Basic Data Mining Sessions

Module 1. Introduction to the Course and Introduction to Data Mining

The learning outcomes of this module are:

- Describe the difference between analytics and analysis and identify viable and profitable business problems for data analytics.
- Apply knowledge of the different application areas of analytics to develop analytics approaches more effectively in the organization.
- Identify common challenges facing the use of analytics to overcome such challenges in data mining projects.
- Understand when to apply descriptive, predictive, or prescriptive analytics

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+1+-+Introduction+to+DataMining-NJIT/1_jycr82py

Module 2. Introduction to Data Mining (continue) and Introduction to Predictive Modeling

Describe the evolution of data mining and the power and applicability of contemporary data mining approaches to organizational business problems.

- Utilize knowledge of the most common data mining application areas to select appropriate data mining tools, techniques, and methodologies for various projects.
- Apply knowledge from various disciplines to handle data analytics tasks more effectively.
- Understand and use the patterns that data mining can discover e.g., associations, classifications, and clustering, and use them more effectively.
- Avoid common traps in data mining.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+2+-+Introduction+to+Predictive+Modeling-NJIT/1_wwdj46po

Module 3. The Data Mining Process, in particular, CRISP-DM

Use the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to carry out data mining projects. The six phases in CRISP-DM are: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+3+-+CRISP-DM+-+Overview+-+NJIT/1_8nmz622e

Module 4. Supervised Segmentation, example Decision Tree

This module delves into one of the main topics of data mining: predictive modeling. We will begin by thinking of predictive modeling as *supervised* segmentation—how can we segment the population into groups that differ from each other with respect to some quantity of interest. In particular, how can we segment the population with respect to something that we would like to predict or estimate. The target of this prediction can be something we would like to avoid, such as which customers are likely to leave the company when their contracts expire, which accounts have been defrauded, which potential

customers are likely not to pay off their account balances (*write-offs*, such as defaulting on one's phone bill or credit card balance), or which web pages contain objectionable content. The target might instead be cast in a positive light, such as which consumers are most likely to respond to an advertisement or special offer, or which web pages are most appropriate for a search query.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+4+-+Supervised+Segmentation/1_m15zbx0y

Module 5. Discriminant Functions

This module specifies the structure of the model with certain numeric parameters left unspecified. Then the data mining calculates the best parameter values given a particular set of training data. A very common case is where the structure of the model is a parameterized mathematical function or equation of a set of numeric attributes. The attributes used in the model could be chosen based on domain knowledge regarding which attributes ought to be informative in predicting the target variable, or they could be chosen based on other data mining techniques, such as the attribute selection procedures. The data miner specifies the form of the model and the attributes; the goal of the data mining is to tune the parameters so that the model fits the data as well as possible. This general approach is called *parameter learning* or *parametric modeling*.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+5+-+Discriminant+Functions-New/1_f2w2vv1i

Module 6. Model Performance Analysis

One of the most important fundamental notions of data science is that of overfitting and generalization. If we allow ourselves enough flexibility in searching for patterns in a particular dataset, we will find patterns. Unfortunately, these "patterns" may be just chance occurrences in the data. As discussed previously, we are interested in patterns that generalize—that predict well for instances that we have not yet observed. Finding chance occurrences in data that look like interesting patterns, but which do not generalize, is called *overfitting* the data.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+6+-+Model+Performance+Analytics-New/1_a9kkufd8

Module 7. Model Performance Evaluation Metrics

What is a good model? For data science to add value to an application, it is important for the data scientists and other stakeholders to consider carefully what they would like to achieve by mining data. Both data scientists themselves and the people who work with them often avoid—perhaps without even realizing it—connecting the results of mining data back to the goal of the undertaking. This may manifest itself in the reporting of a statistic without a clear understanding of why it is the right statistic, or in the failure to figure out how to measure performance in a meaningful way.
Often it is not possible to perfectly measure one's goal, for example because the systems are inadequate, or because it is too costly to gather the right data, or because it is difficult to assess causality. So, we might conclude that we need to measure some surrogate for what we would really like to measure. It is nonetheless crucial to think carefully about what we would really like to measure. If we must choose a surrogate, we should do it via careful, data-analytic thinking.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+7+-+ModelPerformanceEvaluationMetrics/1_ojx6rg5p

