Cyber Data Analytics (CS4035)

Checklist “Do I meet the prerequisites?”

NOTE: You should be able to complete the “Applied Machine Learning Assignment” shown on the next page (approx. 6 hours).

For this course, you are expected to have a **working knowledge** of the following concepts. For each concept, a reference to a specific section explaining the concept in “An Introduction to Statistical Learning” [JWHT] is given.

1. Machine Learning basics [JWHT, Ch. 2]
2. Simple Linear Regression [JWHT, Ch. 3.1]
3. Logistic Regression and LDA [JWHT, Ch. 4]
4. Cross-validation [JWHT, Ch. 5.1]
5. Decision and Regression Trees [JWHT, Ch. 8]
6. Support Vector Machines [JWHT, Ch. 9.1 – 9.3]
7. PCA and Clustering [JWHT Ch. 10]

With working knowledge, we mean that you are able to apply these methods correctly to a dataset, including the required preprocessing steps. The book contains several R scripts that show you how to do this. You may also want to check the scikitlearn Python package: http://scikit-learn.org/stable/index.html, especially the examples.

[JWHT] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013. Available at: http://www-bcf.usc.edu/%7Egareth/ISL/ISLR%20Seventh%20Printing.pdf

Applied Machine Learning Assignment

Expected time required: 6 hours

1. Run, study, and understand the R code samples that come with the [JWHT] book.

2. Download the 1999 KDD Cup dataset from UCI: https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data

3. Preprocess the data using a one-hot encoding to model categorical variables (see, e.g., https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/)

4. Using your favorite machine learning framework (R, Python, Knime, Azure, …), learn the following classifiers from this dataset:

    a. Naïve Bayes
    b. LDA
    c. Decision Tree
    d. SVM

5. Compare their performance using 10-fold cross-validation

6. Repeat after first reducing the dimensionality by normalizing the data and applying PCA

7. Provide a contingency table of the results and conclude which classifier works best