

Statistical Learning Theory

Part 1: Estimation Error and Approximation Error

Cynthia Rudin

Duke

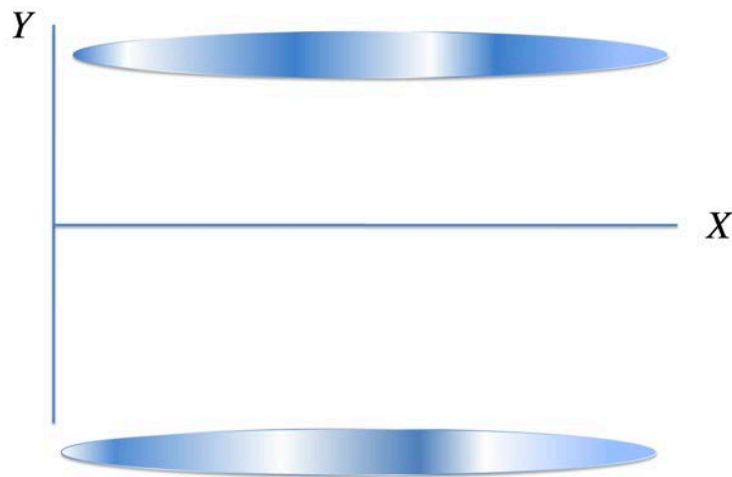


Generalization = Data + Knowledge

credits: Bousquet, Boucheron, Lugosi

$\{(x_i, y_i)\}_{i=1}^n$ drawn iid from D on $\mathcal{X} \times \mathcal{Y}$

\mathcal{X} $\mathcal{Y} = \{-1, 1\}$



Goal:

$f : \mathcal{X} \rightarrow \mathcal{Y}$ that predicts y from x

$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

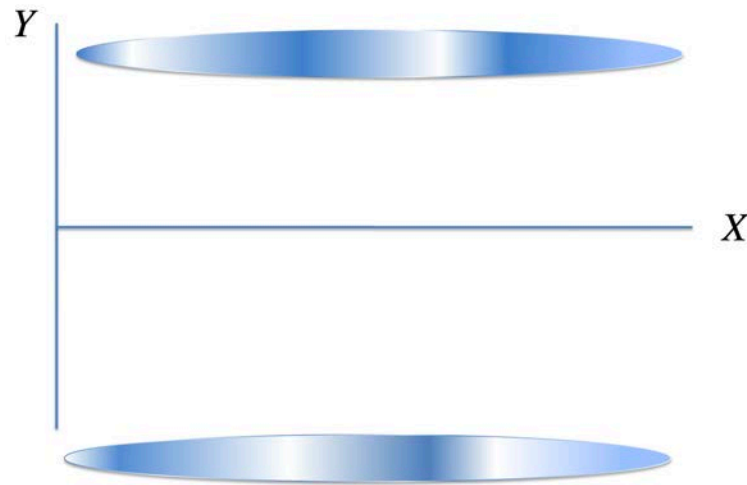
Z **blah** **blah** **drivel** **blah** **drool** **drip** **blah**

1 1 0 1 0 0 1

$$\begin{aligned}\mathbb{E}_{Z \sim D} \mathbf{1}_{[Z=\text{blah}]} &= 1 \times P(\text{outcome 1, when } Z=\text{blah}) \\ &\quad + 0 \times P(\text{outcome 0, when } Z \neq \text{blah}) = 1*(4/7) + 0 \\ &= \mathbb{P}_{Z \sim D}(Z = \text{blah}). = 1*(4/7)\end{aligned}$$

