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.)
Solution: The average of the Y_i's has the same expectation as each individual Y_i:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}\cdot n\cdot\mu = \mu,$$

but less variance than each of the individual Y_i's:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

(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%? **Solution:** Each sample has probability $(1-1/n)^n$ of not being selected. For large n, $(1-1/n)^n \approx \lim_{n\to\infty}(1-1/n)^n = 1/e \doteq 0.368$

(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.

Solution: An optimal number of subsamples T can be found with validation. Alternatively, we can observe the OOB error.

(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 $\forall i \neq j$, $\mathrm{Corr}(Z_i, Z_j) = \rho$. Calculate the variance of their average.

Solution:

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\operatorname{Var}(Z_i) + \sum_{i\neq j}\sum \operatorname{Cov}(Z_i, Z_j)\right)$$

$$= \frac{\sigma^2}{n} + \frac{n(n-1)\sigma^2\rho}{n^2} = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2.$$

We can see that for large n, the first term dominates, which limits the benefit of averaging.

(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.

Solution: Stumps generally have high bias; they are very simple models that cannot fit to anything with reasonable complexity. If we treat $\{Z_i\}$ as the set of possibly correlated predictions the stumps produce,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \mu_z.$$

This tells us if each stump has high bias, averaging the predictions of all stumps will not reduce this bias. Thus a random forest of stumps is generally a bad idea no matter how many stumps we have.

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.

Solution: The probability that it is not considered for splitting in a particular node of a particular tree is $1 - \frac{m}{d}$. The subsampling of m features at each treenode is independent of all others. There is a total of ht treenodes and hence the final answer is $(1 - \frac{m}{d})^{hT}$.

(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.

Solution: The probability that it is not considered in one of the trees is $(1 - \frac{1}{n})^n$, which approaches $1/e$ as $n \to \infty$. Since the choice for every tree is independent, the probability that it is not considered in any of the trees is $(1 - \frac{1}{n})^{nT}$, which approaches e^{-T} as $n \to \infty$.

(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?

Solution: $(\frac{4}{5})^{200} \approx 4.15 * 10^{-20}$ and $(\frac{49}{50})^{1250} \approx 1.07 * 10^{-11}$. It is quite unlikely that a feature will be missed, and extremely unlikely a sample will be missed.

3 Hidden Markov Models: Math Review

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

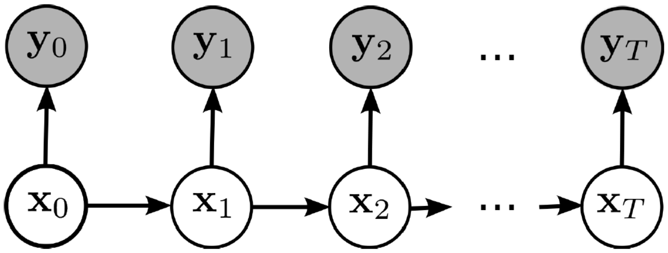

Figure 1: Example Hidden Markov Chain

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.

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

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

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

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 $\sqrt{5} = 2.236$)

Solution: Remember that an eigenvector is a vector $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$, where the constant λ is the eigenvalue corresponding to $\mathbf{v}$. We manipulate the above equation to be $(A - \lambda I)\mathbf{v} = 0$, which implies that $A - \lambda I$ is a singular matrix since it has an eigenvalue of 0.

$$A - \lambda I = \begin{bmatrix} \frac{1}{2} - \lambda & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} - \lambda \end{bmatrix}$$

We can take the determinant of the above matrix and set it to zero in order for the matrix to be singular, giving us the following characteristic polynomial:

$$0 = \left(\frac{1}{2} - \lambda\right)\left(\frac{3}{4} - \lambda\right) - \left(-\frac{1}{4}\right)\left(-\frac{1}{4}\right) = \lambda^2 - \frac{5}{4}\lambda + \frac{3}{8} - \frac{1}{16} = \lambda^2 - \frac{5}{4}\lambda + \frac{5}{16}$$

$$\lambda = \frac{1}{2}\left(\frac{5}{4} \pm \sqrt{\frac{25}{16} - 4\left(\frac{5}{16}\right)}\right) = \frac{1}{8}(5 \pm \sqrt{5})$$

Since $\lambda > 0$ for all possible values, it is a positive-semidefinite matrix (in fact, it is positive definite).

