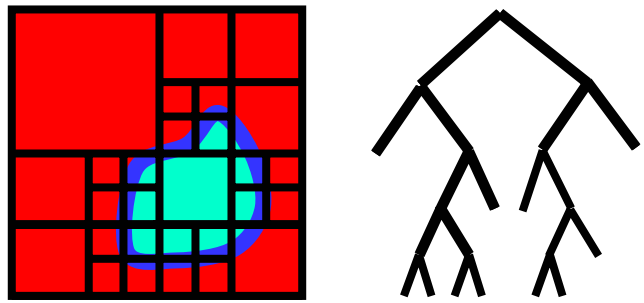
Minimax-Optimal Classification with Dyadic Decision Trees

UCLA IPAM October 27, 2004



Rob Nowak and Clay Scott

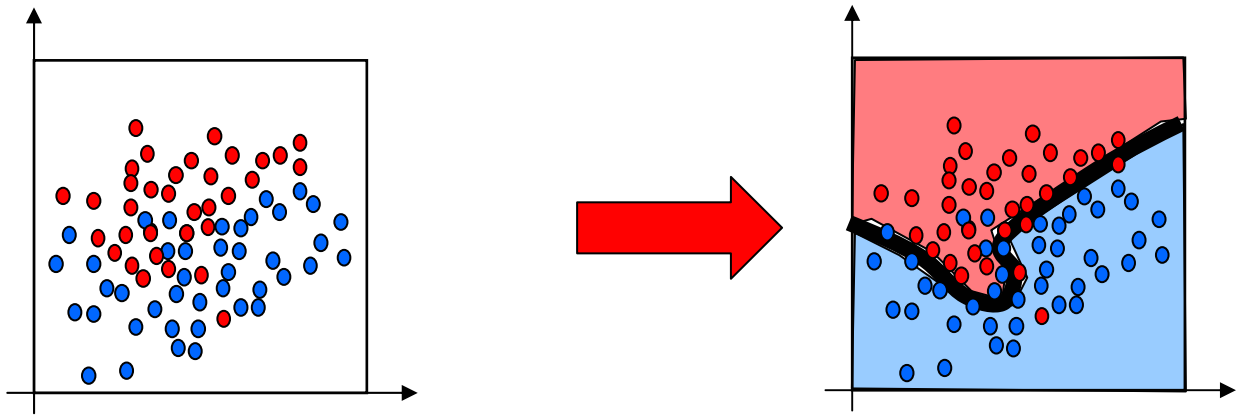
University of Wisconsin-Madison and Rice University

nowak@engr.wisc.edu

Supported by the NSF and the ONR

Basic Problem

Learning and Classification: build a decision rule based on labeled training data



Given n training points, how well can we do ?

Classification

$\mathcal{X}$ = feature space

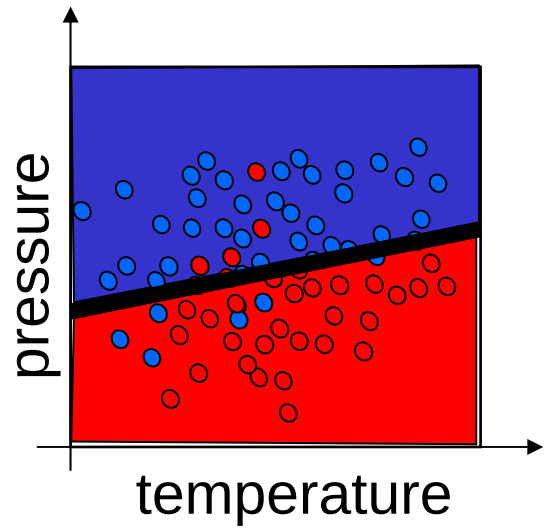
$\mathcal{Y}$ = class labels

Problem: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$

Example: **Weather**

X = (pressure, temperature)

Y = **Rain** or **Shine**



A **classifier** is a mapping

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

Probabilistic Framework

Assume (X, Y) is a **random variable** with probability measure $\mathbf{P}$

The **risk** (**probability of error**) of f is

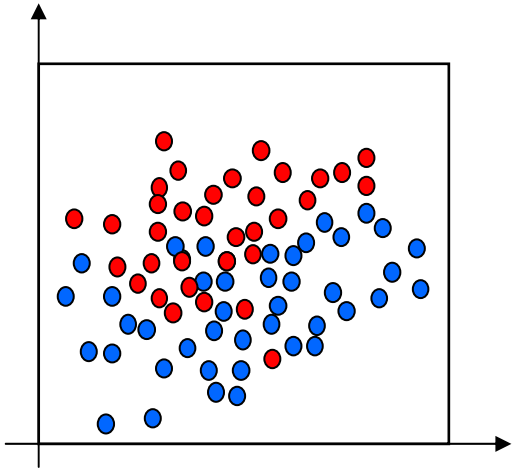
$$R(f) = \mathbf{P}(f(X) \neq Y)$$

The **optimal (Bayes) classifier** is

$$f^* = \arg\min_f R(f)$$

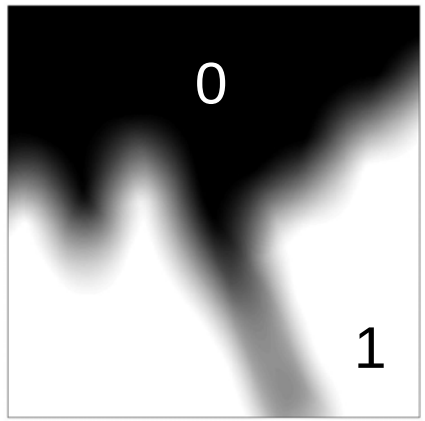
where the min is taken over all measurable f