$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

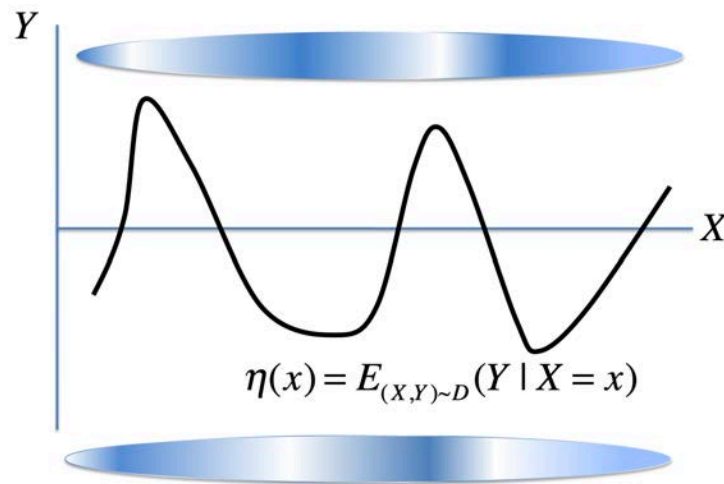
Regression function $\eta(x) = \mathbb{E}_{(X,Y) \sim D}(Y|X = x)$



$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

Regression function $\eta(x) = \mathbb{E}_{(X,Y) \sim D}(Y | X = x)$

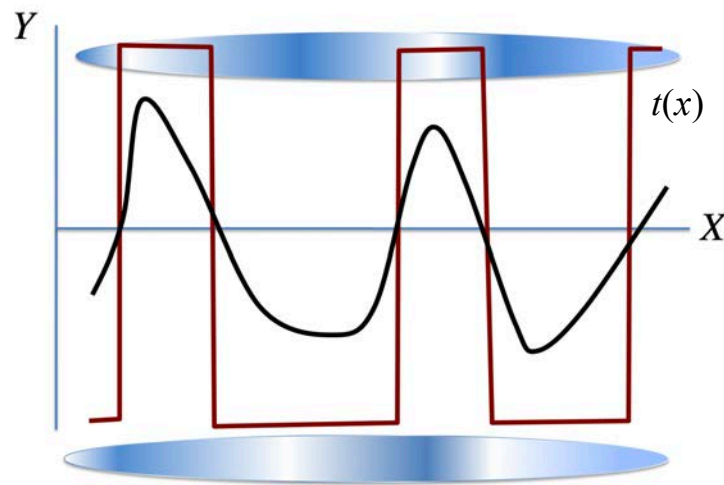
Target function (Bayes classifier) $t(x) = \text{sign } \eta(x)$



$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

Regression function $\eta(x) = \mathbb{E}_{(X,Y) \sim D}(Y|X = x)$

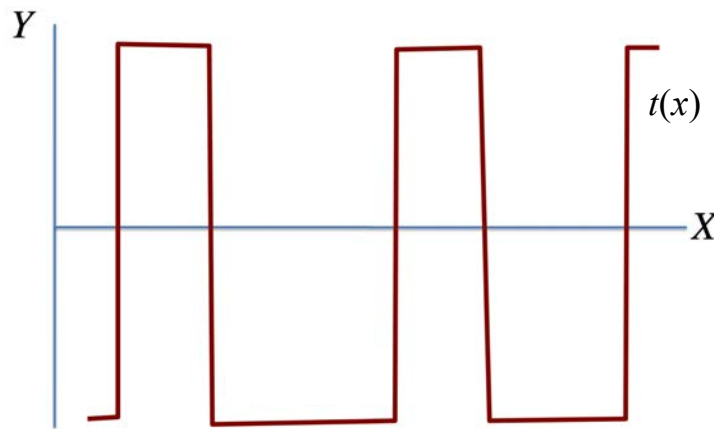
Target function (Bayes classifier) $t(x) = \text{sign } \eta(x)$



