MATH5743M: Statistical Learning - Assessed Practicals

Dr Seppo Virtanen, School of Mathematics, University of Leeds

Semester 2: 2022

Assessed Practical III: Can I eat that mushroom?

Assignment due: 11:59pm Friday 12 May

In this practical we are going to investigate one of the all-time classic machine learning data sets. What is a classic machine learning data set? When researchers create new methods they typically test their performance on data sets people have looked at before, so that the prediction accuracy can be benchmarked against existing methods. This means that certain data sets appear time and time again in research. One of these is the famous 'mushroom data set': a set of observations about different specimens of gilled mushrooms in The Audubon Society Field Guide to North American Mushrooms (1981). Each specimen is measured in terms of some visual and olfactory information, such as its Cap Size and its Odor type. They are also labelled as being edible or poisonous. Our goal is to determine whether a mushroom is edible from its characteristics.

Figure 1: An example of gilled mushrooms. A classic machine learning task is to determine whether or not a particular mushroom is poisonous based on its visual and olfactory characteristics

Download the data from MINERVA: mushrooms.csv

There are 6 attributes in the data, all of which are factors (non-numeric categorical variables). These are: Edible (to be predicted), CapShape, CapSurface, CapColor, Odor and Height.

Tasks

1. Consider logistic regression, decision trees and random forests for predicting ediblity based on all or any subset of the remaining attributes. Focus on tuning each method for maximal predictive performance (see the help functions of the methods). Use the number of mushrooms correctly classified as the criterion for deciding which model is best. Using cross-validation, perform a model selection to determine which model performs best. Explain your approach, and present your results in the most convincing way you can. Imagine that you are reporting your results for a scientific publication. You may use statistical tests to determine if the performance between the methods is statistically significant. Your report is assessed based on the scope, detail and validity of your analysis.