Learning from Data



training data distributed

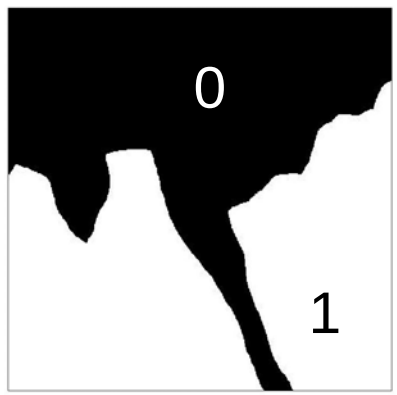
$$\mathbf{P}(X, Y) = P(Y|X)P(X)$$



$$P(Y = 1|X = x)$$

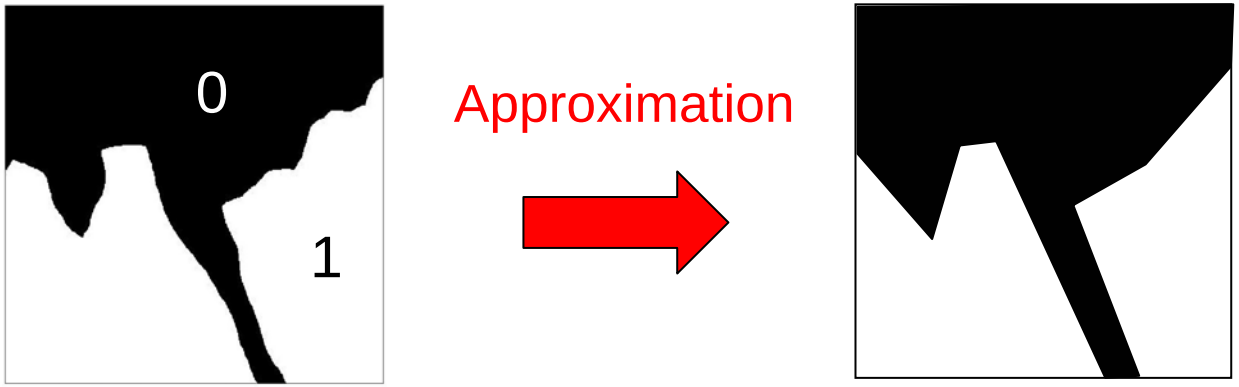
Learn (estimate) set

$$P(Y = 1|X = x) \geq 1/2$$

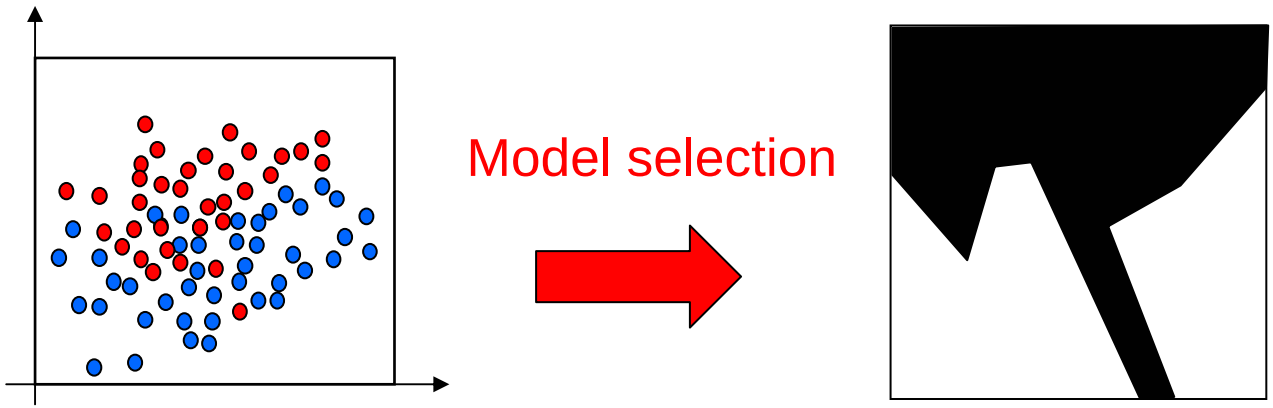


Approximation and Estimation

How easily can f^* be approximated?

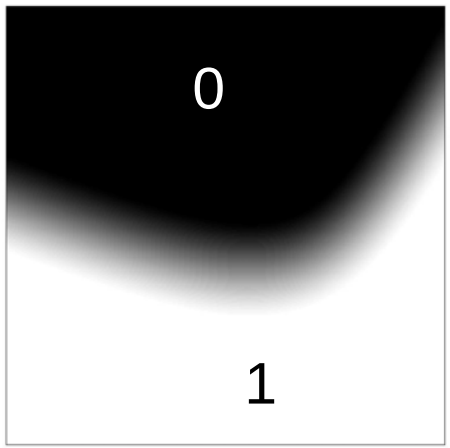


How easily can a "good" approximation be selected?

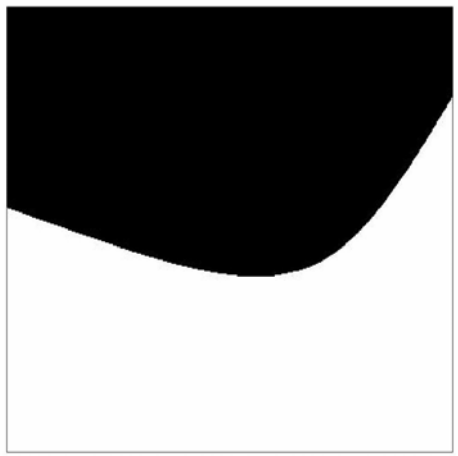


Classifier Approximations

Let $\mathcal{F}$ denote a collection of classifiers. Each $f \in \mathcal{F}$ corresponds to a set $G_f = 1_{\{f(x)=1\}}$



$P(Y = 1|X = x)$

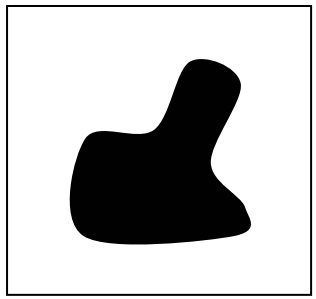


$G_{f^*} = \{x : P(Y = 1|X = x) \geq 1/2\}$

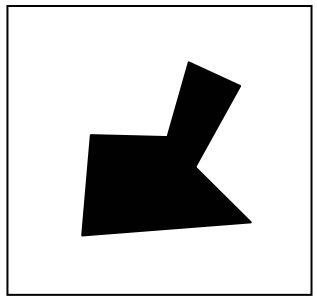
Consider the (set) approximation problem

$$\inf_{f \in \mathcal{F}} R(f) - R(f^*)$$

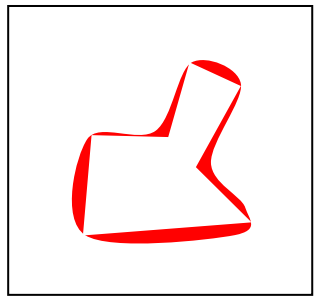
Approximation Error



G_{f^*}



G_f



Δ_{f,f^*}

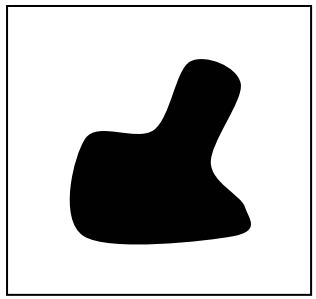
Symmetric difference set

$$\Delta_{f,f^*} = G_f \cap G_{f^*}^c \bigcup G_f^c \cap G_{f^*}$$

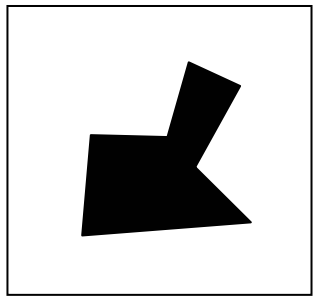
Error

$$\begin{aligned} R(f) - R(f^*) &= P(f(X) \neq Y) - P(f^*(X) \neq Y) \\ &= P(X \in \Delta_{f,f^*}) \times \\ &\quad \Big[P(f(X) \neq Y \mid X \in \Delta_{f,f^*}) \\ &\quad - P(f^*(X) \neq Y \mid X \in \Delta_{f,f^*}) \Big] \end{aligned}$$