$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

Regression function $\eta(x) = \mathbb{E}_{(X,Y) \sim D}(Y|X = x)$

Target function (Bayes classifier) $t(x) = \text{sign } \eta(x)$



$$R^{\text{true}}(t) = \inf_f R^{\text{true}}(f) = R^* \quad \text{Bayes Risk}$$

$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$

$$R^{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{[f(x_i) \neq y_i]}$$

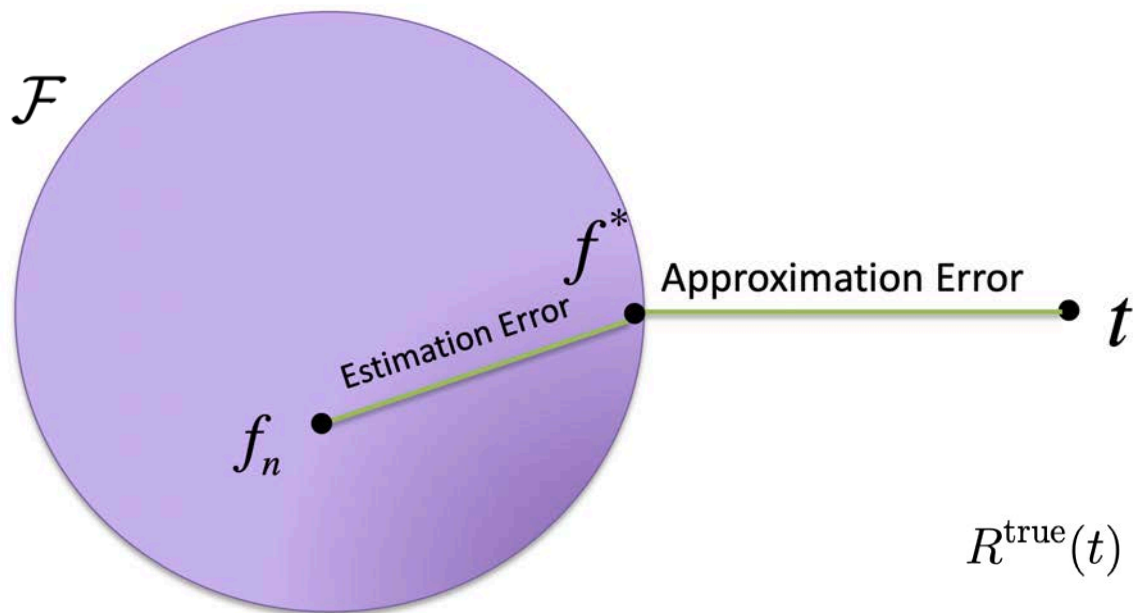
$$f_n \in \operatorname{argmin}_{f \in \mathcal{F}} R^{\text{emp}}(f) + C \|f\|^2$$

Best in practice

$$R^{\text{true}}(f^*) = \inf_{f \in \mathcal{F}} R^{\text{true}}(f) \quad \text{Best in class}$$

$$R^{\text{true}}(t) = \inf_f R^{\text{true}}(f) = R^* \quad \text{Bayes Risk} \quad \text{Best in theory}$$

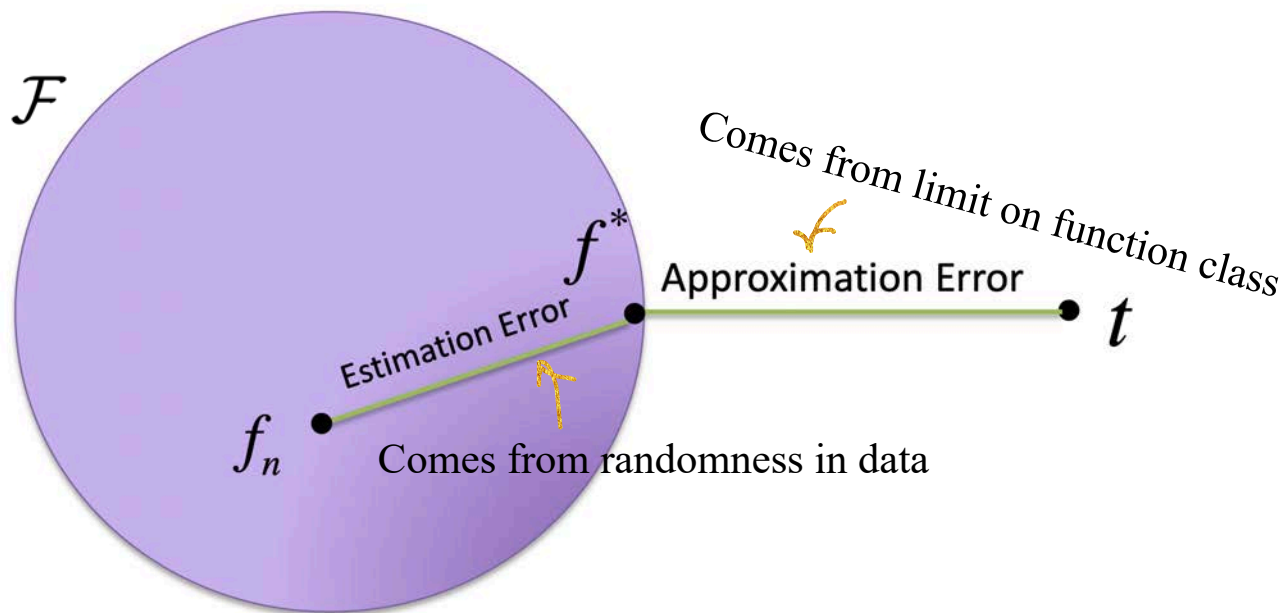
$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y) \sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim D}[\mathbf{1}_{f(X) \neq Y}]$$



