Business Analytics – Homework 6

Due date: Please check LATTE

This assignment consists of 4 exercises the last two **being completely optional**. It has been designed to make sure you master manipulation of dates, strings and data frames in R. As always, you will be asked to write short code snippets or code-chunks in R. You will be graded both on your code, and the written answers you provide. When evaluating the code, the grader will take on the role of a co-worker. Code will be evaluated both in terms of how correct and how clear it is. By correctness, I mean that the code fulfills the requirements of the question. By clarity, I mean that the grader should be able to understand what your code does within 30 seconds of reading it. As discussed in class, this is aided by clear comments, good variable names, proper indentation, and short lines.

Problem 1: Comparing the performance of different classifiers

This exercise is intended for you to understand the importance of the no free lunch theorem. The no free lunch theorem states that

1. In class we have seen how to train
    a. KNN
    b. Naive Bayes
    c. Linear Regression
    d. Regression Trees
    e. Neural Nets

2. Evaluate the predictive performance of each of these methods using confusion matrices seen in class.
    a. Tumor data
    b. Program application
    c. Titanic data set

3. Which predictive model does best? Is any of the remarkably better?

Problem 2: Basic Sentiment Analysis

In machine learning, classification is usually regarded as the problem of predicting a categorical variable Y using as input vectors of explanatory variables (these explanatory variables are sometimes referred to as features, X's, covariates).

Classifiers are extremely useful for companies in the big data era. One of the most prominent uses of classification is sentiment analysis. Companies need to respond fast to phishing attacks, bad reviews and so on.

In this exercise I ask you to build a dictionary based classifier to detect whether a given text is positive or negative or neutral using as features patterns of positive and negative words. You have two data sets named `data_dictionary_tone.csv` and `data_reviews_classified.csv`.

Construct a classifier for the tone of any sentence.