University of Jordan

King Abdullah II for Information Technology

Business Information Technology Department

Course: Data Warehousing and Mining (1904701)

Credit: 3 Credit hours

Semester and Year: 1st semester 2015/2016

Instructors	Office hours	Office Phone	E-Mail
Dr. Hossam Faris	Su, Tu and Th 12:00 – 1:00	22634	hossam.faris@ju.edu.jo

Course Description

This course provides students with an in-depth understanding of the design and implementation of data warehousing and data mining based systems. It will address the opportunities and challenges of applying data mining techniques in academics, industry, businesses, sciences and the Web. Several aspects of the data mining process are covered in this course such as: data gathering and storage, data selection and preparation, model building and testing, results interpretation and validation and models application.

Learning Outcomes

Upon completion of the course, students will be able to:

- Understand what data mining is and how data mining can be employed and applied to solve real problems.
- Recognize wither a data mining solution is feasible alternative for a specific problem.
- Apply basic statistical to evaluate the results of data mining models.
- Develop a comprehensive understanding of how several data mining techniques can be applied to solve problems.
- Understand the common designs and structures of warehouse systems.

Course Outline:

1. **Introduction to Data Mining**
   1.1. Types of Data: Database Data, Data Warehouses, Transactional Data, Other Kinds of Data
   1.2. Applications of Data Mining
   1.3. Labelled and Unlabelled Data
   1.4. Supervised Learning: Classification
   1.5. Supervised Learning: Numerical Prediction
   1.6. Unsupervised Learning: Association Rules
   1.7. Unsupervised Learning: Clustering

2. **Data Warehousing and Online Analytical Processing**
   2.1 Basic Concepts
       2.1.1   What Is a Data Warehouse?
       2.1.2   Differences between Operational Database Systems and Data Warehouses
       2.1.3   Why Data Warehouses?
       2.1.4   Data Warehousing: A Multitiered Architecture
       2.1.5   Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse
       2.1.6   Extraction, Transformation, and Loading
       2.1.7   Metadata Repositories
   2.2 Modeling: Data Cube and OLAP
       2.2.1   Data Cube: A Multidimensional Data Model
       2.2.2   Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models
       2.2.3   Dimensions: The Role of Concept Hierarchies
       2.2.4   Measures: Their Categorization and Computation
       2.2.5   Typical OLAP Operations
       2.2.6   A Starnet Query Model for Querying Multidimensional Databases
   2.3 Design and Usage
       2.3.1   A Business Analysis Framework for Data Warehouse Design
       2.3.2   Data Warehouse Design Process
       2.3.3   Data Warehouse Usage for Information Processing
       2.3.4   From Online Analytical Processing to Multidimensional
   2.4 Implementation
       2.4.1   Efficient Data Cube Computation: An Overview
       2.4.2   Indexing OLAP Data: Bitmap Index and Join Index
       2.4.3   Efficient Processing of OLAP Queries
       2.4.4   OLAP Server Architectures: ROLAP, MOLAP, HOLAP

3. **Data for Data Mining**
   3.1 Types of Variable
       3.1.1   Categorical and Continuous Attributes
   3.2 Data Preparation
   3.3 Data Cleaning
   3.4 Missing Values
   3.5 Discard Instances
   3.6 Replace by Most Frequent/Average Value
   3.7 Exploring Online repositories
   3.8 Feature Reduction Selection
       3.8.1   Wrappers
       3.8.2   Filters
       3.8.3   Embedded

4. **Evaluating Predictive Models**

4.1 Training/ Testing methodologies
4.2 True and False Positives and Negatives
4.3 Confusion Matrix

5. Selected Classification Models
5.1 Naïve Bayes Classifiers
5.2 Nearest Neighbour Classification
5.3 Decision Trees
5.4 Artificial Neural Networks

6. Clustering
6.1 K-means clustering
6.2 Hierarchical Clustering

7. Ensemble Classification
7.1 Estimating the Performance of a Classifier
7.2 Selecting a Different Training Set for Each Classifier
7.3 Selecting a Different Set of Attributes for Each Classifier
7.4 Combining Classifications: Alternative Voting Systems
7.5 Parallel Ensemble Classifiers

8. Comparing Classifiers
8.1 The Paired t-Test
8.2 Choosing Datasets for Comparative Evaluation
8.3 Confidence Intervals
8.4 Sampling

Text Books:

- Principles of Data Mining, Max Bramer, Springer
- Data Mining Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank and Mark A. Hall, Elsevier
- Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier