Module 5
Advanced Big Data Analysis & Science

Exploratory Data Analysis

- data reduction
  - dimensionality reduction
    - forward selection
    - backward elimination
    - decision tree induction
    - feature extraction
  - data discretization
    - binning
    - clustering

Classification

- logistic regression
- linear regression
- decision trees
- pre-pruning
- post-pruning
- feature splitting
- entropy
- information gain
- classification rules
- rule-based model
  - one rule (1R) algorithm
  - k nearest neighbor (kNN)
- naïve Bayes
- Bayes' theorem
- Laplace smoothing

Outlier Detection

- outlier types
  - global
  - contextual
  - collective
- statistical techniques
  - parametric
  - non-parametric
  - distance-based/unsupervised
  - supervised
  - semi-supervised
  - clustering
    - k-means
    - cluster-based local outlier factors (CBLOF)
  - non-clustering

Modeling

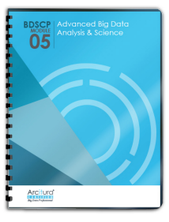
- machine learning algorithms
- statistical models
- predictive modeling
- feature vector
- instance/example
- target
- concept

Pattern Identification

- association rules
  - actionable
  - trivial
  - inexplicable
- Apriori algorithm
- clustering
  - k-means
    - assign
    - update
    - reassignment
  - stages
- text analytics
  - bag of words
  - term frequency
    - inverse document frequency (IDF)
    - term frequency inverse document frequency (TFIDF)
  - cosine distance
  - n-grams
  - named entity extraction
  - text representation
    - token/term
    - document
    - corpus

Model Evaluation Measures

- sensitivity
- specificity
- recall
- precision
- accuracy
- error rate
- f-score
- confusion matrix
- cross-validation
- bias-variance



Module 5: Advanced Big Data Analysis & Science
Official Mind Map Supplement

Arcitura CERTIFIED Big Data Scientist

Arcitura Big Data School - Big Data Scientist Certification Program
Copyright © Arcitura Education Inc. www.arcitura.com

Arcitura TRAINING Big Data Science

Arcitura CERTIFIED Big Data Scientist