Analysis of Histogram and Decision Tree classifiers

1 Empirical Risk Minimizer (ERM)

We will consider the classifier $\widehat{h}(x)$ that minimize the empirical risk over a class of classifiers $\mathcal{H}$, i.e.

$$\widehat{h}(x) = \arg\min_{h \in \mathcal{H}} \widehat{R}(h)$$

We can use concentration of measure arguments (e.g. the VC theorem) to get a uniform bound on the deviation of true risk $R(h)$ and empirical risk $\widehat{R}(h)$ for all classifiers $h \in \mathcal{H}$: With probability $> 1 - \delta$,

$$\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| \leq \epsilon_{\mathcal{H}}$$

where $\epsilon_{\mathcal{H}}$ depends on the complexity of the class $\mathcal{H}$ such as its VC dimension, cardinality, growth function, etc., and number of training data n.

We can use this bound to analyze the risk for ERM as follows: With probability $> 1 - \delta$,

$$\begin{aligned} R(\widehat{h}) &\leq \widehat{R}(\widehat{h}) + \epsilon_{\mathcal{H}} \\ &\leq \widehat{R}(h) + \epsilon_{\mathcal{H}} \quad \forall h \in \mathcal{H} \\ &\leq R(h) + 2\epsilon_{\mathcal{H}} \quad \forall h \in \mathcal{H} \end{aligned}$$

This implies that with probability $> 1 - \delta$,

$$R(\widehat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + 2\epsilon_{\mathcal{H}}$$

i.e. the ERM classifier is almost as good as the best classifier in the class $\mathcal{H}$ and the gap depends on the complexity of the class. Smaller the class, smaller the gap (easier it is to *find* the classifier).

2 Decomposing the overall error

Ideally, we want the risk of our classifier to be as close to the Bayes risk as possible, i.e. we are interested in bounding the excess risk:

$$\mathcal{E}(\widehat{h}) = R(\widehat{h}) - R^*$$

To obtain bounds on the excess risk, we typically analyze its decomposition into two terms:

$$\mathcal{E}(\widehat{h}) = R(\widehat{h}) - R^* = \left[R(\widehat{h}) - \inf_{h \in \mathcal{H}} R(h) \right] + \left[\inf_{h \in \mathcal{H}} R(h) - R^* \right]$$

The first term denotes the *approximation error* (equivalent to bias) and the second term denotes the *estimation error* (equivalent to variance). Notice that the more complex the class $\mathcal{H}$, smaller is the approximation error, but larger the estimation error (as discussed in previous section, estimation error is bounded by $2\epsilon_{\mathcal{H}}$). This is akin to bias-variance tradeoff.

3 Histogram Classifier

Let $\mathcal{H}_m := \{\text{all classifiers built on a uniform partition of the domain into } m^d \text{ bins}\}$, i.e. a classifier $h \in \mathcal{H}_m$ is either 0 or 1 on each bin. Consider the ERM classifier:

$$\widehat{h}(x) = \arg\min_{h \in \mathcal{H}_m} \widehat{R}(h)$$

It is easy to check that the Histogram classifier $\widehat{h}$ is essentially a majority vote on each bin. To analyze its performance, notice that the VC dimension of $\mathcal{H}_m$ is m^d since the class can shatter (generate all possible labelings for) a set of m^d points. Alternatively, notice that $|\mathcal{H}_m| = 2^{m^d}$. Therefore, we have using VC theorem, with high probability

$$R(\widehat{h}) \leq \inf_{h \in \mathcal{H}_m} R(h) + O\left(\sqrt{\frac{m^d \log n}{n}}\right)$$

This provides a bound on the estimation error for the Histogram classifier.

