Recursive Feature Elimination by Sensitivity Testing

Anonymous Authors[1]

1. Supplementary Material

1.1. Generalization of Theorem 2.1

In this Section, we will prove a stronger version of Theorem 2.1, generalizing it to apply to a product distribution $\mathcal{D}$ and to a function other than parity.

There are two parameters that are important in generalizing Theorem 2.1, ρ and $I_{\min}$. Under a uniform distribution, each feature j has equal probability of being either 1 or 0. Under a product distribution, one of these two probabilities may be larger than the other. We use $\rho > 0$ to denote the maximum, over all features j, of the ratio between the larger and the smaller of these two probabilities, for product distribution $\mathcal{D}$. Thus, for example, if each feature j is 1 with probability 3/4 and 0 with probability 1/4, then $\rho = 3$.

When the examples are labeled according to a parity function (on a subset of the variables), flipping the value of a relevant feature j in a random example drawn from $\mathcal{D}$ always changes the value of the function. For other functions g, flipping the value of a relevant feature j in a random example drawn from $\mathcal{D}$ will change the value of g with some non-zero probability. We denote the minimum of that probability, over all relevant j, by $I_{\min}$. This is the minimum *influence* of a relevant variable of g, with respect to distribution $\mathcal{D}$ (cf. (Hellerstein & Servedio, 2007)).

For the uniform distribution with g being a parity function, $\rho = 1$ and $I_{\min}$ = 1.

The generalized theorem replaces the polynomial dependence of m on $\frac{1}{\frac{1}{2}-\epsilon}$ in Theorem 2.1 with a polynomial dependence on $\frac{1}{\frac{1}{2}I_{\min}-\rho\epsilon}$.

Theorem 1.1. *Suppose a machine learning algorithm is used to learn a classifier M for a Boolean target concept f defined on n Boolean features, where the target concept labels examples according to the value of a Boolean function g, computed on a fixed subset of the features. Suppose M has true error rate $\epsilon < \frac{1}{2}$, with respect to a product distribution $\mathcal{D}$, where $2\rho\epsilon \leq I_{\min}$. Then there is a quantity t that is polynomial in $n, \ln\frac{1}{\delta}$, and $\frac{1}{\frac{1}{2}I_{\min}-\rho\epsilon}$, with the following property: for all $0 < \delta < 1$, if the $\tilde{R}(j)$ values for all n features are computed using M and an i.i.d. sample of size t, drawn from distribution $\mathcal{D}$, then with probability least $1-\delta$, the computed $\tilde{R}(j)$ values for all the relevant features will be higher than the computed $\tilde{R}(j)$ values for the irrelevant features.*

Proof. Consider a random example a drawn from $\mathcal{D}$. Flipping any relevant bit in a reverses the output of f with probability at least $I_{\min}$.

Let $P(a)$ denote the probability of drawing assignment a from distribution $\mathcal{D}$. By the definition of ρ, for any bit j, $\frac{1}{\rho}P(a) \leq P(a_{\neg j}) \leq \rho P(a)$. Here $a_{\neg j}$ denotes the assignment produced by flipping bit j of a.

Let A denote the set of assignments in $\{0,1\}^n$ such that $M(a) \neq f(a)$.

Consider a relevant variable j of f. First, we will lower bound the probability, for random a drawn from distribution $\mathcal{D}$, that $f(a) \neq M(a_{\neg j})$. It is easy to see that $f(a) \neq M(a_{\neg j})$ iff one of the following two conditions holds: (1) $f(a) \neq f(a_{\neg j})$, and $a_{\neg j} \notin A$, or (2) $f(a) = f(a_{\neg j})$, and $a_{\neg j} \in A$. Thus the probability that $f(a) \neq M(a_{\neg j})$ is lower bounded by the probability that Condition (1) holds. We will now lower bound that probability.

$$\begin{aligned} & Prob[f(a) \neq f(a_{\neg j}) \text{ and } a_{\neg j} \notin A] \\ & \geq Prob[f(a) \neq f(a_{\neg j})] - Prob[a_{\neg j} \in A] \\ & \geq I_{\min} - \rho\epsilon \end{aligned} \tag{1}$$

The last inequality above uses the fact that the total probability mass of A is ϵ, and therefore the total probability mass of assignments a such that $a_{\neg j} \in A$ is at most $\rho\epsilon$.

Thus, for relevant variable j, for random a drawn from $\mathcal{D}$, $Prob[f(a) \neq M(a_{\neg j})] \geq I_{\min} - \rho\epsilon$.

Now consider the case where j is an irrelevant variable. In this case, the only way that $f(a) \neq M(a_{\neg j})$ is if $a_{\neg j} \in A$, which happens with probability at most $\rho\epsilon$. Therefore, $Prob[f(a) \neq M(a_{\neg j})] \leq \rho\epsilon$.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

In the statement of the theorem, we assumed that $I_{\min} > 2\rho\epsilon$. Let $\tau = \frac{1}{2}I_{\min} - \rho\epsilon$.