$$R^{\text{true}}(t) = \inf_f R^{\text{true}}(f)$$

$$R^{\text{true}}(f^*) = \inf_{f \in \mathcal{F}} R^{\text{true}}(f)$$

$$f_n \in \operatorname{argmin}_{f \in \mathcal{F}} R^{\text{emp}}(f) + C\|f\|^2$$



Test error of (Best in practice) – Test error of (Best in theory)

$$\begin{aligned}
 \text{Test error of (Best in practice)} - \text{Test error of (Best in theory)} &= R^{\text{true}}(f_n) - R^* \\
 &= \underbrace{[R^{\text{true}}(f^*) - R^*]}_{\text{Approximation Error}} + \underbrace{[R^{\text{true}}(f_n) - R^{\text{true}}(f^*)]}_{\text{Estimation Error}}
 \end{aligned}$$

$$R^{\text{true}}(f_n) = R^{\text{emp}}(f_n) + \underbrace{[R^{\text{true}}(f_n) - R^{\text{emp}}(f_n)]}_{\substack{\backslash \wedge \\ \text{Stuff}(n, \mathcal{F})}}$$

What comes next:

A bound for a single f

The reason why a bound for a single f is no good.

The Ockham's Razor Bound.

The VC Bound.

Statistical Learning Theory

Part 2

A bound for a single function

Cynthia Rudin

Duke

For each f in \mathcal{F} create a loss function g :

✦ $g(x, y) = \mathbf{1}_{f(x) \neq y}$

g is 1 if f makes a mistake on (x, y)

✦ $\mathcal{G} = \{g : (x, y) \rightarrow \mathbf{1}_{f(x) \neq y} : f \in \mathcal{F}\}$

There is a bijection between \mathcal{F} and \mathcal{G}

If $g(x, y) = 1$ set $f(x) = -y$, otherwise set $f(x) = y$

$$P^{\text{true}}g = \mathbb{E}_{(X, Y) \sim D}[g(X, Y)] \quad (\text{true risk again})$$

$$P^{\text{emp}}g = \frac{1}{n} \sum_{i=1}^n g(X_i, Y_i) \quad (\text{empirical risk again})$$

Want: $P^{\text{true}}g_n - P^{\text{emp}}g_n$

Change notation yet again: $Z = (X, Y)$
 $Z_i = (X_i, Y_i)$

$$\mathbf{Z} \sim D^n$$

In the new notation:

$$P^{\text{true}}g - P^{\text{emp}}g = \mathbb{E}_{\mathbf{Z} \sim D^n}[g(Z)] - \frac{1}{n} \sum_{i=1}^n g(Z_i)$$

As $n \rightarrow \infty$, the above approaches 0. But we can do better.

