STAT2450, Introduction to Data Mining with R

1 Calendar Description

This course provides an introduction to data mining and R programming, suited for science students. Data mining methods include a vast set of tools developed in different areas for identifying the patterns in data. Students will learn programming methods for manipulating and exploring data through learning the basic ideas of some clustering, regression and classification methods. No prior programming knowledge is assumed.

2 Prequisites

The course is open to anyone who has successfully completed MATH 1000 and either STAT/MATH 1060 or STAT/MATH 2060.

3 Course Description

The primary aim of this course offering is to attract lower-year students intending to major in statistics, mathematics, computer science, neuroscience, biology, and psychology, into the areas of machine learning, data science, and artificial intelligence. There is a lot of interest revolving around the data science with applications in many disciplines; one goal of this course is to provide an accessible path for younger students to study these fields closely and become acquainted with basic data analysis skills and software.

The second goal of this course is to teach students R programming and general scientific computing skills. A course covering core statistical programming and data analytics skills for non-technical students is lacking from the course selection at Dalhousie. There is a demonstrated need in various departments at Dalhousie including psychology and biology to provide non-technical students with the opportunity to obtain these useful data analytics skills so that they can apply them in their respective fields effectively. This course fills the gap by teaching data science skills through an R programming environment, motivated by basic concepts in statistical learning.

Throughout the course, we make use of many clever visual demos to deliver information so that the basic ideas are comprehensible. The focus of the course will be to teach students where and how these basic techniques can be applied and how to interpret the results, with as little as possible technical details behind the methods. Additionally, practical applications will be emphasized throughout

lectures with focuses on health science and financial applications, attempting to highlight as many real-world scenarios as possible.

4 Evaluation

6 – 8 assignments (65%), Final exam (35%)

5 Textbook

An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 2013; Publisher: Springer. (Available freely online, in Dalhousie library, and for purchase in the Dalhousie Book Store).

6 Topics

This course goes through the main ideas and basic algorithms of statistical learning through the R programming language. We will learn data visualization techniques, and core data analytics skills through interactive exercises, which can be applied in a variety of fields. Some topics listed below will be covered:

1. Introduction to Statistical Learning: Explain the problems of Classification, Clustering, Regression, Dimensionality Reduction; Describe supervised, and unsupervised learning paradigms; Describe the purpose of a feature vector representation.
2. Supervised Learning
   a) Model Selection: Describe the bias-variance trade-off; Discuss Occams Razor in context of hypothesis selection; Explain the problems of overfitting and underfitting.
   b) K-nearest Neighbours: Describe the K-nearest Neighbours algorithm for classification and regression; Apply weighted K-nearest Neighbours to datasets; Explain the issues of noisy observations in data.
   c) CART-based Decision Trees: Describe how decision tree is constructed through CART; Describe the effect of adjusting tree hyperparameters; Describe how CART is used for regression and classification.
   d) Support Vector Machines: Explain how the SVM finds the optimal hyperplane; Describe the difference between Black-box vs. White-box learning algorithms; Describe how to use kernel methods to solve non-linear problems using linear methods; Describe how to introduce a soft-margin with the cost hyperparameter.
   e) Artificial Neural Networks: Explain the single-layer perceptron learning algorithm and the perceptron learning rule; Describe how the multilayer perceptron computes output values; Discuss problems associated

with training neural networks; Explain the use of momentum and early-stopping in neural networks; Explain the differences between online learning and offline learning approaches.

f) Other in Supervised Learning: Compare and contrast the pros and cons of varying learning algorithms; Describe the implications of the No Free Lunch theorem in the context of machine learning; Describe the difference between interpolation and extrapolation; Describe the purpose of learning curves; Discuss the purpose of regularization in machine learning algorithms; Describe how we can extend binary classifiers with One-vs-all and one-vs-one techniques.

3. Unsupervised Learning: Apply DBSCAN for density-based clustering; Apply K-means for centroidbased clustering; Describe inertia and silhouette as cluster quality evaluation metrics; Describe the elbow method for choosing the optimal K-value for K-means clustering; Describe the difference between agglomerative and divisive approaches for hierarchical clustering.
4. Data Preparation: Explain when and why we should apply normalization, standardization, or minmax scaling; Discuss the importance of feature engineering in machine learning; Describe forward selection and backward selection for feature selection; Describe some problems associated with the curse of dimensionality; Discuss simple data imputation techniques; Transforming categorical input features.