**Course 2. Data Mining (4 credits)**

By the time you finish this course, you’ll know how to preprocess data for data mining, explore the data before applying data mining techniques, discover association patterns in a dataset, use advanced predictive modeling techniques to solve problems in real world, Use different advanced clustering techniques to summarize data. You will also analyze the complexity of these algorithms and use them appropriately to solve problems in the real world within available resources

Textbook: Introduction to Data Mining, PANG-NING TAN Michigan State University, MICHAEL STEINBACH University of Minnesota, VIPIN KUMAR University of Minnesota, Pearson/Addison Wesley

Module 1: Introduction

Learning Objectives:

* Explain how Data Mining has evolved
* Differentiate Information Retrieval and Data Mining
* Differentiate Data Mining tasks in to descriptive and predictive
* Discover challenges of finding knowledge from data
* Discover applications of Data Mining in different field

Module 2: Data Preprocessing

Learning Objectives:

* Discuss data-related issues that are important for successful data mining
* Describe different types of data
* Explain general characteristics of datasets
* Use sampling to select subset of data
* Explain the curse of dimensionality
* Discover methods for dimensionality reduction
* Discuss Feature subset selection, feature engineering
* Explain Discretization and Binarization
* Discuss Measures of Similarity and Dissimilarity of Attributes and data objects

Module 3: Exploring Data

Learning Objectives:

* Produce Statistical summary of the data
* Visualize data
* Explain how to visualize higher dimensional data
* Explain different multidimensional analysis technique

Module 4: Association Analysis: Basic Concepts, Apriori Algorithm

Learning Objectives:

* Define Association Analysis problem
* Explain Apriori Principle
* Use Apriori Algorithm to generate frequent Item sets and Rule generation
* Define Maximal frequent Itemset
* Define closed frequent itemset

Module 5: Association Analysis: Algorithms FP-Growth Algorithm

Learning Objectives:

* Explain FP-Growth Algorithm
* Use FP-Growth Algorithm to generate frequent Item sets and Rule generation
* Describe how to evaluate Association Patterns

Module 6: Association Analysis: Handling different types of attributes, Sequential patterns Learning Objectives:

* Transform the categorical and symmetric binary attributes into "items"
* Describe quantitative association rule
* Explain the three types of methods to generate quantitative association rules
* Discuss concept hierarchy in Itemset
* Use DAG to represent concept hierarchy
* Formulate sequential pattern problem
* Discover sequential patterns in a temporal dataset

Module 7: Association Analysis: Subgraph patterns, Infrequent patterns

Learning Objectives:

* Describe the application of association analysis methods to more complex entities beyond item sets and sequence (chemical compounds, 3-D protein structures, network topologies, and tree structured XML document)
* Explain graph representation of entities in various applications
* Use frequent subgraph mining to derive a set of common substructures among the collection of graphs.
* Define infrequent pattern
* Compare Infrequent Patterns, Negative Patterns, and Negatively Correlated Patterns

Module 8: Classification: Basic Concepts, Decision Trees

Learning Objectives:

* Explain classification task in datamining with examples
* Explain how a classification model can serve as an explanatory tool to distinguish between objects of different classes
* Extend classification model to a predictive model
* Explain how confusion matrix can be used for model evaluation
* Describe how a decision tree works as a classification model
* Use Hunt’s algorithm to build d decision tree
* Discover design issues of decision tree induction
* Explain overfitting and its causes
* Describe methods for estimating the generalizationerror of a model during training

Module 9: Classification: Model Evaluation and Comparing Classifiers

Learning Objectives:

* Explain evaluation methods of classifiers
* Evaluate the performance of a classifier by using Holdout method
* Evaluate the performance of a classifier by using Cross Validation method
* Evaluate the performance of a classifier by using Boot Strap method
* Discover Methods for Comparing Classifiers

Module 10: Classification: Nearest-Neighbor classifiers

Learning Objectives:

* Differentiate eager learners and lazy learners
* Explain the steps of Nearest-Neighbor classifier model building
* Describe characteristics of Nearest-Neighbor classifier

Module 11: Classification: Bayesian Classifiers

Learning Objectives:

* Explain the probabilistic relationships between the attribute set and the class variable
* State Bayes’ theorem
* Explain how Bayes’ theorem can be used for classification
* Describe how Naive Bayes Classifier works
* Describe how Bayesian Belief Network works

Module 12: Classification: Ensemble methods

Learning Objectives:

* Describe techniques ensemble or classifier combination methods for classification.
* Explain rationale for ensemble method
* Describe Methods for Constructing an Ensemble Classifier
* Explain Bagging and Boosting techniques
* Describe Random Forest classifier

Module 13: Classification: Class imbalance problem

Learning Objectives:

* Define Class Imbalance problem
* Discover alternating classifier evaluating methods such as ROC curve and cost sensitive learning
* Explain Sampling-Based Approaches to handle class imbalance problem

Module 14: Cluster Analysis: Basic concepts and K-means algorithm

Learning Objectives:

* Explain how cluster analysis helps for data summarization, data compression, efficiently finding nearest neighbors
* Describe cluster analysis applications to into different fields study
* Define what is cluster analysis
* Explain different types of clustering
* Explain methods of measure quality of clustering
* Explain issues in K-means clustering and how to overcome them
* Discuss strengths and weaknesses of k-means clustering
* Use K-means to find clusters in a dataset

Module 15: Cluster Analysis: Agglomerative Hierarchical Clustering

Learning Objectives:

* Explain Hierarchical Clustering
* Describe different types of Hierarchical Clustering
* Use Hierarchical Clustering to find clusters in a dataset
* Discuss strengths and weaknesses of Hierarchical Clustering

Module 16: Cluster Analysis: DB Scan

Learning Objectives:

* Explain what Density-Based clustering is
* Describe center-based approach for DB Scan
* Explain DB Scan algorithm
* Discuss strengths and weaknesses of DB Scan

Module 17: Cluster Analysis: Cluster evaluation

Learning Objectives:

* Differentiate evaluation methods for classification and cluster models
* Explain cohesion and separation to evaluate supervised clusters
* Define the Silhouette Coefficient
* Explain different methods to evaluate unsupervised clusters
* Compare K-means and DB Scan

Module 18: Self Organizing maps

Learning Objectives:

* Discuss the characteristics of data that can strongly affect cluster analysis
* Explain self-organizing maps (SOM)
* Describe SOM algorithm

Module 19: Graph based clustering

Learning Objectives:

* Explain Graph Based clustering
* Describe MST clustering algorithm
* Describe Chameleon Algorithm

Module 20: Anomaly detection

Learning Objectives:

* Define anomalies in a dataset
* Illustrate some applications for which anomalies are of considerable interest
* Discuss causes of anomalies
* Discuss different techniques of finding anomalies
* Discuss important issues that need to be addressed when dealing with anomalies