Abraao Lourenco
CS4445-A04
Homework #1
Due 9/9/2004

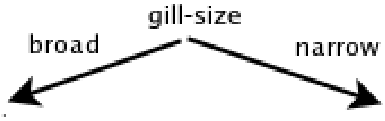
Homework #1

1. Constructing the ID3 tree.

Sample Mushrooms		**Edible**	**Poisonous**	**Total**
cap-surface				
fibrous	6/20 *	[- 4/6 * $\log_2$(4/6)	- 2/6 * $\log_2$(2/6)]	0.275489
grooves	0/20 *	[- 0	- 0]	0
scaly	9/20 *	[- 4/9 * $\log_2$(4/9)	- 5/9 * $\log_2$(5/9)]	0.445984
smooth	5/20 *	[- 2/5 * $\log_2$(2/5)	- 3/5 * $\log_2$(3/5)]	0.242738
entropy total				**0.964211**
bruises?				
bruises	7/20 *	[- 5/7 * $\log_2$(5/7)	- 2/7 * $\log_2$(2/7)]	0.302092
no	13/20 *	[- 5/13 * $\log_2$(5/13)	- 8/13 * $\log_2$(8/13)]	0.624804
entropy total				**0.926896**
gill-size				
broad	13/20 *	[- 10/13 * $\log_2$(10/13)	- 3/13 * $\log_2$(3/13)]	0.506577
narrow	7/20 *	[- 0/7 * $\log_2$(0/7)	- 7/7 * $\log_2$(7/7)]	0
entropy total				**0.506577**
habitat				
grasses	5/20 *	[- 3/5 * $\log_2$(3/5)	- 2/5 * $\log_2$(2/5)]	0.242738
leaves	4/20 *	[- 2/4 * $\log_2$(2/4)	- 2/4 * $\log_2$(2/4)]	0.200000
meadows	0/20 *	[- 0	- 0]	0
paths	3/20 *	[- 1/3 * $\log_2$(1/3)	- 2/3 * $\log_2$(2/3)]	0.137744
urban	0/20 *	[- 0	- 0]	0
waste	1/20 *	[- 1/1 * $\log_2$(1/1)	- 0/1 * $\log_2$(0/1)]	0
woods	7/20 *	[- 3/7 * $\log_2$(3/7)	- 4/7 * $\log_2$(4/7)]	0.344830

entropy total				**0.925312**

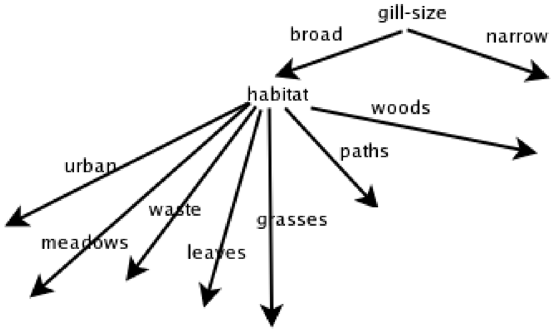
Since the attribute gill-size has the lowest entropy, the root of the ID3 tree will be gill-size:



Now for gill-size = broad we measure the entropy for the rest of the three attributes: cap-surface, bruises?, habitat

Sample Mushrooms		**Edible**	**Poisonous**	**Total**
cap-surface				
fibrous	6/13 *	[- 4/6 * $\log_2$(4/6)	- 2/6 * $\log_2$(2/6)]	0.423829
grooves	0/13 *	[- 0	- 0]	0
scaly	5/13 *	[- 4/5 * $\log_2$(4/5)	- 1/5 * $\log_2$(1/5)]	0.277665
smooth	2/13	[- 2/2 * $\log_2$(2/2)	- 0/2 * $\log_2$(0/2)]	0
entropy total				**0.701494**
bruises?				
bruises	5/13	[- 5/5 * $\log_2$(5/5)	- 0/5 * $\log_2$(0/5)]	0
no	8/13	[- 5/8 * $\log_2$(1/1)	- 3/8 * $\log_2$(3/8)]	0.587344
entropy total				**0.587344**
habitat				
grasses	3/13	[- 3/3 * $\log_2$(3/3)	- 0/3 * $\log_2$(0/3)]	0
leaves	2/13	[- 2/2 * $\log_2$(2/2)	- 0/2 * $\log_2$(0/2)]	0
meadows	0/13 *	[- 0	- 0]	0
paths	3/13	[- 1/3 * $\log_2$(1/3)	- 2/3 * $\log_2$(2/3)]	0.211914
urban	0 *	[- 0	- 0]	0
waste	1/13	[- 1/1 * $\log_2$(1/1)	- 0/1 * $\log_2$(0/1)]	0
woods	4/13	[- 3/4 * $\log_2$(3/4)	- 1/4 * $\log_2$(1/4)]	0.249624
entropy total				**0.461538**