4 Decision Tree Classifier

Let $\mathcal{T}_k := \{\text{all decision tree classifiers with k leaves i.e. partition of the domain into } k \text{ bins}\}$, (a classifier $t \in \mathcal{T}_k$ is either 0 or 1 on each bin. Consider the ERM classifier:

$$\widehat{t}(x) = \arg\min_{t \in \mathcal{T}_k} \widehat{R}(t)$$

To analyze its performance, notice that the VC dimension of $\mathcal{T}_k$ can be upper bounded as $k(d+1)$ since each split is a linear threshold classifier and hence can shatter (generate all possible labelings for) a set of additional $d+1$ points. Therefore, we have using VC theorem, with high probability

$$R(\widehat{t}) \leq \inf_{t \in \mathcal{T}_k} R(t) + O\left(\sqrt{\frac{kd \log n}{n}}\right)$$

This provides a bound on the estimation error for the ERM Decision Tree classifier.

5 Box-counting dimension and Lipschitz boundaries

To characterize the approximation error $\inf_{h \in \mathcal{H}} R(h) - R^*$, we need to place some mild assumptions on the Bayes decision boundary (the decision boundary of the Bayes optimal classifier). Just as in regression, we assume the regression function has a bounded derivative or is Lipschitz, in classification we will assume that the Bayes decision boundary is essentially Lipschitz.

However, notice that the boundary may not have a functional form (see Figure 5), so we may not be able to assume that the boundary is Lipschitz. But we use an equivalent notion and say that the Bayes decision boundary has box-counting dimension $d-1$. Formally, if we assume that the domain is a hypercube i.e. $x \in [0,1]^d$, then a curve has box-counting dimension α if the number of bins of size $1/m$ needed to cover the boundary is $O(m^\alpha)$. Essentially, this says that the boundary is a $d-1$ dimensional curve embedded in d dimensions.

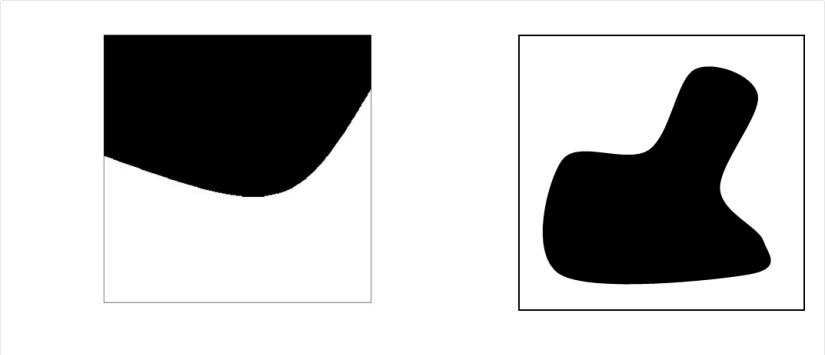

Figure 1: Two figures showing Bayes decision boundaries: $h^*(x) = 1$ if x lies in black region and is 0 otherwise. In the first figure, the decision boundary can be characterized as a function in one of the coordinates. In the second figure, it is not possible to describe the boundary in a functional form.

To see the equivalence to Lipschitz functional assumption, consider $d = 2$ and let the boundary be a Lipschitz function in one of the two features, i.e. $|f(x)-f(x')| \leq C|x-x'|$. Now if that feature changes by $1/m$, then the Lipschitz function can only change by $O(1/m)$ by definition. Therefore, the function restricted to $1/m$ of the domain can be covered by $O(1)$ boxes. To cover the entire function on the domain, we need $O(m)$ boxes. More generally, in d-dimensions, we can consider the boundary be a Lipschitz function in $d-1$ of the d features. Now if each of the $d-1$ features changes by $1/m$, then the Lipschitz function can only change by $O(1/m)$ by definition. Therefore, the function restricted to $1/m^{d-1}$ of the domain can be covered by $O(1)$ boxes. To cover the entire function on the domain, we need $O(m^{d-1})$ boxes. The box-counting assumption is more general and holds for boundaries that may not have a functional form.

