CS 189: Introduction to Machine Learning - Discussion 9

1. Maximum Entropy Distribution

   Suppose we have a discrete random variable that has a Categorical distribution described by the parameters $p_1, p_2, \ldots, p_d$. Recall that the definition of entropy of a discrete random variable is

   $$H(X) = E[-\log p(X)] = -\sum_{i=1}^{d} p_i \log p_i$$

   Find the distribution (values of the p_i) that maximizes entropy. (Hint: remember that $\sum_{i=1}^{d} p_i = 1$. Don't forget to include that in the optimization as a constraint!)

2. Decision Trees

Recall that training a decision tree requires looking at every feature to find the best split, where the best split greedily maximizes the information gain. The information gain is defined as

$$H - \left[\frac{n_1 H_1 + n_2 H_2}{n_1 + n_2}\right]$$

where H is the entropy at the current node, H_1 is the entropy at the "left" split, and H_2 is the entropy at the "right" split. n_1 and n_2 are the number of data points at the "left" and "right" splits.

(a) What are good values to choose to test the splits?

(b) What is the running time for the naive approach to finding the best split (just finding the split, not training the entire tree)?

(c) What is a smarter way to search for the best split, and what is the running time of this?

3. Random Forests

   (a) In terms of the bias-variance tradeoff, where does a single deep decision tree fall?

   (b) How does bootstrap aggregating, or bagging, help? How are predictions made using bagging?

   (c) How do random forests extend bagging? What problem do random forests help with?