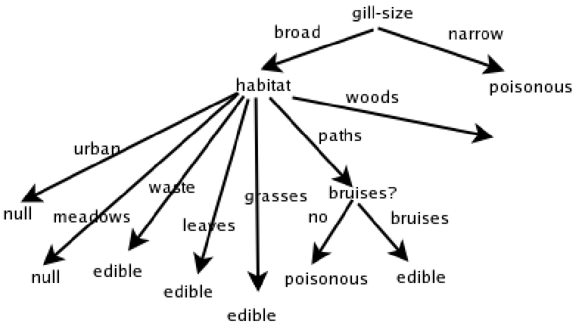
So the next node in the tree will be habitat.



In the branch grasses all examples have class edible; likewise for the branches leaves and waste. There are no instances containing meadows or urban, so that leaves paths and woods. Now for habitat = paths, we measure the entropy of the rest of the two attributes, cap-surface, bruises?

Sample Mushrooms		**Edible**	**Poisonous**	**Total**
cap-surface				
fibrous	2/3 *	[- 0/2 * $\log_2$(0/2)	- 2/2 * $\log_2$(2/2)]	0
grooves	0/3 *	[- 0	- 0]	0
scaly	1/3 *	[- 1/1 * $\log_2$(1/1)	- 1/1 * $\log_2$(1/1)]	0
smooth	0/3 *	[- 0	- 0]	0
entropy total				**0**
bruises?				
bruises	1/3	[- 1/1 * $\log_2$(1/1)	- 0/1 * $\log_2$(0/1)]	0
no	2/3	[- 0/2 * $\log_2$(0/2)	- 2/2 * $\log_2$(2/2)]	0
entropy total				**0**

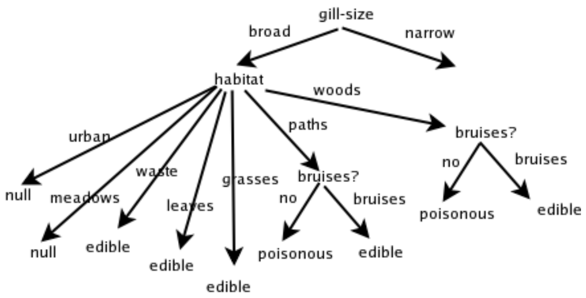
Since both cap-surface and bruises? have the same entropy value, 0, we can select either one for the next level of the tree. I'll select bruises?. Now for branch (gill-size = broad, habitat = paths, bruises? = bruises) all examples have class edible and for this branch the tree will have leaf edible. Now for branch (gill-size = broad, habitat = paths, bruises? = no) all examples have class poisonous and for this branch the tree will have leaf poisonous.



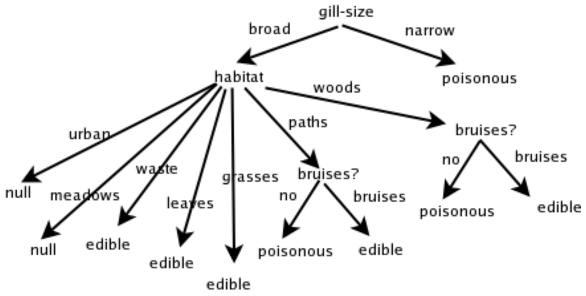
Now for habitat = woods, we measure the entropy of the rest of the two attributes, cap-surface, bruises?