6 Rate of convergence comparison for Histogram vs. Decision Tree Classifier

We are now ready to investigate the approximation error of histogram and decision tree classifiers under the assumption that the box-counting dimension of the Bayes decision boundary is $d-1$, i.e. the Bayes decision boundary is a $d-1$ dimensional curve in d dimensions. Without loss of generality, we also assume that the domain is $[0,1]^d$. Extension to any compact domain is straight-forward. Also, we assume that the marginal density is bounded $p(x) \leq B$.

Lets first consider the histogram case. Notice that the approximation error is $\inf_{h\in\mathcal{H}_m} R(h) - R^* = R(\bar{h}) - R^*$, the best *fit* of a classifier built on a uniform partition of the domain into m^d bins. The best classifier $\bar{h}$ in the histogram class will only differ from the Bayes classifier in the bins that intersect the boundary (since the best classifier is restricted to have a constant label in each bin). Thus, the approximation error is given as: (here $1(A)$ denotes indicator of set A)

$$\begin{aligned}\inf_{h\in\mathcal{H}_m} R(h) - R^* &= \mathbb{E}[|2m(X) - 1/2|1(\bar{h} \neq h^*)] \leq P(\bar{h} \neq h^*) \\ &\leq B\text{vol}(\{x : \bar{h}(x) \neq h^*(x)) \preceq B \times m^{d-1} \times m^{-d} = O(m^{-1})\end{aligned}$$

Thus, for histogram classifier we have the following bound on the excess risk: With high probability,

$$R(\widehat{h}) - R^* = O\left(\frac{1}{m} + \sqrt{\frac{m^d \log n}{n}}\right)$$

Thus, the best bin-width which balances the approximation and estimation error is given as $1/m \asymp n^{-1/(d+2)}$. And the risk of the histogram classifier converges to the Bayes risk at a rate $n^{-1/(d+2)}$. (Notice that the rate of MSE convergence of a kernel regression estimator was $n^{-2/(d+2)}$) and the classification error of the plug-in histogram classifier is indeed square-root of this.)

Now consider decision trees. Specifically, lets restrict attention to dyadic decision trees, where the splits only occur at mid-points. Let $\bar{t}$ denote the best *fit* of a classifier built on a non-uniform partition of the domain into k bins, i.e. best fit of a tree classifier with k leaves. Sine the trees are dyadic, the best classifier $\bar{t}$ in the tree class can be obtained by taking the corresponding histogram classifier $\bar{h}$ and pruning any leaves that don't intersect the boundary. Since the number of bins that intersect the boundary is $O(m^{d-1})$, it can be shown that the best tree classifier $\bar{t}$ has $k = O(m^{d-1})$ leaves[1]. Again $\bar{t}$ will only differ from the Bayes classifier in the bins that intersect the boundary (since the best classifier is restricted to have a constant label in each bin). Thus, the approximation error is

$$\inf_{t \in \mathcal{T}_k} R(h) - R^* = O(m^{-1})$$

Thus, for histogram classifier we have the following bound on the excess risk: With high probability,

$$R(\widehat{h}) - R^* = O\left(\frac{1}{m} + \sqrt{\frac{m^{d-1} \log n}{n}}\right)$$

Thus, the best #splits $k \asymp m^{d-1}$ which balances the approximation and estimation error is given by $1/m \asymp n^{-1/(d+1)}$. And the risk of the histogram classifier converges to the Bayes risk at a rate $n^{-1/(d+1)}$. Thus, decision trees offer a better bias-variance tradeoff than histogram classifier for decision boundaries that have box-counting dimension $d-1$.

For tree classifiers, the rate $n^{-1/(d+1)}$ can be improved further to $n^{-1/d}$ using better procedures. In fact, $n^{-1/d}$ is the minimax optimal rate of convergence for any classifier under the box-counting assumption. Thus, decision tree classifiers are *minimax optimal*. However, they are computationally more challenging - the computational complexity scales exponentially in the dimension d and hence are typically used for low-dimensional settings only.

[1]This follows by bounding the number of leaves that $\bar{t}$ has at each level by the number of leaves at each level that can intersect the boundary, and summing over all levels