CSCI446/946 Big Data Analytics

Summarization

Subject Description

- This subject covers the principles, techniques and applications of processing and analysing big data.
- This subject will equip students with the fundamental knowledge on big data analytics and the skills to appropriately choose and apply algorithms to resolve practical analytics problems.

Subject Description

- The subject is organized in three parts:
  - The first part introduces fundamental concepts, platforms, systems, and tools for Big Data analysis, to give students a basic understanding of this field.
  - The second part introduces main analytics algorithms that are used to uncover the underlying patterns and rules in big data; including classification, regression, clustering, recommendation and retrieval algorithms.
  - The third part focuses on case studies and practical applications of big data analytics to help students gain a better understanding of these algorithms.

Subject Learning Outcomes

- On successful completion of this subject, students will be able to
  - Understand basic concepts of big data analytics.
  - Correctly apply the related core algorithm(s).
  - Choose appropriate algorithms to design and build systems to complete a big data analysis task.
  - Interpret and explain results.
  - Integrate the knowledge and skills learned in this subject to resolve practical issues.

Topics in delivered Lectures

- Big Data Overview
- Data Analytics Lifecycle
- Data Analytic Methods Using R
- Clustering
- Association Rules
- Regression
- Classification
- Text Analysis
- Image Analysis
- Time Series Analysis
- MapReduce and Hadoop
- Deep Graph Learning

Big Data Overview

- The definition and attributes of Big Data
- State of the practice in Analytics
- Key Roles for the New Big Data Ecosystem

Data Analytics Lifecycle

- Key Roles for a Successful Analytics Project
- Six phases in Analytics

Data Analytic Methods Using R

- Introduction to R
- Visualization
- Statistical Methods for Evaluation
  - Hypothesis Testing, ANOVA

Clustering

- Overview of Clustering
- K-means clustering
- Hierarchical Clustering

Association Rules

- Overview of Association Rules
- Apriori Algorithm
- Evaluation of Candidate Rules
- Validation and Testing

Regression

- Overview of Regression
- Linear Regression
- Logistic Regression
- ROC Curve
- Histogram of the Probabilities
- Deviance

Classification

- Overview of Classification
- Decision Tree
- Naïve Bayes
- Diagnostics of Classifiers

Text Analysis

- Overview of Text Analysis
- Representing Text
- Term Frequency --- Inverse Document Frequency (TFIDF)
- Categorizing Documents by Topics
- Determining Sentiments

Image Analysis

- Overview of Image Analysis
- Representing Image
- Bag-of-Visual-Words model
- Deep Convolutional Neural Networks

Time Series Analysis

- Overview of Time Series Analysis
- Box-Jenkins Methodology
- Autoregressive (AR) Models
- Moving Average (MA) Models
- Building and evaluating ARIMA models

MapReduce and Hadoop

- MapReduce Steps
- The concept of Apache Hadoop
- Analytics for Unstructured Data
- The Hadoop Ecosystem

Deep Graph Learning

- Basic Concepts of a Graph
- Basic Measures of a Graph
- Fundamentals of GNN
- Fundamentals of Deep learning in Graphs

Prepare for the final exam

- The final exam will
  - Focus on concepts, principles, basic algorithms, and basic applications (explanation & description)
  - Have no programming tasks
  - Have some calculations

Sample questions

- The final exam generally consists of short-answer questions and long-answer questions. Here are some example questions.
- **Example** short-answer questions:
  1. Explain what “analytic sandbox” is.
  2. Explain what “hypothesis testing” is.
  3. Describe the concepts of “support” and “confidence” in association analysis

Sample questions

- **Example** long-answer questions (I):
  - In big data analytics, clustering analysis plays an important role. Please answer:
    1. Describe the purpose of clustering and when it shall be used
    2. Describe the following two clustering algorithms and their differences: K-means clustering and agglomerative hierarchical clustering
    3. As data scientist, what shall you pay attention to before and after applying a clustering algorithm to your data? Please discuss.

Additional Information

- Closed-book exam.
- 3 hours duration.
- Calculator without programming function is allowed. But strictly prohibit using calculator apps in mobile phones

Try Hard

&

Good Luck