Sample Mushrooms		**Edible**	**Poisonous**	**Total**
cap-surface				
fibrous	1/4 *	[- 1/1 * log$_2$(0/2)	- 0/1 * log$_2$(0/1)]	0
grooves	0/4 *	[- 0	- 0]	0
scaly	3/4 *	[- 2/3 * log$_2$(2/3)	- 1/3 * log$_2$(1/3)]	0.688722
smooth	0/4 *	[- 0	- 0]	0
entropy total				**0.688722**
bruises?				
bruises	1/3	[- 1/1 * log$_2$(1/1)	- 0/1 * log$_2$(0/1)]	0
no	2/3	[- 0/2 * log$_2$(0/2)	- 2/2 * log$_2$(2/2)]	0
entropy total				**0**

So the next node in the tree will be bruises?. Now for branch (gill-size = broad, habitat = woods, bruises? = bruises) all examples have class edible and for this branch the tree will have leaf edible. Now for branch (gill-size = broad, habitat = woods, bruises? = no) all examples have class poisonous and for this branch the tree will have leaf poisonous



Now for gill-size = narrow all examples have class poisonous and for this branch the tree will have leaf poisonous.



2. Computing the accuracy

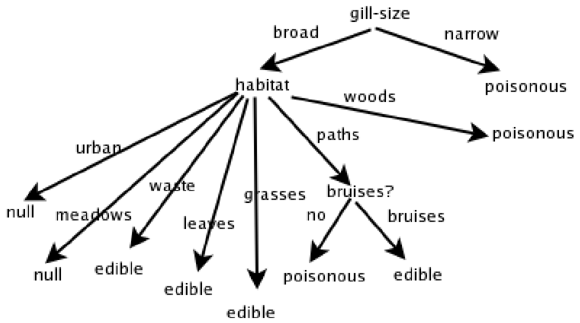
fibrous,no,broad,grasses,poisonous	Tree's prediction: edible (incorrect)
scaly,bruises,broad,grasses,edible	Tree's prediction: edible (correct)
scaly,no,broad,grasses,poisonous	Tree's prediction: edible (incorrect)
scaly,no,broad,paths,poisonous	Tree's prediction: poisonous (correct)
smooth,bruises,broad,grasses,edible	Tree's prediction: edible (correct)
smooth,bruises,broad,waste,edible	Tree's prediction: edible (correct)
smooth,no,broad,grasses,edible	Tree's prediction: edible (correct)
smooth,no,broad,leaves,edible	Tree's prediction: edible (correct)
smooth,no,narrow,leaves,poisonous	Tree's prediction: poisonous (correct)
smooth,no,narrow,paths,poisonous	Tree's prediction: poisonous (correct)

The accuracy of your decision tree on this test data is: 80%

3. Subtree replacement pruning technique

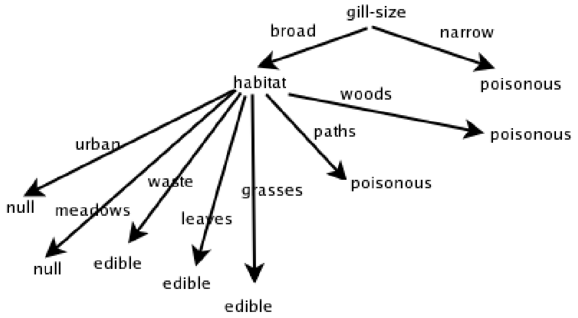
a) As the handout states, “pruning a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training examples affiliated with that node. Nodes are removed only if the resulting pruned tree performs no worse than the original over the testing set”. So the rightmost “bruises?” node is the first candidate for pruning.

b) Implementing this pruning technique and pruning the first candidate will give us the following tree:



The node will be pruned because the obtained tree performs no worse than the original tree on the testing data (i.e. the accuracy of the pruned decision tree is 80% as well).

c) After the rightmost "bruises?" node has been replaced, the next "bruises?" node is considered. Pruning this node and replacing it with the "poisonous" classification yields a



tree that performs no worse than the original tree:

The next node considered for pruning is habitat. There is no most common classification; there are ten examples classified edible and ten examples classified poisonous. Replacing habitat with the edible classification results in a tree results in a tree that has a 70% accuracy over the testing set (i.e. the first, third, and fourth instances in the training set are the only ones predicted incorrectly by the tree). Replacing habitat with the poisonous classification yields a tree that has a 50% accuracy (i.e. it predicts the second, fifth, sixth, seventh, and eight instances in the training set incorrectly). Since pruning habitat results in a tree with a worse performance, the final tree is the one displayed above this paragraph.