Theorem (Hoeffding's Inequality). *Let $Z_1 \dots Z_n$ be n iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:*

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[\left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}_{\mathbf{Z} \sim D^n}[h(Z)] \right| > \epsilon \right] \leq 2 \exp \left(-\frac{2n\epsilon^2}{(b-a)^2} \right).$$

$$\delta = 2 \exp \left(-\frac{2n\epsilon^2}{(b-a)^2} \right) \quad \longrightarrow \quad \epsilon = (b-a) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Plug in, and apply Hoeffding's to P 's:

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[|P^{\text{emp}} g - P^{\text{true}} g| > (b-a) \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \right] \leq \delta.$$

“2-sided Hoeffding's”

Theorem (Hoeffding's Inequality). *Let $Z_1 \dots Z_n$ be n iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:*

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[\left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}_{\mathbf{Z} \sim D^n} [h(Z)] \right| > \epsilon \right] \leq 2 \exp \left(-\frac{2n\epsilon^2}{(b-a)^2} \right).$$

After “inversion”:

with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathbf{Z} \sim D^n}[h(Z)] - \frac{1}{n} \sum_{i=1}^n h(Z_i) \leq (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

$$P^{\text{true}}g - P^{\text{emp}}g \leq (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

“1-sided Hoeffding’s”

Theorem (Hoeffding’s Inequality). *Let $Z_1 \dots Z_n$ be n iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:*

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[\mathbb{E}_{\mathbf{Z} \sim D^n}[h(Z)] - \frac{1}{n} \sum_{i=1}^n h(Z_i) > \epsilon \right] \leq \exp \left(-\frac{2n\epsilon^2}{(b - a)^2} \right) = \delta$$

After “inversion”:

with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathbf{Z} \sim D^n}[h(Z)] - \frac{1}{n} \sum_{i=1}^n h(Z_i) \leq (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

$$P^{\text{true}}g - P^{\text{emp}}g \leq (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

with probability at least $1 - \delta$,

$$P^{\text{true}}g \leq P^{\text{emp}}g + (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

Want small

Is small

If n big, is small

- Interestingly, the bound doesn't apply when f comes from any reasonable learning algorithm!

with probability at least $1 - \delta$,

$$P^{\text{true}}g \leq P^{\text{emp}}g + (b - a)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

↑
Want small

↑
Is small

↑
If n big, is small

Statistical Learning Theory

Part 3

Why the bound doesn't work

Cynthia Rudin

Duke

Interestingly, the bound doesn't apply when f comes from any reasonable learning algorithm!

with probability at least $1 - \delta$,

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

↑
Want small

↑
Is small

↑
If n big, is small

For f , chosen in advance (without knowledge of the data),

with probability at least $1 - \delta$,

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$



Want small



Is small



If n big, is small

with probability at least $1 - \delta$,

For f_n

$$R^{\text{true}}(f_n) \leq R^{\text{emp}}(f_n) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Want

For f , chosen in advance (without knowledge of the data),

with probability at least $1 - \delta$,

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$



Want small



Is small



If n big, is small

with probability at least $1 - \delta$,

For all f in \mathcal{F}

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Want

For f , chosen in advance (without knowledge of the data),

with probability at least $1 - \delta$,

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$



Want small



Is small



If n big, is small

For each fixed function $g \in \mathcal{G}$, there is a large set of “good” datasets where

$$P^{\text{true}}g - P^{\text{emp}}g \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