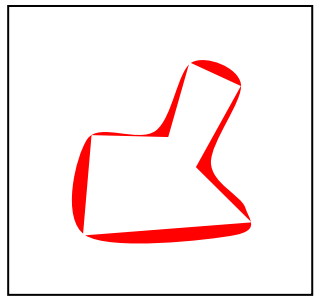
Approximation Error



G_{f^*}



G_f



Δ_{f,f^*}

$$R(f) - R(f^*) \;=\; p(\Delta_{f,f^*})q(\Delta_{f,f^*})$$

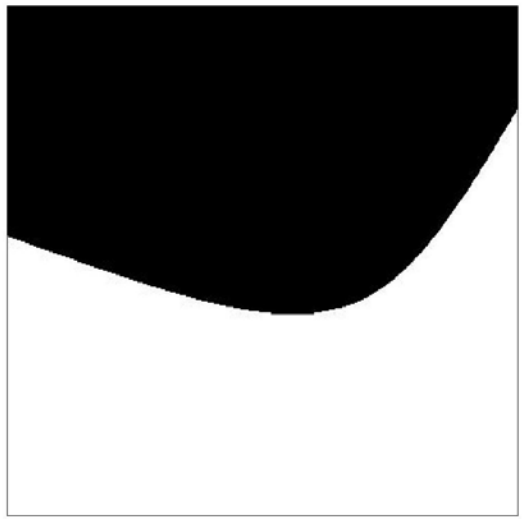
boundary smoothness

$$p(\Delta_{f,f^*}) \;=\; P(X \in \Delta_{f,f^*})$$

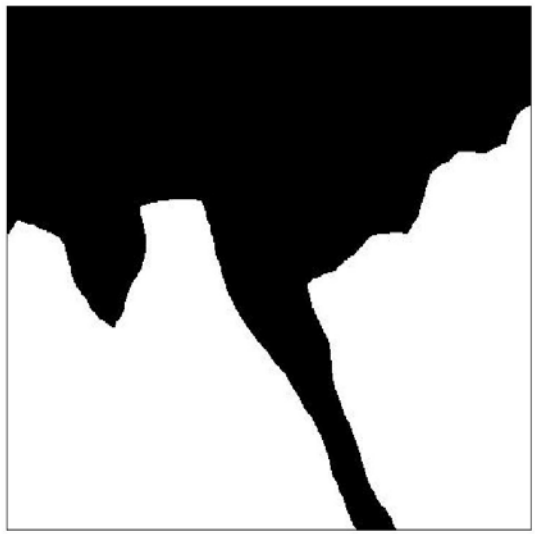
risk functional (transition) smoothness

$$q(\Delta_{f,f^*}) \;=\; \Big[P(f(X) \neq Y \,|\, X \in \Delta_{f,f^*}) - P(f^*(X) \neq Y \,|\, X \in \Delta_{f,f^*})\Big]$$

Boundary Smoothness $\gamma > 0$



smooth γ_1

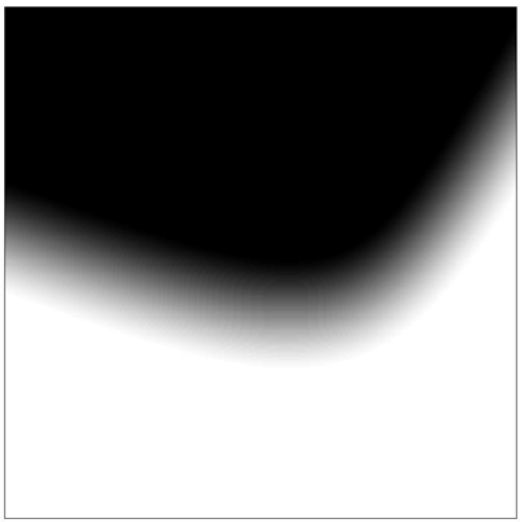


irregular γ_2

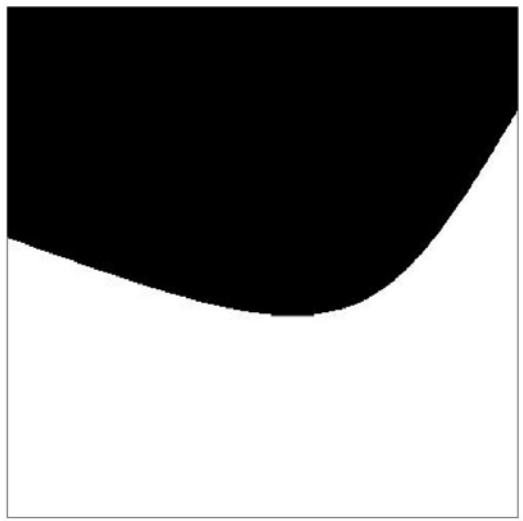
$$\gamma_1 > \gamma_2$$

smoother boundary $\Rightarrow$ faster error decay

Transition Smoothness $\kappa \geq 1$



smooth κ_1



irregular κ_2

$$\kappa_1 > \kappa_2$$

smoother transition $\Rightarrow$ faster error decay

Fundamental Limit to Learning



γ = Hölder smoothness of boundary

κ = transition smoothness

d = dimension of $\mathcal{X}$

For certain classes of distributions $\mathcal{D}(\gamma, \kappa, d)$

$$\inf_{\widehat{f}_n} \sup_{\mathbf{P} \in \mathcal{D}} \left[\mathbf{E} \left\{ R(\widehat{f}_n) \right\} - R(f^*) \right] \asymp n^{-\kappa/(2\kappa + (d-1)/\gamma - 1)}$$

where inf is over all classification learning schemes

Mammen & Tsybakov (1999)

Related Work

Global smoothness constraints on densities: Optimal rates with plug-in density estimates

- Marron (1983)
- Yang (1999)

