CSCI 6350/7350 **Name** _____

Homework 2 (due beginning of class, Wednesday Feb 8th)

(Please type your solution for the homework)

We have collected a data set of 14 data objects representing 14 different mushrooms. Each mushroom is labeled by domain expert whether it is edible or poisonous. We would like to learn a decision tree that will help us determine for any mushroom we may find in the future whether it is edible or poisonous. The three attributes chosen for describing the mushrooms, together with the possible values for each attribute are shown below:

- Cap Shape: bell, flat, or convex
- Cap Color: brown, grey
- Odor: almond, spicy, foul

Here is the data set:

Data:

Object	**Cap Shape**	**Cap color**	**Odor**	**class**
X1	bell	brown	almond	edible
X2	flat	grey	almond	edible
X3	convex	grey	spicy	poisonous
X4	bell	brown	almond	edible
X5	flat	grey	almond	edible
X6	flat	grey	spicy	edible
X7	convex	grey	almond	edible
X8	bell	brown	almond	edible
X9	convex	brown	foul	poisonous
X10	bell	brown	spicy	edible
X11	bell	grey	almond	edible
X12	convex	grey	spicy	poisonous
X13	flat	brown	almond	edible
X14	flat	grey	foul	poisonous

1. Use *information gain* as the attribute selection criterion to build a decision tree for the data.
   Show step-by-step computation involved in the attribute selection that results in a partial decision tree with 2 levels (root is the first level). Draw the partial decision tree and label each node in the 2nd level of tree with a class label.

2. Enter the data into the DecisionTree.py program and learn a complete decision tree from this data. (Use Information gain criterion, and no pruning)
   a. Draw (or copy and paste) the resulting tree in your answer.
   b. Classify the following two mushrooms using the decision tree learned from the previous step:

   Data:

Object	**Cap Shape**	**Cap color**	**Odor**	**class**
q1	bell	grey	almond	??
q2	convex	grey	spicy	??
q2	flat	brown	foul	??

3. Perform chi square test to determine if the first level split performed in question 1 is statistically significant? (ρ=0.05) Should it be pruned away? Show all the computations involved.

4. Describe how to compute the classification accuracy of the DecisionTree program used in problem 2 using "leave one out" method with the data set provided.