Now suppose we compute the $\tilde{R}(j)$ values for all features j using an i.i.d. random sample $\mathcal{X}$ drawn from $\mathcal{D}$ and labeled according to f. Let $t = \frac{1}{2\tau^2}\ln\frac{n}{\delta}$ be the size of this sample. Recall that $\tilde{R}(j)$ is the difference between the accuracy of M on $\mathcal{X}$, and the accuracy of M on the sample derived from $\mathcal{X}$ by flipping j in each example. This second accuracy measures the percentage of examples a for which $f(a) = M(a_{\neg j})$. Let $d(j)$ be the percentage of examples a for which $f(a) \neq M(a_{\neg j})$. It follows that for any pair of features j' and j'', $\tilde{R}(j') \geq \tilde{R}(j'')$ iff $d(j') \geq d(j'')$. We will prove the following claim: with probability at least $1-\delta$, $d(j) > \frac{1}{2}I_{min}$ for each relevant feature j, and $d(j) < \frac{1}{2}I_{min}$ for each irrelevant feature j. This suffices to prove the theorem.

To prove the claim, consider a random a drawn from $\mathcal{D}$. We can view the test of whether $f(a) \neq M(a_{\neg j})$ as a Bernoulli trial, with success when the inequality holds. Thus if j is a relevant variable, the probability of success is at least $I_{min} - \rho\epsilon$. If j is an irrelevant variable, the probability of success is at most $\rho\epsilon$.

With this view, we can apply a standard bound of Hoeffding. Consider a sequence of m independent Bernoulli trials, each with probability p of success. Suppose that out of these m trials, the observed fraction of successes is $\hat{p}$. The bound of Hoeffding states that for any $c > 0$, $Prob[\hat{p} \geq p + c] \leq e^{-2mc^2}$ (Hoeffding, 1963). By exchanging the role of failures and successes, it immediately follows that the inequality $Prob[\hat{p} \leq p - c] \leq e^{-2mc^2}$ also holds. Thus if $m \geq \frac{1}{2c^2}\ln\frac{1}{\delta}$, we have the following two inequalities

$$Prob[\hat{p} \geq p + t] \leq \delta \tag{2}$$

$$Prob[\hat{p} \leq p - t] \leq \delta \tag{3}$$

We apply these two inequalities to the tests performed in computing $d(j)$ from $\mathcal{X}$. Consider a random assignment a drawn from $\mathcal{D}$. If j is relevant, then the probability of success (i.e., that $f(a) \neq M(a_{\neg j})$) is at least $(I_{min} - \rho\epsilon)$. If j is irrelevant, then the probability of success is at most $\rho\epsilon$. The assignments in $\mathcal{X}$ correspond to $\frac{1}{2\tau^2}\ln\frac{n}{\delta}$ Bernoulli trials. Because $\tau = \frac{1}{2}I_{\min} - \rho\epsilon$, applying the above bounds with $c = \tau$ and $s = \frac{1}{2\tau^2}\ln\frac{n}{\delta}$ implies that the following holds for each feature j: If j is relevant, then $Prob[d(j) \leq \frac{1}{2}I_{\min}] \leq \frac{\delta}{n}$, and if j is irrelevant, then $Prob[d(j) \geq \frac{1}{2}I_{\min}] \leq \frac{\delta}{n}$.

Since there are n features, it follows that with probability at least $1-\delta$, the $d(j)$ values for the relevant variables will all be greater than $\frac{1}{2}I_{\min}$, and the $d(j)$ values for the irrelevant features will be less then $\frac{1}{2}I_{\min}$. $\square$

The condition $\epsilon < I_{min}/(2\rho)$ in the above theorem limits its applicability to arbitrary functions g, even under the uniform distribution. For example, consider the consensus function (which is correlation immune): $g(x_1, \ldots, x_k) = 1$ iff $x_1 = x_2 = \ldots = x_k$. Under the uniform distribution, the value of I_{min} for the consensus function is $1/2^{k-2}$. For $k = 4$, the condition $\epsilon < I_{min}/(2\rho)$ would then be satisfied only if the error ϵ of model M was less than $1/8$.

We note that while it might be possible to prove a version of the theorem with a somewhat less restrictive condition, there are inherent limits as to what can be proved. For example, suppose g is a function on k variables that classifies at least 75% of its 2^k possible examples as negative. (The consensus function on 3 variables has this property.) Then the model that predicts negative on all examples has exactly 75% accuracy. Using RFEST with such a model, there is no hope of distinguishing relevant from irrelevant variables.

References

Hellerstein, L. and Servedio, R. A. On pac learning algorithms for rich boolean function classes. *Theoretical Computer Science*, 384(1):66 – 76, 2007. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2007.05.018. Theory and Applications of Models of Computation.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459.