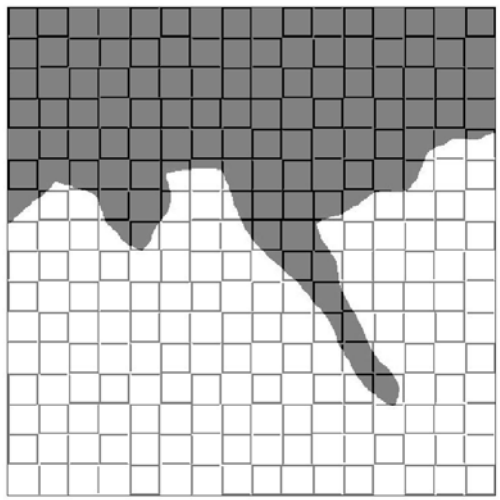
Smoothness constraints on/near Bayes decision boundary: Impractical estimators achieving minimax rates

- Mammen and Tsybakov (1999)
- Tsybakov (2004)
- Tsybakov and van de Geer (2004)
- Audibert (2004)

Box-Counting Class

$\mathcal{D}_{\text{box}}(d)$ = all distributions with bounded densities and Bayes decision boundaries with finite box-counting dimension (box $\Leftrightarrow \gamma = 1$)

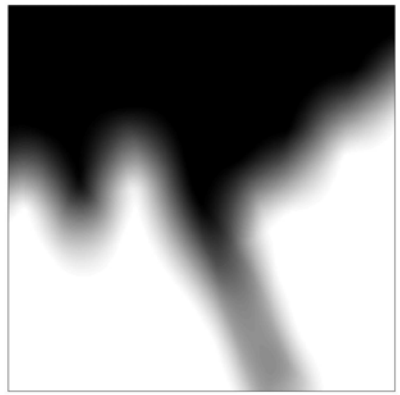
Uniform partition into m^d boxes



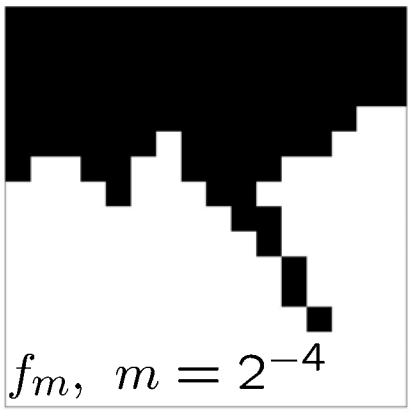
decision boundary passes through $O(m^{d-1})$ boxes

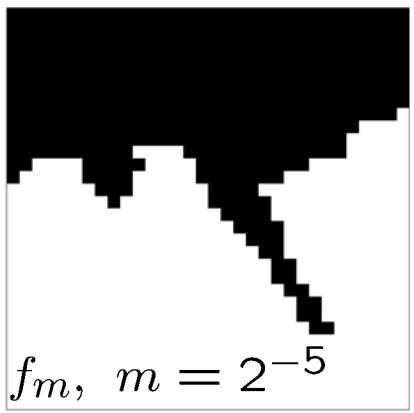
Box-Counting Sub-Classes $\mathcal{D}_{\text{box}}(\kappa, d)$

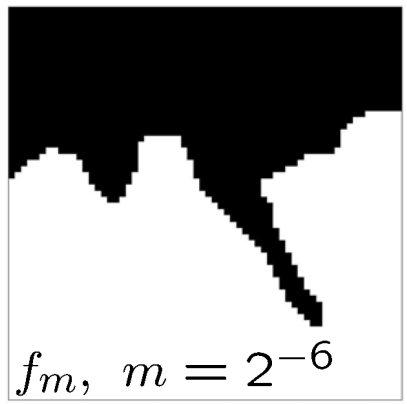
$\kappa = 1$



$\kappa > 1$

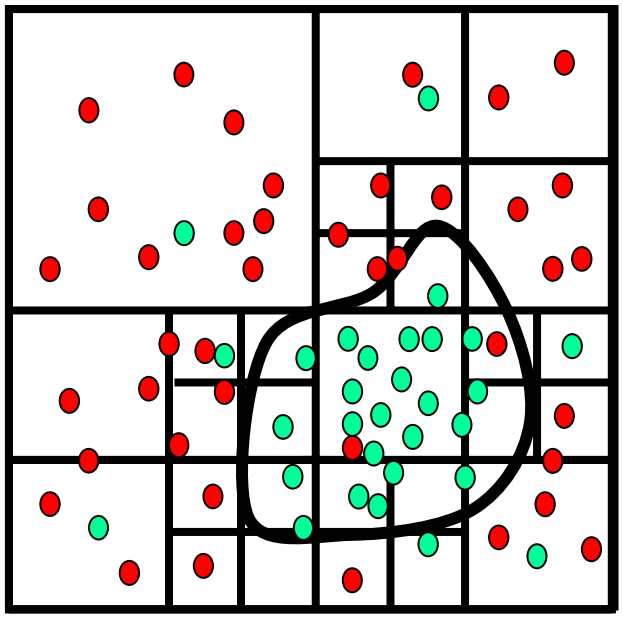




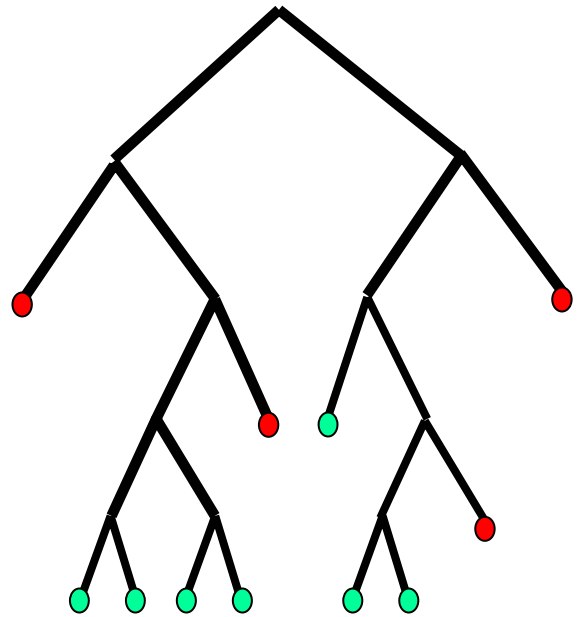


$$R(f_m) - R(f^*) \ = \ O(m^{-\kappa})$$

Dyadic Decision Trees



recursive dyadic partition



corresponding dyadic decision tree
- majority vote at each leaf

Dyadic Decision Trees



$\gamma = 1$ (Lipschitz smooth boundary)

