CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

DIS8

1 Curse of Dimensionality in Nearest Neighbor Classification

We have a training set: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$. To classify a new point $\mathbf{x}$, we can use the nearest neighbor classifier:

$$\text{class}(\mathbf{x}) = y^{(i^*)} \quad \text{where } \mathbf{x}^{(i^*)} \text{ is the nearest neighbor of } \mathbf{x}.$$

Assume any data point $\mathbf{x}$ that we may pick to classify is inside the Euclidean ball of radius 1, i.e. $\|\mathbf{x}\|_2 \leq 1$. To be confident in our prediction, in addition to choosing the class of the nearest neighbor, we want the distance between $\mathbf{x}$ and its nearest neighbor to be small, within some distance $\epsilon > 0$:

$$\|\mathbf{x} - \mathbf{x}^{(i^*)}\|_2 \leq \epsilon \quad \text{for all } \|\mathbf{x}\|_2 \leq 1. \tag{1}$$

What is the minimum number of training points we need for inequality (1) **to hold** (assuming the training points are well spread to cover the maximum amount of space)**?**

How does this lower bound depend on the dimension d**?**

Hint: Think about the volumes of the hyperspheres in d dimensions. A d-dimensional hypersphere B with radius r has volume $\text{vol}(B) = c \cdot r^d$ for some scalar c.

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:

$$IG(\textbf{node}) = H(S) - H_{\text{after}} = H(S) - \frac{|S_\ell| H(S_\ell) + |S_r| H(S_r)}{|S_\ell| + |S_r|}$$

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

3 Concerns about Randomness

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. In this problem we will be exploring this phenomenon.

(a) 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 to be the available features (features outside this sampled set of size m are not allowed to make splits for this 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.**

(b) Now let us investigate the possibility that some sample point might never be selected when training bagged decision trees. 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.

(c) **Compute the values of the two probabilities you obtained in parts (a) and (b)** 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 you had starting the problem?