(b) What is the $||\mathbb{E}[Y_\infty]||_2$? Prove your assertion. **Solution:** Lets look at the first several expressions of the true state X

$$\begin{aligned}
X_1 &= AX_0 + \mathcal{N}(0, I) \\
X_2 &= A(AX_0 + N(0, I)) + \mathcal{N}(0, I) \\
X_3 &= A(A(AX_0 + \mathcal{N}(0, I)) + \mathcal{N}(0, I)) + \mathcal{N}(0, I)
\end{aligned}$$

We note that a particular state can be defined by our original state as follows $X_n = A^n X_0 + \sum_{i=0}^{n-1} A^i N(0, I)$. Thus, our observation of that is $Y_n = A^n X_0 + N(0, I) + \sum_{i=0}^{n-1} A^i N(0, I)$. Remember that since matrix A is a real symmetric matrix, we can use spectral decomposition to prove that $A^N = (UDU^\top)^N = UD^N U^\top$, where U is a unitary matrix and D is a diagonal matrix of eigenvalues. Note that our eigenvalues are such that $0 < \lambda < 1$. Therefore, $D^N = 0 \Rightarrow A^N = 0$.

Thus, when we take expectations and norm, we see that

$$\begin{aligned}
||\lim_{n\to\infty} \mathbb{E}[Y_n]||_2 &= ||\mathbb{E}[A^n X_0 + \mathcal{N}(0, I) + \sum_{i=0}^{n-1} A^i \mathcal{N}(0, I)]||_2 \\
&= ||\mathbb{E}[N(0, I) + \sum_{i=0}^{n-1} A^i \mathcal{N}(0, I)]||_2 \\
&= ||0||_2 \\
&= 0
\end{aligned}$$

(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 $||\text{Var}[Y_\infty]||_F$ 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: $\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r} \qquad \forall r$ s.t. $0 < r < 1; a, r \in \mathbb{R}$

Solution: We will approach this part in the same way as part b). Remember from discussion that for multidimensional i.i.d. random variables X,Y with variance I, and constant matrix B:

Var$[BX] = BIB^\top = BB^\top \quad Var[B+X] = Var[X] \quad Var[X+Y] = Var[X]+Var[Y] = I+I = 2I$

Therefore, if we examine $\lim_{n\to\infty} Var[Y_n]$, where we define N(0,I) = Q and note that A is symmetric ($A = A^\top$), we see that:

$$\begin{aligned}
\lim_{n\to\infty} \text{Var}[Y_n] &= \text{Var}[A^n X_0 + \sum_{i=0}^{n-1} A^i Q + Q] \\
&= \text{Var}[\sum_{i=0}^{n-1} A^i Q] + \text{Var}[Q] = \sum_{i=0}^{n-1} \text{Var}[A^i Q] + I \\
&= \sum_{i=0}^{n-1} A^i I (A^\top)^i + I = \sum_{i=0}^{n-1} (AA^\top)^i + I \\
&= \sum_{i=0}^{n-1} (A)^{2i} + I = \sum_{i=0}^{n-1} (UDU^\top)^{2i} + I \\
&= \sum_{i=0}^{n-1} UD^{2i}U^\top + I = U(\sum_{i=0}^{n-1} D^{2i})U^\top + I
\end{aligned}$$

We could stop here and note that $\sum_{i=0}^{n-1} D^{2i}$ is finite since $0 < D_{1,1}, D_{2,2} < 1$. Thus, since D is a diagonal matrix and D^n is also diagonal we can apply the geometric sum formula for each term $\sum_{i=0}^{n-1}(D_{1,1})^{2i}$ and $\sum_{i=0}^{n-1}(D_{2,2})^{2i}$. We then note that the sum is finite, that U and $U^\top$ will preserve magnitude, and I is finite. Therefore, the above limit is finite, which means that $\|\text{Var}[Y_\infty]\|_F$ is also finite.

If we want to decompose further, we can use the Triangle Inequality and Cauchy Schwarz Inequality:

$$\begin{aligned}
\|\lim_{n\to\infty} \text{Var}[Y_n]\|_F &= \|U\left(\sum_{i=0}^{n-1} D^{2i}\right)U^\top + I\|_F \\
&\leq \|I\|_F + \|U\left(\sum_{i=0}^{n-1} D^{2i}\right)U^\top\|_F \\
&\leq \|I\|_F + \|U\|_F \|\left(\sum_{i=0}^{n-1} D^{2i}\right)\|_F \|U^\top\|_F
\end{aligned}$$

We then use the same argument as before to show that the sum of diagonal matrices is a geometric series, and note that I and U are finite matrices. Therefore, both have finite norms and the sum must be finite.