$\kappa \geq 1$ transition smoothness

$d =$ dimension of $\mathcal{X}$

$$\inf_{\widehat{f}_n} \sup_{\mathbf{P} \in \mathcal{D}_{\text{box}}(\kappa,d)} \left[\mathbf{E}\left\{ R(\widehat{f}_n) \right\} - R(f^*) \right] \asymp n^{-\kappa/(2\kappa+d-2)}$$

$$\sup_{\mathbf{P} \in \mathcal{D}_{\text{box}}(\kappa,d)} \left[\mathbf{E}\left\{ R(T_n) \right\} - R(f^*) \right] \preccurlyeq \left(\frac{\log n}{n} \right)^{\kappa/(2\kappa+d-2)}$$

The Classifier Learning Problem

Training Data:

$\{X_i, Y_i\}_{i=1}^n \sim$ i.i.d. $\mathbf{P}(X, Y)$, unknown
$X_i \in \mathcal{X}$ and $Y_i \in \{0, 1\}$

Model Class:

$\mathcal{T}$ = set of all "dyadic decision trees"

Problem:

select a "good" T from $\mathcal{T}$ based on $\{X_i, Y_i\}_{i=1}^n$

"good" $T \Leftrightarrow \mathbf{P}(T(X) \neq Y)$ is small

Empirical Risk

$R(T)$ = true error probability of T

$\widehat{R}(T)$ = empirical error probability of T

$$\widehat{R}(T) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T(X_i)\neq Y_i\}}$$

$\widehat{R}(T)$ is an average of i.i.d. Bernoulli random variables, and $R(T) = \mathbf{E}\left\{\widehat{R}(T)\right\}$

Chernoff's bound tells us that for $0 \leq \delta \leq 1$

$$\mathbf{P}\left(R(T) > \widehat{R}(T) + \sqrt{\log(1/\delta)/2n}\right) \leq \delta$$

actual risk is probably not much larger than empirical risk

Error Deviation Bounds

Chernoff's bound holds for a single, fixed tree.

Need a bound that holds uniformly for all trees.

assign a confidence δ_T to each $T \in \mathcal{T}$

$$\mathbf{P}\left(\exists T \in \mathcal{T} : R(T) > \hat{R}(T) + \sqrt{\log(1/\delta_T)/2n}\right)$$

$$\leq \sum_{T \in \mathcal{T}} \mathbf{P}\left(R(T) > \hat{R}(T) + \sqrt{\log(1/\delta_T)/2n}\right)$$

$$\leq \sum_{T \in \mathcal{T}} \delta_T$$

Uniform Deviation Bound

If $\sum_{T \in \mathcal{T}} \delta_T \leq \delta \in [0, 1]$, then with probability at least $1 - \delta$

$$R(T) \leq \widehat{R}(T) + \sqrt{\frac{\log(1/\delta_T)}{2n}}, \quad \forall\, T \in \mathcal{T}$$

Recall: We want to select $T \in \mathcal{T}$ with small risk – suggests selecting T that minimizes upper bound

$\log(1/\delta_T)$ can be viewed as a penalty term for tree T

Setting Penalties

It is reasonable to penalize larger trees:

$$\log(1/\delta_T) \propto |T|, \text{ number of leafs in } T$$

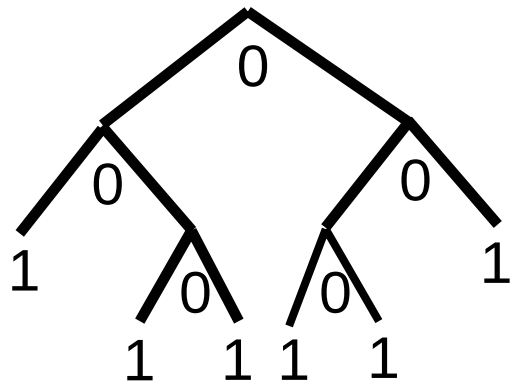
We require $\sum \delta_T \leq \delta$

Claim: if $\delta_T \equiv \delta\, 2^{-3|T|}$, then $\sum \delta_T \leq \delta$

$$\boxed{\text{penalty} = 3|T| + \log(1/\delta)}$$

Setting Penalties

prefix codes for trees:



code: 0001001111
+ 6 bits for leaf labels

code-length $\ell(T) = 3|T| - 1$

Kraft inequality $\Rightarrow \sum_{T \in \mathcal{T}} 2^{-\ell(T)} \leq 1$

Uniform Deviation Bound

If $\delta_T = \delta 2^{-3|T|}$, with probability at least $1 - \delta$

$$R(T) \;\leq\; \widehat{R}(T) + \sqrt{\frac{3|T| + \log(1/\delta)}{2n}}\,, \quad \forall T \in \mathcal{T}$$

smaller trees are less penalized
(i.e., receive more a priori weight)

Decision Tree Selection

Set $\delta = 1/n$ and

$$T_n \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \sqrt{\frac{3|T| + \log(n)}{2n}} \right\}$$

Compare with CART: $T_{\mathsf{CART}} \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \lambda |T| \right\}$

Oracle Bound:

$$\mathbf{E}\left\{R(T_n)\right\} - R(f^*) \leq \min_{T \in \mathcal{T}} \left\{ R(T) - R(f^*) + \sqrt{\frac{3|T| + \log(n)}{2n}} \right\}$$

Approximation Error

Estimation Error

Rate of Convergence

$$\inf_{\widehat{f}_n} \sup_{\mathbf{P} \in \mathcal{D}_{\text{box}}(\kappa, d)} \left[\mathbf{E} \left\{ R(\widehat{f}_n) \right\} - R(f^*) \right] \asymp n^{-\kappa/(2\kappa+d-2)}$$

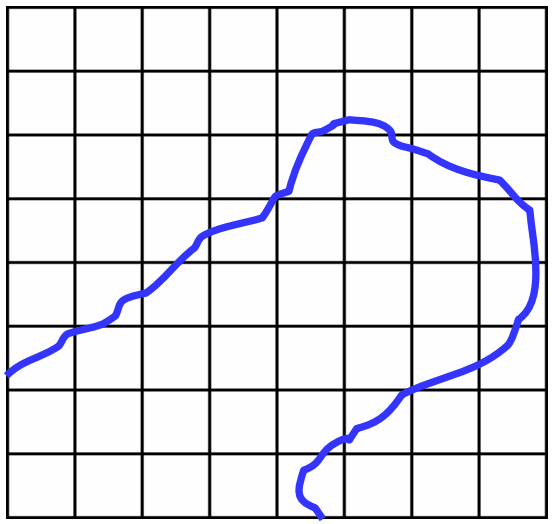
BUT...

