Assessed Practical: Can I eat that mushroom?

Assignment due: 12pm Friday May 10

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. Fit Random Forest models using each possible input on its own to predict edibility. Evaluate the quality of fit by using the **predict** function to calculate the predicted class for each mushroom (edible or poisonous) (hint, you need type='response'). Which input fits best? (i.e. which classifies the most mushrooms correctly?)
2. Using cross-validation, perform a model selection to determine which features are useful for making predictions using a Random Forest. As above, use the number of mushrooms correctly classified as the criterion for deciding which model is best. You might try to find a way to loop over all 32 possible models (ignore the possibility of no input variables. Hint: you can use **allCombs** in the **dagR** package to generate all combinations of the numbers 1 to n). Or select features 'greedily', by picking one at a time to add to the model. Present your results in the most convincing way you can.
3. Would you use this classifier if you were foraging for mushrooms? Discuss with reference to factors that you identified as important and the probability of posioning yourself.