NOVA College-wide Course Content Summary

ITD 245 – Advanced Applied Data Science Techniques (3 cr.)

Course Description

This course provides a broad survey of Big Data and data analytics, including demos and applications of widely used tools and methods. Topics include descriptive statistics, basic data analysis, common data extraction/translation/loading methods and tools from varied data sources and types, data visualizations, as well as machine learning (supervised and unsupervised). This course includes theory and practice, heavily emphasizing practical applications through case studies. Lecture 3 hours per week.

General Course Purpose

To prepare the student to derive meaningful and expressive information from a multitude of raw data sources, including the application of basic statistics, analysis tools and techniques, data extraction and cleaning, creation of visualizations, as well as the application of machine learning to analysis problems.

Course Prerequisites/Corequisites

Prerequisite: ITD 145 - Intro to Applied Data Science Techniques
Recommended: ITP 150 - Python Programming (or Python experience)

Course Objectives

A. Describe and use basic statistics on data.
B. Describe and work with datasets from a multitude of sources and formats.
C. Describe and manipulate datasets that are within the definition of "Big Data."
D. Extract, transfer and clean up data from raw data sources, transforming them into usable forms.
E. Describe and generate various visualizations from raw and derived data.
F. Describe and use supervised and unsupervised machine learning.
G. Define and apply feature engineering techniques in the process of developing machine learning models.

Major Topics to be Included

A. Basic descriptive statistics
B. Statistical distributions
C. Data manipulation and cleaning/'wrangling'
D. Big Data theory/applications; extraction and manipulation tools; ETL/ELT
E. Data visualization
F. Machine learning, supervised and unsupervised
G. Computer Vision (CV) and Natural Language Processing (NLP)
H. Feature engineering

Student Learning Outcome

A. Explain the purpose of statistics and define:
    1. qualitative and quantitative variables
    2. continuous and discrete quantitative variables
B. Define, obtain and use a dataset
C. Define and examine distributions, including Gaussian/normal distributions
D. Measuring Central Tendency
    1. Define and calculate the mean, median and mode
E. Measuring Dispersion
    1. Define and calculate range

2. Define and assess skew
3. Define and calculate variability, outliers, variance (σ^2)
4. Define and calculate standard deviation (σ)

F. Define, explain and calculate correlations
G. Define and classify independent and dependent variables
H. Explain the purpose of and create visualization plots
I. Define a random variable and explain random variable distributions
J. Extract/Translate/Load (ETL and ELT); Data Wrangling; Basic Analysis
    1. Define and explain the process of extraction, translation and loading (ETL)
    2. Use a relational database and SQL to perform basic ETL
    3. Extract data from multiple sources and formats, including CSV, JSON, XML, Web APIs, SQL and NoSQL database systems
    4. Define, explain and apply methods of data 'wrangling' and cleaning
    5. Apply basic tools to perform ETL, data wrangling/cleaning as well as analysis on 'cleaned' datasets
    6. Describe and explain the alternative process of ELT, involving data lakes
K. Define and explain the purpose of machine learning (ML)
L. Define and apply basic feature engineering
    1. Define imputation and apply basic imputation techniques
    2. Define and classify nominal and ordinal attributes
M. Define, explain and apply basic machine learning approaches, including regression, classification, clustering, etc.
N. Supervised machine learning
    1. Define and explain the purpose of supervised learning
    2. Examine supervised learning algorithms and identify appropriate applications
    3. Define and apply regression as a supervised learning prediction task
    4. Define classification and identify appropriate applications of classification
    5. Define and apply various classification algorithms, including, e.g., decision trees, k-nearest neighbors, logistic regression, random forests, neural networks, etc.
    6. Define and examine bias, variance, bias-variance tradeoff, overfitting, underfitting
    7. Define and explain the purpose of hyperparameters
    8. Define, explain and apply both traditional and deep neural networks (deep learning)
    9. Use advanced techniques in computer vision (CV) and natural language processing (NLP), such as tokenization, vector embeddings, CNNs, RNNs, and transformers, et al.
    10. Apply supervised learning algorithms to analyze and solve real world problems through case studies
O. Unsupervised machine learning
    1. Define, explain and apply unsupervised learning
    2. Define, explain and apply clustering methods, e.g., k-means, DBSCAN, etc.
    3. Define, explain and apply dimensionality reduction, e.g. PCA
    4. Demonstrate when dimensionality reduction is appropriate
    5. Apply unsupervised learning algorithms to analyze and solve real world problems through case studies
P. Apply statistics, Python and GUI tools, as well as ML theory and applications to analyze real world problems through case studies

Required Time Allocation per Topic

To standardize the core topics of this course, the following student contact hours per topic are required. Each syllabus should be created to adhere as closely as possible to these allocations. Topics are not necessarily to be taught in the order shown.

There are normally 45 student contact-hours per semester for a three-credit course (14 weeks of instruction, excluding final exam week: 14*3.2 = 45 hours). Sections of the course offered in alternative formats (i.e., not standard 15-week) still meet for the same number of contact hours. The final exam is not included in the timetable.

The quickly evolving nature of data analytics means that some content noted in this document may be superseded or made obsolete. As such, it is important to include such changes in individual syllabi. Additionally, time is allocated for additional and optional topics in order to provide instructors flexibility in tailoring the course to special needs or resources.

Topics	**Hours**	**Percentage**
Basic descriptive statistics	3	6.67%
Statistical distributions	2	4.44%
Data manipulation and cleaning/'wrangling'	6	13.33%
Big data theory/applications; extraction and manipulation tools; ETL/ELT	6	13.33%
Data visualization	5	11.11%
Machine learning, supervised and unsupervised	6	13.33%
Computer Vision (CV) and Natural Language Processing (NLP)	5	11.11%
Feature engineering	5	11.11%
Testing to include quizzes, tests and exams (excluding final exam)	3	6.67%
Other optional topics	4	8.89%
Total	**45**	**100%**