$$\begin{aligned}
&\mathbf{E}\{R(T_n)\} - R(f^*) \\
&\quad \leq \min_{T \in \mathcal{T}} \left\{ R(T) - R(f^*) + \sqrt{\frac{3|T| + \log(n)}{2n}} \right\} \\
&\quad \leq C_1 m^{-\kappa} + C_2 \sqrt{m^{d-1} \log n / n} \\
&\quad \asymp \left(\frac{\log n}{n} \right)^{\kappa/(2\kappa+d-1)}
\end{aligned}$$

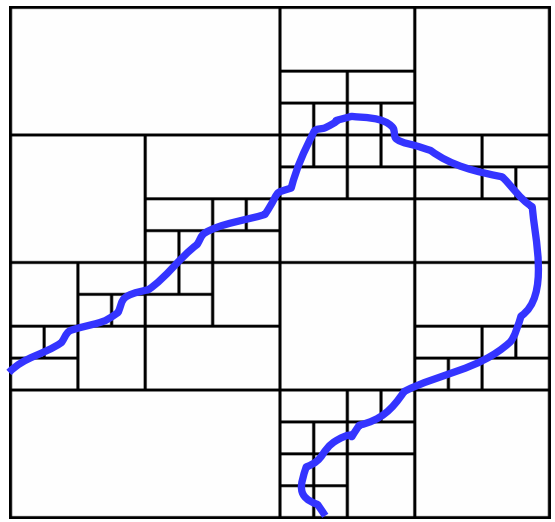
Why too slow ?

Balanced vs. Unbalanced Trees

same number of leafs



T_1



T_2

$$T_n \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \sqrt{\frac{3|T| + \log(n)}{2n}} \right\}$$

all |T| leaf trees are equally favored

Spatial Adaptation

Observe

$$R(T) - \hat{R}(T) = \sum_{L \in \text{leaf}(T)} R(L) - \hat{R}(L)$$

where

$$R(L) = \mathbf{P}(T(X) \neq Y, X \in L)$$

local error

$$\hat{R}(L) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{T(X_i) \neq Y_i, X_i \in L\}}$$

local empirical error

Bernoulli$(R(L))$

Relative Chernoff Bound

With probability at least $1 - \delta_L$

$$\begin{aligned} R(L) - \widehat{R}(L) &\leq \sqrt{\frac{2R(L)\log(1/\delta_L)}{n}} \\ &\leq \sqrt{\frac{2P_L\log(1/\delta_L)}{n}} \end{aligned}$$

where $P_L = \mathbf{P}(X \in L)$

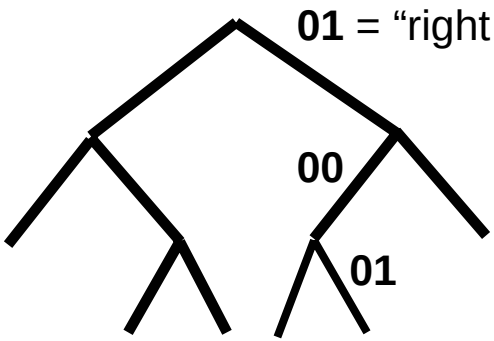
"small" leaf $\Leftrightarrow$ small P_L $\Leftrightarrow$ small deviation

Designing Leaf Penalties

Let $\mathcal{L}$ denote every possible dyadic cell L. Assign costs $\ell(L) = 3j(L)$ where $j(L)$ is the depth of L in tree. Then

$$\sum_{L \in \mathcal{L}} 2^{-\ell(L)} \leq 1$$

prefix code construction :



code for L:

010001110

code length:

$$2j(L) + 1 < 3j(L)$$

Uniform Deviation Bound

With probability at least $1-\delta$, $\forall\ T \in \mathcal{T}$

$$R(T) \ \leq \ \hat{R}(T) \ + \ \sum_{L \in T} \sqrt{2 P_L \frac{3j(L) + \log(1/\delta)}{n}}$$

If density of $\mathbf{P}_X$ is bounded above by $C/2$, then

$$R(T) \ \leq \ \hat{R}(T) + \sum_{L \in T} \sqrt{C 2^{-j(L)} \ \frac{3j(L) + \log(1/\delta)}{n}}$$

Spatial Adaptivity

Non-Adaptive Bound:

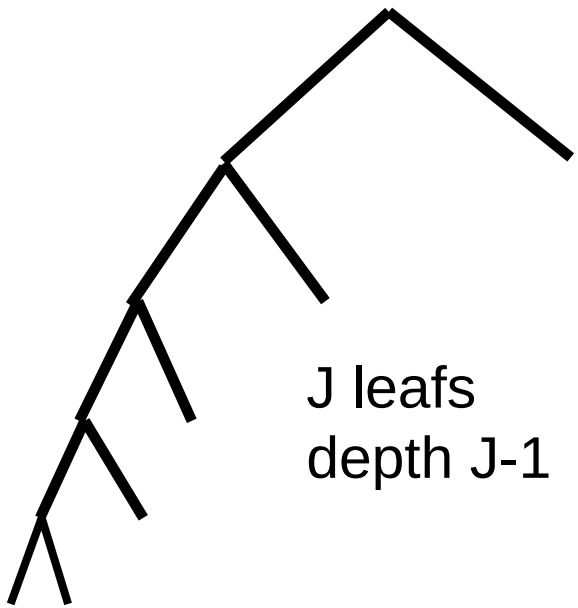
$$R(T) \leq \hat{R}(T) + \sqrt{\frac{3|T| + \log(1/\delta)}{2n}}$$

Spatially Adaptive:

$$R(T) \leq \hat{R}(T) + \sum_{L \in T} \sqrt{C 2^{-j(L)} \frac{3j(L) + \log(1/\delta)}{n}}$$

Key: local complexity is offset by small volumes!

