Tutorial 5 Introduction to Machine Learning

Overview

Statistics, statistical learning, machine learning and data science are overlapping terms that seem to have definitions that vary with time and space. For our purposes, we can focus on two of these terms:

Statistics: the branch of mathematics that deals with data organization, presentation, analysis and interpretation. With this broad definition, it certainly covers the other terms! In particular, we refer to "**descriptive statistics**" when we characterize a dataset by its mean, standard deviation, correlation coefficients etc, and when we move into plotting the data, we can also apply the trendy term "data visualization". In "**statistical inference**" we are attempting to extract some higher truth from the numbers, such as the likelihood that two means are different, or the probability that a bolt size is 6mm when the measurement says 5.5 mm (see Tutorial 3). "**Data modelling**" typically refers to finding a simplified representation of the data to be used for interpolation or prediction; "regression" is the prime example.

Machine Learning: defined on Wikipedia as the scientific study of algorithms and statistical models that computer systems use to perform a task without using explicit instruction. Other people insist that ML implies that the computer is doing a task T with measurable performance level that improves with more experience E. This definition seems to apply well to **"supervised learning"** (regression and classification), but not to **"unsupervised learning"** such as cluster analysis which is intended to find patterns in complex data sets.

If we applied a Naïve Bayes algorithm to identify "Machine Learning", it would probably be strongly influenced by the involvement of computers, very large complex data sets, training, and complex algorthms. ML is certainly built on the principles of "traditional" statistical analysis, so the first 10 weeks of the course were essential preparation for the activities in this tutorial.

In this tutorial you will build a predictive model of wine quality (as assessed by wine expert tasters) using objective chemical and physical characteristics of wine. Part of your grade will depend on the performance of your model!

Objectives:

a) Explore a large dataset using a few "traditional" tools to identify some key features.

b) Use Naïve Bayes and several classification tools to develop a predictive model for one of the variables in the dataset (in this case, wine quality).

c) Test the predictions on a portion of the data that was withheld from the training.

Pre-tutorial questions:

There are no text questions associated with this tutorial, but there is some recommended reading.

Download the text An Introduction to Statistical Learning with Applications in R, but James, Witten, Hastie and Tibshirani, Springer, 8th printing 2017. Read Chapter 2, with particular attention to section 2.2.1 and 2.2.3. Read section 4.1, 5.1 (especially 5.1.3); Ch. 8 focus on 8.1

Background

Wine quality and price are controversial subjects: you can pay a lot for a poor wine. Even experts can be influenced by the presentation of the wine (wineglass vs paper cup). It would be nice if we could determine the wine quality class using objective measurements. This may be impossible, but we will try!

The dataset we will work with provided on the UC Irvine data repository and the subject of a paper by Cortez et al (2009).[1] The data relate to wines from northern Portugal, of white and red varieties. The governing wine organization there tests samples of the wines for physical and chemical characteristics. In addition, the quality of each sample is rated (on an integer scale from 1-10, 10 being best). You will be given 2/3 of the red wine data, with which you will build the best classifier model you can.

The performance of the model will be assessed using the portion of the data that you did NOT have for model training. The model predictions will be classified as “perfect” (correct category), “near” (one category off) and “bad” (more than one category off).

The overall score is $S=(2\,N_{\text{perfect}}+N_{\text{near}}-2\,N_{\text{bad}})/N$ where $N=N_{\text{perfect}}+N_{\text{near}}+N_{\text{bad}}$.

The variables in the dataset are:

Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):
12 - quality (score between 0 and 10)

[1] https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub

Tutorial Assignment

The questions and file uploads will be on Canvas, but the assignments are explained below.

1. Complete the quiz on Canvas related to the readings assigned above.
2. Load the training data trainRW.txt into matlab and use corrplot to identify the most likely predictors of wine quality (variable 12).
3. Using the Matlab Classification Learner App, build 10 model variants (including 3-6 model types, and 2-3 validation options). Importantly, build models using only 5 of the 11 predictors.
4. List the model variants and their accuracy, and also examine the confusion matrix for each one.
5. Select the model that you think will perform the best on the training data. As requested on Canvas, explain why you chose this model, what the performance is (including a confusion matrix for the training set), and
6. Use the second file "validRWpredictors.csv" to create a vector of predicted classes. The vector you upload must be a list with numbers (0-10) after the first row. The exact format is specified in the script Assignment_RedWine.m. Our plan is to take these predictions and compare with the correct results.
7. Submit the assignment above by Monday, **midnight**, April 1.