Appendix

Computer Science 6981

Data Preparation Techniques

Department of Computer Science

Instructor: Dr. Amilcar Soares

Email: amilcarsj@mun.ca **(please use this email to contact your instructor, not D2L)**

Course information (e.g., lecture notes, announcements, grades, etc) can be found on Brightspace (D2L). However, please note that your instructor **DOES NOT check email in Brightspace (D2L)**. The course number must be included in the subject line of any emails to the instructor or the instructional staff.

Current Course Prerequisites/Credit Restrictions:

Basic knowledge of Statistics and programming in Python 3 are required skills.

Course Description:

Students will learn several data preparation techniques for preprocessing your data set for data analytics tasks such as data mining, machine learning and data visualization. The course includes data cleaning, scaling, normalizing, discretizing, and imputing data and feature engineering, feature selection, and dimensionality reduction. The course will also include how to scale up the processing with distributed frameworks such as Apache Spark to handle large datasets. Finally, the students will see a high-level overview of some traditional data mining algorithms, such as linear regression and classification, decision trees, k-means and DBSCAN that will be eventually introduced to evaluate the impact of the techniques being taught.

List of topics

- iPython, Jupyter notebooks, and NumPy basics.
- Pandas (mapping, sorting and ranking, and descriptive statistics), Matplotlib
- Data cleaning
    - Why to clean your data?
    - Identify values for cleaning, formatting, finding outliers and duplicates.
- Data scaling, normalization, and discretization
    - Min-max scaler, standard scaler, max abs scaler, robust scaler, quantile transformer scaler, power transformer scaler, unit vector scaler.
    - Range, clipping, log and z-score normalization.
    - Equal width discretization and equal-frequency discretization.
    - Binning histogram and correlation analysis for data discretization.
- Scikit-learn basics, Supervised Learning (Bayesian, k-Nearest neighbors, Decision trees, Linear models)
    - Basics of the scikit-learn package, how to prepare your data, load and execute models.
    - Using basic models such as Bayesian, kNearest neighbors, Decision trees, Linear models.
    - How cleaning, scaling, normalization and discretization affects supervised learning?

- Scikit-learn, Unsupervised Learning (Kmeans and DB-SCAN)
    - How cleaning, scaling, normalization and discretization affects unsupervised learning.
- Scikit-learn, Dimensionality reduction (PCA and TSNe)
    - How cleaning, scaling, normalization and discretization affects dimensionality reduction.
- Scikit-learn, Feature selection
    - Statistics for filter feature selection method, correlation statistics, selection method, and transform variables.
- Data integration and encodings
    - A data integration primer. How to combine data sets with join, merge and concatenation.
    - One-hot encoding.
- Map Reduce
    - Scaling up the data analysis with the Map Reduce framework. Apache spark basics and examples

Course Objectives:

To give students basic knowledge on how preprocess raw data. We will cover both processing techniques for small and big data. While advancing in the content, some simple models would be introduced in the course for evaluating the impact of the addition of the preprocessing techniques.

At the end of this course the students will be able to:

- Perform data pre-processing in small and large data sets
- Evaluate the effect of pre-processing techniques using data mining/machine learning methods
- Scale the up pre-processing of large datasets using distributed frameworks

Recommended Readings

1. Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

2. Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Evaluation:

The final grade in this course will be determined by three project iterations as follows:

Quizzes (6)	30%
Assignments (3)	40%
Final Exam	30%

Assignments and Quizzes are individual. You must pass the final exam to pass the course.