CS 189 Introduction to Machine Learning
Spring 2020 Jonathan Shewchuk

DIS9

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. How surprised are you if you take out a black ball?

(b) With the same bag of balls, how surprised are you if you take 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 a measure of expected surprise from choosing an element from S; that is,

$$H(S) = -\sum_C p_C \log_2(p_C), \text{ where } p_C = \frac{|i \in S : y_i = C|}{|S|}.$$

(d) Draw the graph of entropy $H(p_c)$ when there are only two classes C and D, with $p_D = 1 - p_C$. 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 and Random Forests

Random forests are a specific ensemble method where the individual models are decision trees trained in a randomized way so as to reduce correlation among them. Because the basic decision tree building algorithm is deterministic, it produces the same tree every time if we give it the same dataset and use the same hyperparameters (stopping conditions, etc.).

Consider constructing a multi-class binary decision tree on n training points with d real-valued features. The splits are chosen to maximize the information gain. We only consider splits that form a linear boundary orthogonal to one of the coordinate axes.

(a) Give an example or disprove: For every $n \geq 3$, there exists some discrete probability distribution on n objects whose entropy is negative.

(b) One may be concerned that the randomness introduced in random forests may cause trouble. For example, some features or sample points may never be considered at all. We investigate this phenomenon in parts (b)–(d). Consider n training points in a feature space of d dimensions. Consider building a random forest with T binary trees, each having exactly h internal nodes. Let m be the number of features randomly selected (from among d input features) at each treenode. For this setting, compute the probability that a certain feature (say, the first feature) is never considered for splitting in any treenode in the forest.

(c) Now let us investigate the possibility that some sample point might never be selected. Suppose each tree employs $n' = n$ bootstrapped (sampled with replacement) training sample points. Compute the probability that a particular sample point (say, the first sample point) is never considered in any of the trees.

(d) Compute the values of the two probabilities you obtained in parts (b) and (c) for the case where there are $n = 2$ training points with $d = 2$ features each, $T = 10$ trees with $h = 4$ internal nodes each, and we randomly select $m = 1$ potential splitting features in each treenode. You may leave your answer in a fraction and exponentiated form, e.g., $\left(\frac{51}{100}\right)^2$. What conclusions can you draw about the concern mentioned in part (b)?

3 Decision Boundary Visualization on Decision Tree and Random Forest

In this problem, we will visualize the decision boundaries of decision tree, random forest, and adaboost with decision tree. Please go to the Jupyter Notebook part and visualize the decision boundaries of the above approaches. You do not need to write code in the Jupyter Notebook. (Use this notebook to open the file in the Google drive and follow instructions therein).