Bound Comparison for Unbalanced Tree



Non-adaptive bound:

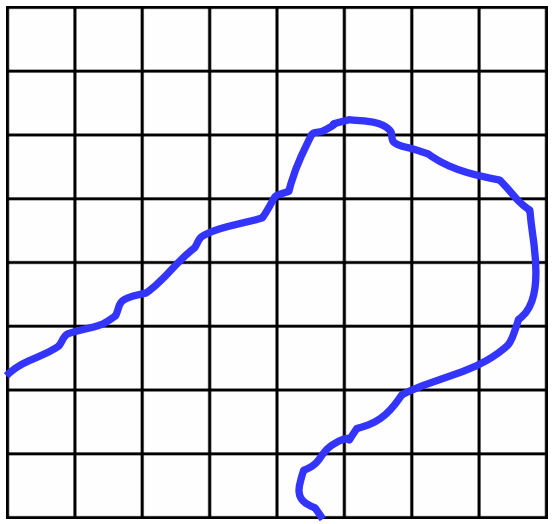
$$R(T) - \widehat{R}(T) = O\left(\sqrt{\frac{J + \log(1/\delta)}{n}}\right)$$

Adaptive bound:

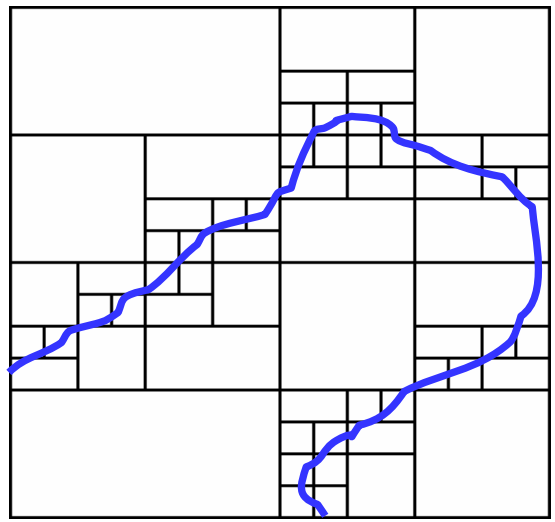
$$\begin{aligned} R(T) - \widehat{R}(T) &\leq 2\sum_{j=1}^{J}\sqrt{\frac{C2^{-j}(3j + \log(1/\delta))}{n}} \\ &= O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \end{aligned}$$

Balanced vs. Unbalanced Trees

same number of leafs



T_1



T_2

$$\sum_{L \in T_1} \sqrt{C 2^{-j(L)} \frac{3j(L) + \log(1/\delta)}{n}} \gg \sum_{L \in T_2} \sqrt{C 2^{-j(L)} \frac{3j(L) + \log(1/\delta)}{n}}$$

Decision Tree Selection

Set $\delta = 1/n$ and

$$T_n \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \sum_{L \in T} \sqrt{C 2^{-j(L)} \frac{3j(L) + \log(n)}{n}} \right\}$$

Oracle Bound:

$$\mathbf{E}\{R(T_n)\} - R(f^*) \leq$$

$$\min_{T \in \mathcal{T}} \left\{ R(T) - R(f^*) + \sum_{L \in T} \sqrt{C 2^{-j(L)} \frac{3j(L) + \log(n)}{n}} \right\}$$

Approximation Error

Estimation Error

Rate of Convergence

$$\inf_{\widehat{f}_n} \sup_{\mathbf{P} \in \mathcal{D}_{\mathsf{box}}(\kappa,d)} \left[\mathbf{E}\left\{ R(\widehat{f}_n) \right\} - R(f^*) \right] \asymp n^{-\kappa/(2\kappa+d-2)}$$

$$\begin{aligned}
&\mathbf{E}\left\{R(T_n)\right\} - R(f^*) \\
&\quad \leq \min_{T \in \mathcal{T}} \left\{ R(T) - R(f^*) + \sum_{L \in T} \sqrt{\frac{C 2^{-j(L)} [3j(L) + \log(n)]}{n}} \right\} \\
&\quad \leq C_1 m^{-\kappa} + C_2 \sqrt{m^{d-2} \log(n)/n} \\
&\quad \asymp \left(\frac{\log n}{n} \right)^{\kappa/(2\kappa+d-2)}
\end{aligned}$$

Computable Penalty

In practice bound C on density is unknown

Instead replace with empirical counterpart:

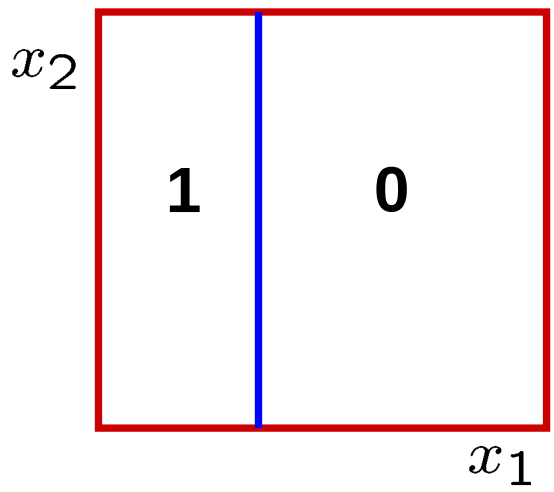
$$T_n \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \sum_{L \in T} \sqrt{\widehat{P}'_L \, \frac{3j(L) + \log(n)}{n}} \right\}$$

$$\widehat{P}'_L = 4 \max\left(\widehat{P}_L, \frac{3j(L) + \log(n)}{n} \right)$$

achieves same rate of convergence

$$\mathbf{E}\{R(T_n)\} - R(f^*) \preccurlyeq \left(\frac{\log n}{n} \right)^{\kappa/(2\kappa + d - 2)}$$

Adapting to Dimension - Feature Rejection

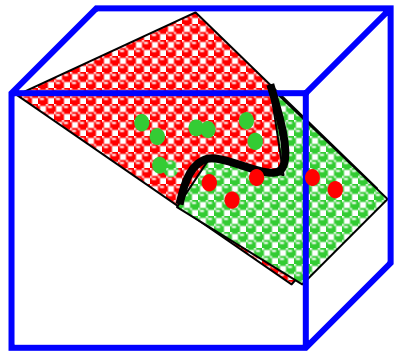


If $\mathbf{P} \in \mathcal{D}_{\mathsf{box}}(\kappa, d)$ and all but $d' \leq d$ features X^i are statistically independent of Y, then

$$\mathbf{E}\{R(T_n)\} - R(f^*) \preccurlyeq \left(\frac{\log n}{n}\right)^{\kappa/(2\kappa+d'-2)}$$

$\Rightarrow$ reduces to d' dimensional problem

Adapting to Dimension - Data Manifold



If support of $\mathbf{P} \in \mathcal{D}_{\mathsf{box}}(\kappa, d)$ is a manifold with box-counting dimension $d' < d$, then

$$\mathbf{E}\left\{R(\widehat{T}_n)\right\} - R(f^*) \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{(2\kappa+d'-2)}}$$

$\Rightarrow$ reduces to d' dimensional problem

Computational Issues

additive penalty

$$T_n \equiv \arg\min_{T \in \mathcal{T}} \left\{ \widehat{R}(T) + \sum_{L \in T} \sqrt{\widehat{P}'_L \, \frac{3j(L) + \log(n)}{n}} \right\}$$

