data Profiling: Types of Profiling*
- *Descriptive Statistics*
  - Descriptive Statistics Definition
  - Examples
  - Additional Examples
- *Results of Data Profiling and Descriptive Statistics*
- *Data Relationships*
  - Data Relationships
  - Outliers and Anomalies
- *Findings – Important Variables*
- *Visualization Techniques*
  - Basic
  - Distribution
  - Advanced
- *Outcomes and Interpretations*
- *Sampling Size*
- *Sample Quality*
- *Big Data Considerations*
- *Feature Selection*
- *EDA Checklist*
- *Summary*

Module 4. Data Mining Models and Algorithms (71min)

- *Overview*
- *Build the Model*
- *Anatomy of a Model*
- *What is a Classification Problem*
- *Classification*
  - Decision Trees
  - K Nearest Neighbor
  - Probability Bayes Classification
  - Neural Networks
  - Support Vector Machine
- *Predictive Data Mining*
  - Predictive Data Mining
  - Linear Regression
  - Linear Regression Example
  - Linear Regression – Error Difference
  - Logistic Regression
  - Logistic Regression Uses
  - Other Regression Options
- *Ensemble Methods*
  - *Introduction*
  - *Bagging*
  - *Boosting*
  - *Bagging and Boosting Summary*
  - *Ensemble Method Uses*
- *Clustering*

- *Clustering Uses*
- *Association−Market Basket*
- *Association Rules*
- *Association Uses*
- *Anomaly Detection*
- *Application of Data Mining Models*
- *Model Selection*
- *Summary*

Module 5. Model Validation Techniques (18min)

- *Module Overview*
- *The Validation Process*
- *Fitting a Model*
- *Bias/Variance Tradeoff*
- *Regression – Mean Squared Error*
- *Linear Regression – Confidence and Prediction Intervals*
- *Logistic Regression – Significance Test*
- *Classification Accuracy*
- *Classification Accuracy – Other Measures*
- *Prediction Error Methods*
- *Hold-Out Cross Validation*
- *K-Fold Cross Validation Method*
- *Summary*

Module 6. Deploying Data Mining Models (9min)

- *Deploying Data Mining Models*
  - The Real Value
  - Scheduled Execution
  - Model Support
  - Review and Revision
  - Deploying Data Mining Models Summary
- *Course Summary Parts 1 & 2*
- *References*