NJ ACTS Machine Learning with Python Professional Training

Advanced Data Mining Sessions

Module 8. Support Vector Machine In-Depth

In this module, we will describe one of the most widely used machine learning methods, i.e., Support Vector Machine (SVM). SVM is an approach for classification that was developed in the computer science community, and it has been shown to perform well in a variety of settings and is often considered one of the best "out of the box" classifiers. In addition, Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM). SVR performs regression-based prediction. We focus on the classification method, i.e., SVM.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+8+-+Support-Vector-Machine/1_a8fu2ev2

Module 9. Prediction via Evidence Combination

Let's now examine a different way of looking at drawing such conclusions. We could think about the things that we know about a data instance as *evidence* for or against different values for the target. The things that we know about the data instance are represented as the features of the instance. If we knew the strength of the evidence given by each feature, we could apply principled methods for combining evidence probabilistically to reach a conclusion as to the value for the target. We will determine the strength of any particular piece of evidence from the training data.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+9+-+Prediction+via+Evidence+Combination/1_2atp92cj

Module 10. Representing and Mining Text: Text and Sentiment Analysis

Data are represented in ways natural to problems from which they were derived. If we want to apply the many data mining tools that we have at our disposal, we must either engineer the data representation to match the tools or build new tools to match the data. Top-notch data scientists employ both strategies. It generally is simpler to first try to engineer the data to match existing tools, since they are well understood and numerous. In this module, we will focus on one sort of data that has become extremely common as the Internet has become a ubiquitous channel of communication: text data. Examining text data allows us to illustrate many real complexities of data engineering and helps us to better understand a very important type of data.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+10-+Representing+and+Mining+Text/1_8z39j6dg

Module 11. Unsupervised Learning Algorithms

Similarity underlies many data science methods and solutions to business problems. If two things (people, companies, products) are similar in some ways they often share other characteristics as well. Data mining procedures often are based on grouping things by similarity or searching for the "right" sort of similarity. We saw this implicitly in previous chapters where modeling procedures create boundaries for grouping instances together that have similar values for their target variables. In this module, we will look at similarity directly, and show how it applies to a variety of different tasks. We may want to group similar items together into *clusters*, for example

to see whether our customer base contains groups of similar customers and what these groups have in common. Previously we discussed supervised segmentation; this is unsupervised segmentation. After discussing the use of similarity for classification, we will discuss its use for clustering.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+11+-+Unsupervised+Data+Mining+and+Clustering/1_29zsu16i

Module 12. Deep Learning

Deep Learning (DNN) is a branch of machine learning methods and originates from artificial neural networks (ANN). Given the advancements in computer hardware, algorithms, and big data abundance, deep learning has gained rapid developments since the early 2000s. It brings new approaches, new methodologies into bioinformatics and provides a new angle to approach challenging bioinformatics problems. Deep learning has many architectures, such as Deep Multilayer perceptron's (DMLP), deep belief networks (DBN), graph neural networks (GNN), recurrent neural networks (RNN), and convolutional neural networks (CNN). They bring state-of-the-art performance to biomedical research, clinical patient care, electronic health record (EHR) analysis. Sometimes, many think that it is a "magic bullet" for solving any challenging problem. In the module, we demystify deep learning and uncover the fundamental connection between Machine Learning and Deep Learning and explain deep learning with the terminologies introduced in the previous modules.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+12+-+Deep-Learning.mp4/1_yd6vp7o7

Module 13. Causal Inference

In this module, we will use Machine Learning to predict the causal effects and introduce key concepts in machine learning-based causal inference. There will be a specific highlight on the methods for estimating causal effects in historical data to estimate the impact of treatments.

*Pre-recorded Lesson: https://rutgers.mediaspace.kaltura.com/media/Module+13-+Causal-Inference_1920x1440/1_pftececu