Cyclic DDT: force coordinate splits in cyclic order

Scott and Nowak (2002): Simple bottom-up CART-like pruning procedure

T_n computed in $O(nd \log(n))$ ops

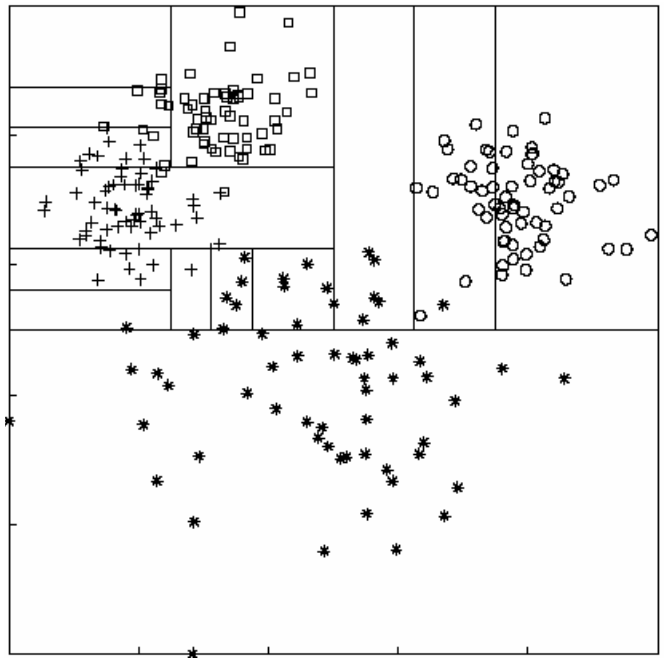
Free-Split DDT: no order enforcement in splits

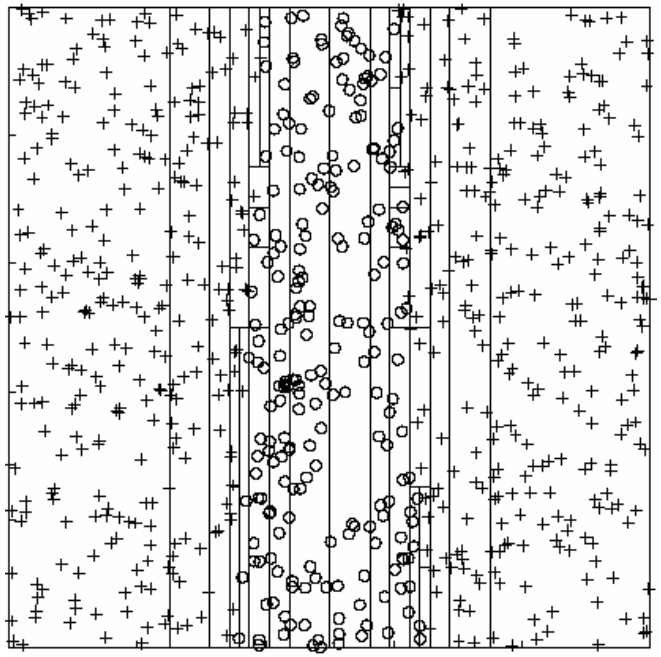
Blanchard et al. (2004): Modified version of Donoho (1997) that exploits data sparsity

T_n computed in $O(ndJ^d \log(nJ^d))$ ops

where J max # splits on each coordinate

DDTs in Action

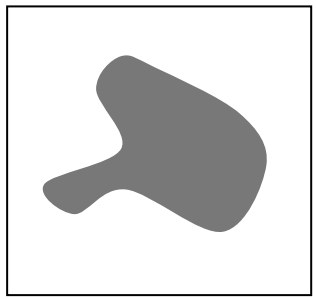


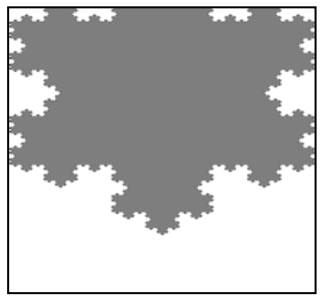


Schäfer, Blanchard, Rozenholc and Müller (2004)

Decision Boundary Smoothness

Box-counting class essentially requires **Lipschitz** smoothness

$\gamma = 1$ 

$\gamma < 1$ 

If Bayes boundary has regularity $\gamma \leq 1$ then

$$\mathbf{E}\{R(T_n)\} - R(f^*) \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+(d-1)/\gamma-1}}$$

Smooth Curves

Bayes boundary is Hölder smooth with $\gamma > 1$

$$\mathcal{D}(\gamma, \kappa, d) \subset \mathcal{D}_{\text{box}}(\kappa, d)$$

van de Geer & Tsybakov (2004)

assume smooth boundary is a "fragment" (function); impractical, high-dim, non-convex optimization over wavelet approximation

Scott & Nowak (2003)

polynomial-decorated trees; potentially practical implemented via SVMs; suboptimal rates

$$\mathbf{E}\left\{R(T_n^{\text{SVM}})\right\} - R(f^*) \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\gamma}{2\gamma+d-2}}$$

Summary

minimax lower bound

$$\inf_{\widehat{f}_n} \sup_{\mathbf{P}\in\mathcal{D}_{\text{box}}(\kappa,\rho^*)} \left[\mathbf{E}\left\{R(\widehat{f}_n)\right\} - R(f^*)\right] \asymp n^{-\kappa/(2\kappa+\rho^*-1)}$$

$$\gamma \leq 1\,, \quad \kappa \geq 1\,, \quad \rho^* = (d^*-1)/\gamma$$

d^* = dim of manifold supporting "informative" features

upper bound for DDT

$$T_n \equiv \arg\min_{T\in\mathcal{T}} \left\{\widehat{R}(T) + \sum_{L\in T} \sqrt{\widehat{P}'_L \, \frac{3j(L)+\log(n)}{n}}\right\}$$

$$\sup_{\mathbf{P}\in\mathcal{D}_{\text{box}}(\kappa,\rho^*)} \left[\mathbf{E}\left\{R(T_n)\right\} - R(f^*)\right] \preccurlyeq \left(\frac{\log n}{n}\right)^{\kappa/(2\kappa+\rho^*-1)}$$

Conclusions and Future Work

Dyadic decision trees automatically adapt to

- boundary smoothness $\gamma \leq 1$
- transition smoothness $\kappa \geq 1$
- feature distribution manifold
- "informative" feature dimensions

Tech Report:

www.ece.wisc.edu/~nowak/ddt.pdf

Open Problem:

Practical schemes achieving minimax optimal rates $n^{-\kappa/(2\kappa+(d-1)/\gamma-1)}$ when $\gamma > 1$