CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik

DIS9

1 Random Forest Motivation

Ensemble learning is a general technique to combat overfitting, by combining the predictions of many varied models into a single prediction based on their average or majority vote.

(a) **The motivation of averaging.** Consider a set of uncorrelated random variables $\{Y_i\}_{i=1}^n$ with mean μ and variance σ^2. Calculate the expectation and variance of their average. (In the context of ensemble methods, these Y_i are analogous to the prediction made by classifier i.)

(b) **Ensemble Learning – Bagging.** In lecture, we covered bagging (Bootstrap AGGregatING). Bagging is a randomized method for creating many different learners from the same data set.

Given a training set of size n, generate T random subsamples, each of size n', by sampling with replacement. Some points may be chosen multiple times, while some may not be chosen at all. If $n' = n$, around 63% are chosen, and the remaining 37% are called out-of-bag (OOB) samples.

(a) Why 63%?

(b) If we use bagging to train our model, How should we choose the hyperparameter T? Recall, T is the number of subsamples, and typically, a few dozen to several thousand trees are used, depending on the size and nature of the training set.

(c) In part (a), we see that averaging reduces variance for uncorrelated classifiers. Real-world prediction will of course not be completely uncorrelated, but reducing correlation among decision trees will generally reduce the final variance. Reconsider a set of correlated random variables $\{Z_i\}_{i=1}^n$. Suppose $\forall i \neq j$, $\text{Corr}(Z_i, Z_j) = \rho$. Calculate the variance of their average.

(d) Is a random forest of stumps (trees with a single feature split or height 1) a good idea in general? Does the performance of a random forest of stumps depend much on the number of trees? Think about the bias of each individual tree and the bias of the average of all these random stumps.

2 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 tree node. For this setting, compute the probability that a certain feature (say, the first feature) is never considered for splitting in any tree node in the forest.

(b) 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.

(c) Compute the values of the two probabilities you obtained in parts (b) and (c) for the case where there are $n = 50$ training points with $d = 5$ features each, $T = 25$ trees with $h = 8$ 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 concerns of not considering a feature or sample mentioned at the beginning of the problem?

3 Hidden Markov Models: Math Review

A Hidden Markov Model is a Markov Process with unobserved (hidden) states.

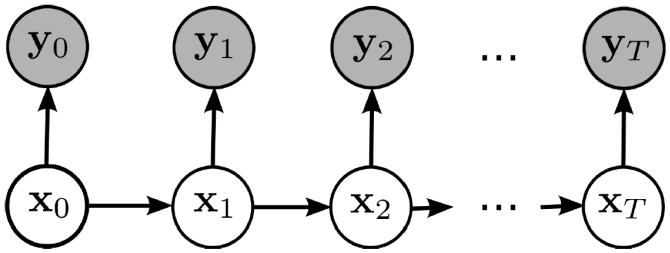

Figure 1: Example Hidden Markov Chain

Consider the following system in $\mathbb{R}^2$, where X_n is the true state at any given time n and Y_n is our observation. Given an initial state X_0, we move to future states by recursively multiplying our current state with transformation matrix A and adding i.i.d. Standard Normal Gaussian noise. When we take an observation Y_n of the true state X_n, we are also exposed to i.i.d. Standard Normal Gaussian Noise.

$$X_{n+1} = AX_n + N(0, I) \tag{1}$$
$$Y_n = X_n + N(0, I) \tag{2}$$

Where we have the $2x2$ transformation matrix A defined as follows:

$$A = \begin{bmatrix} .5 & -.25 \\ -.25 & .75 \end{bmatrix} \tag{3}$$

If we restrict the initial state X_0 to be a unit vector ($||X_0||_2 = 1$), determine the following

(a) What are the eigenvalues of A? Is A a positive semi-definite matrix? (Note that $\sqrt{5} = 2.236$)

(b) What is the $\| E[Y_\infty] \|_2$? Prove your assertion.

(c) Consider the Frobenius Norm of an arbitrary M x N matrix Q, defined as $\|Q\|_F = \sqrt{\sum_i \sum_j |Q_{i,j}|^2}$, which indicates the "magnitude" or "largeness" of a matrix. Is $\| Var[Y_\infty] \|_F$ finite or infinite? Prove your assertion.

You may find the following facts to be useful:

(i) Triangle Inequality: $\|X + Y\| \leq \|X\| + \|Y\|$

(ii) Cauchy Schwarz: $\|XY\| \leq \|X\|\|Y\|$

(iii) Geometric Sum: $\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r} \qquad \forall r \text{ s.t. } 0 < r < 1; a, r \in \mathbb{R}$