INF283 - Compulsory Assignment 2

Submission deadline: Tuesday, November 14, 2017

Updated: 2017-11-04

Classifying mushrooms as poisonous or edible

In this assignment, you will implement a decision tree algorithm for identifying mushrooms as poisonous or edible.

The data set

The data set is a table (CSV file, `agaricus-lepiota.data`) of mushrooms, where the first column is the class, 'e' for edible and 'p' for poisonous. The other columns represent properties (features) that can be used for deciding.

Table 1: Feature name and legal values for each column in the data set.

1	cap-shape	bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s
2	cap-surface	fibrous=f,grooves=g,scaly=y,smooth=s
3	cap-color	brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p, purple=u,red=e,white=w,yellow=y
4	bruises?	bruises=t,no=f
5	gill-attachment	attached=a,descending=d,free=f,notched=n
6	gill-spacing	close=c,crowded=w,distant=d
7	gill-size	broad=b,narrow=n
8	gill-color	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
9	stalk-shape	enlarging=e,tapering=t
10	stalk-root	bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r
11	stalk-surface-above-ring	fibrous=f,scaly=y,silky=k,smooth=s
12	stalk-surface-below-ring	fibrous=f,scaly=y,silky=k,smooth=s
13	stalk-color-above-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
14	stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15	veil-type	partial=p,universal=u
16	veil-color	brown=n,orange=o,white=w,yellow=y
17	ring-number	none=n,one=o,two=t
18	ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
19	spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
20	population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
21	habitat	grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

Note that some values are missing (marked with a question mark)

Tasks

1. Calculate entropy gain for each feature

Modify the code from Marsland (p253) to deal with missing values, and calculate Information Gain for each of the features in the data set, and producde the results in a table.

2. ID3 Decision Tree

Decide on a suitable training and validation scheme (and describe it), and use the code from Marsland (p255 and p256, modifying as necessary) to implement a decision tree for this data set. What is your estimated accuracy of your classfier? Does it overfit? Also calculate precision and recall for the classifier.

3. Accuracy vs other measures

Classifying an edible mushroom as poisonous is a less grave error than classifying a poisonous mushroom as edible. How would you take this into account when constructing the decision tree?

4. Regularization

ID3 builds the tree until it runs out of features or the set of data is unanimous. Marsland mentions three methods for constructing trees that are shallower, and thus less likely to overfit the data: early stopping, post-pruning, or rule-based pruning.

Decide on a regularization scheme, and implement it. What is the size of the smallest tree that can classify mushrooms with the same classification performance? What if you allow for (somewhat) poorer performance?

Submission

Please submit your paper answering the questions above, as well as your modified source code to MittUIB by Tuesday, November 14.