1 Surprise and Entropy

In this section, we will clarify the concepts of surprise and entropy. Recall that entropy is one of the standards for us to split the nodes in decision trees until we reach a certain level of homogeneity.

(a) Suppose you have a bag of balls, all of which are black. What's the surprise of taking out a ball whose color is black?

(b) With the same bag of balls, what's the surprise of taking out a white ball?

(c) Now we have 10 balls in the bag, each of which is black or white. Under what color distribution(s) is the entropy of the bag minimized? And under what color distribution(s) is the entropy maximized? Calculate the entropy in each case.
Recall: The entropy of an index set S is the expected surprise of choosing an element from S. For a set S, the entropy

$$H(S) = -\sum p_c \log_2(p_c), \text{ where } p_c = \frac{|i \in S : y_i = c|}{|S|}$$

(d) Draw the graph of entropy $H(p_c)$ when there are only two classes. Is the entropy function strictly concave, concave, strictly convex, or convex? Why? What is the significance?
Hint: For the significance, recall the information gain.

2 Decision Trees

Consider constructing a decision tree on data with d features and n training points where each feature is real-valued and each label takes one of m possible values. The splits are two-way, and are chosen to maximize the information gain. We only consider splits that form a linear boundary parallel to one of the axes. We will only consider a standalone decision tree and not a random forest (hence no randomization). Recall the definition of information gain:

$$IG(\textbf{node}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|},$$

where S is set of samples considered at **node**, S_l is the set of samples remaining in the left subtree after **node**, and S_r is the set of samples remaining in the right subtree after **node**.

(a) Prove or give a counter-example: In any path from the root to a leaf, the same feature will never be split on twice. If false, can you modify the conditions of the problem so that this statement is true?

(b) Prove or give a counter-example: The information gain at the root is at least as much as the information gain at any other node.
Hint: Think about the XOR function.

(c) Suppose that a learning algorithm is trying to find a consistent hypothesis when the labels are actually being generated randomly. There are d Boolean features and 1 Boolean label, and examples are drawn uniformly from the set of 2^{d+1} possible examples. Calculate the number of samples required before the probability of finding a contradiction in the data reaches $\frac{1}{2}$.
(A contradiction is reached if two samples with identical features but different labels are drawn.)

(d) Intuitively, how does the bias-variance trade-off relate to the depth of a decision tree?