Error Rate of RF

One of Many Possible Solutions to Part of Fun HW Problem

Cynthia Rudin

I am working with a distribution where $P(X = 0) = 1 - \delta$ and $P(X = 1) = \delta$, and $P(Y = 1|X = 0) = 0$ and $P(Y = 1|X = 1) = 1$.

Say a draw a sample from the distribution. The sample possesses n_+ positives and n_- negatives, with n being the total. Note that n_+ and n_- are random.

Bootstrap samples: Choose T bootstrap samples of size $\tilde{n}$. (We generally have $\tilde{n} = n$.)

If a bootstrap sample has at least one positive, the overfitted decision tree will create a leaf for it. Same for negatives. We choose parameters so that all features are available on each split to keep things simple. The only times there are mistakes made is when the bootstrap sample has all positives in more than half the bootstrap samples (or if it has all negatives).

What is the probability that one bootstrap sample contains all positives? $(n_+/n)^{\tilde{n}}$. What is the probability that one bootstrap sample contains all negatives? $(n_-/n)^{\tilde{n}}$.

To make an error, either more than T/2 of the bootstrap samples has no positives (in which case RF votes negative on everything), or more than T/2 of the samples has no negatives (in which case RF votes positive on everything).

Those two types of errors are disjoint - you can't make an error on the positives and also make an error on the negatives. If that were true RF wouldn't have voted either positive or negative, which is not possible.

The probability of either these things happening is: P(all pos at least T/2 times out of T) and P(all neg at least T/2 times out of T).

$$P_{\text{Error}}(\text{Misses all positives}|N_+ = n_+) = \sum_{t=T/2}^{T} \text{Bin}\left(t; T, \left(\frac{n_-}{n}\right)^{\tilde{n}}\right)$$

$$P_{\text{Error}}(\text{Misses all negatives}|N_+ = n_+) = \sum_{t=T/2}^{T} \text{Bin}\left(t; T, \left(\frac{n_+}{n}\right)^{\tilde{n}}\right).$$

$$\begin{aligned} &P_{\text{Error}}(\text{Mistakes} = 0|N_+ = n_+) \\ &= 1 - P(\text{Misses all positives}|N_+ = n_+) - P(\text{Misses all negatives}|N_+ = n_+). \end{aligned}$$

If we have $\tilde{n}$ very large, then the probabilities $\left(\frac{n_-}{n}\right)^{\tilde{n}}$ and $\left(\frac{n_+}{n}\right)^{\tilde{n}}$ are very small, unless n_+ is 0 or n_- is 0. That will drive the sums to 0.

If we have a large number of bootstrap samples T, the error rate is also driven to 0. This is because the standard deviation grows as $\sqrt{T}$ (the variance of the binomial is $Tp(1-p)$, where p is the probability of success). This means that the two binomial distributions will be more concentrated around the smaller values $T\left(\frac{n_-}{n}\right)^{\tilde{n}}$ and $T\left(\frac{n_+}{n}\right)^{\tilde{n}}$, and the probability to be above $T/2$ will diminish as T grows.

That's the main part of the proof.

Technically we're not done though. We don't know in advance what n_+ and n_- are. So we need to compute the distribution of errors. We get a test error of δ if we miss all the positives or an error of $1 - \delta$ if we miss all the negatives.

$$\begin{aligned} P(\text{RF's test error is } \delta) &= P(\text{Misses all positives}) \\ &= \sum_{n_+=0}^{n-1} P(N_+ = n_+) P(\text{RF misses all positives}|N_+ = n_+) \\ P(\text{RF's test error is } 1-\delta) &= P(\text{Misses all negatives}) \\ &= \sum_{n_+=1}^{n} P(N_+ = n_+) P(\text{RF misses all negatives}|N_+ = n_+) \\ P(\text{RF's test error is } 0) &= \sum_{n_+=1}^{n-1} P(N_+ = n_+) P(\text{Mistakes} = 0|N_+ = n_+) \end{aligned}$$

We know that the first two sets of terms are basically 0 and the last one is basically 1. To fill in the details, we recall that data are generated so the probability

of being positive in the distribution is δ, so

$$P(N_+ = n_+) = \text{Bin}(n_+; n, \delta)$$

and

$$P(N_- = n_-) = \text{Bin}(n_-; n, 1 - \delta).$$

Plugging back in, now we have the whole distribution of errors. Let's figure out how RF's distribution of errors compares to DT's distribution of errors.

Recall DT has a high probability of having error δ and a small probability of having error $1 - \delta$. It never has 0 error.

RF has the probability of having error δ is almost 0, the probability of having error $1 - \delta$ is almost 0, and it almost always has 0 error.

When the number of samples in the bootstrap (which is usually n) is large, then the bound gets better. If the number of bootstrap trees is large, RF's advantage increases.