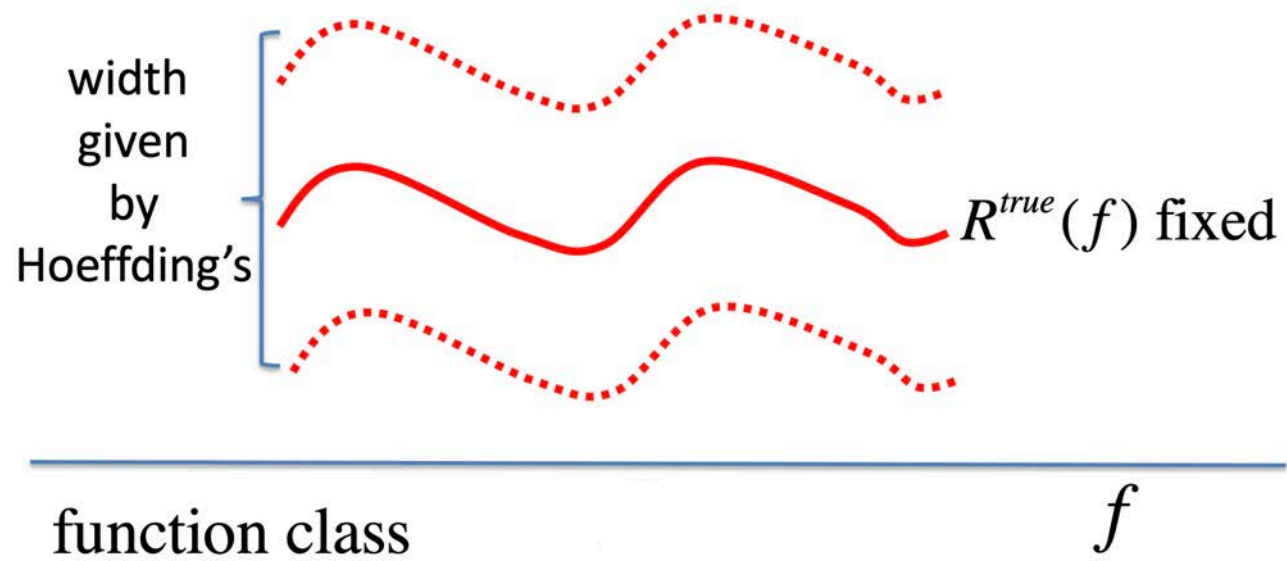
this set of datasets has measure $\mathbb{P}_{\mathbf{Z} \sim D^n}[\mathbf{Z} \in S] \geq 1 - \delta$.

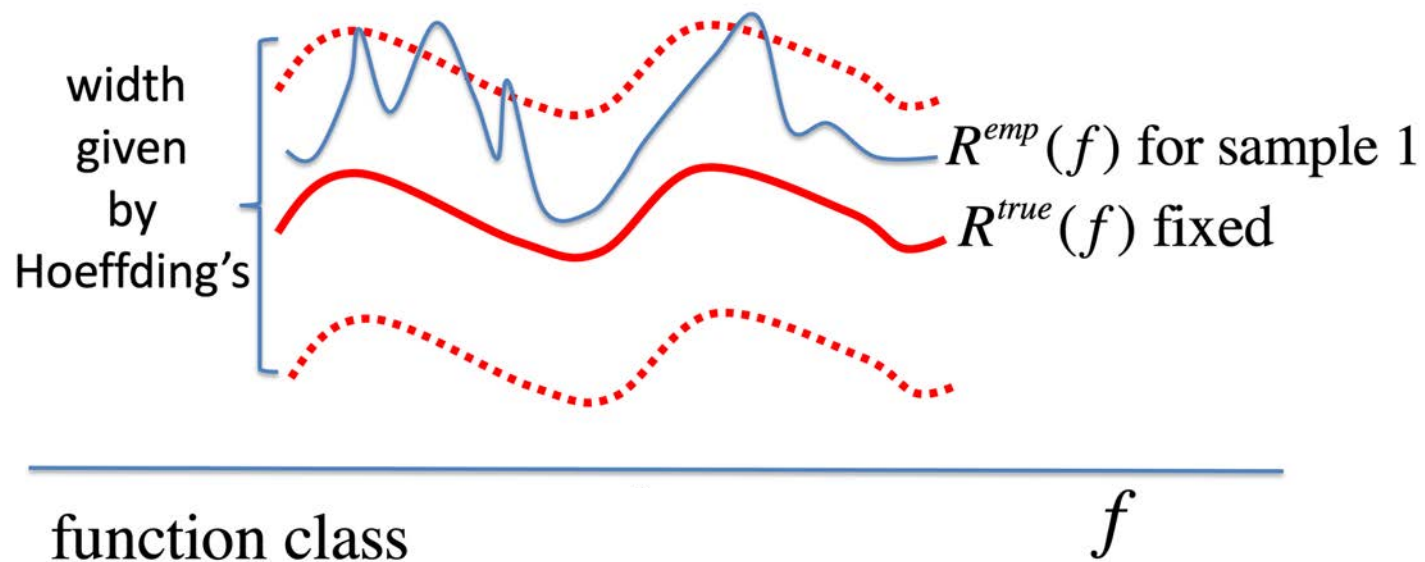
But this set can be different for different g !

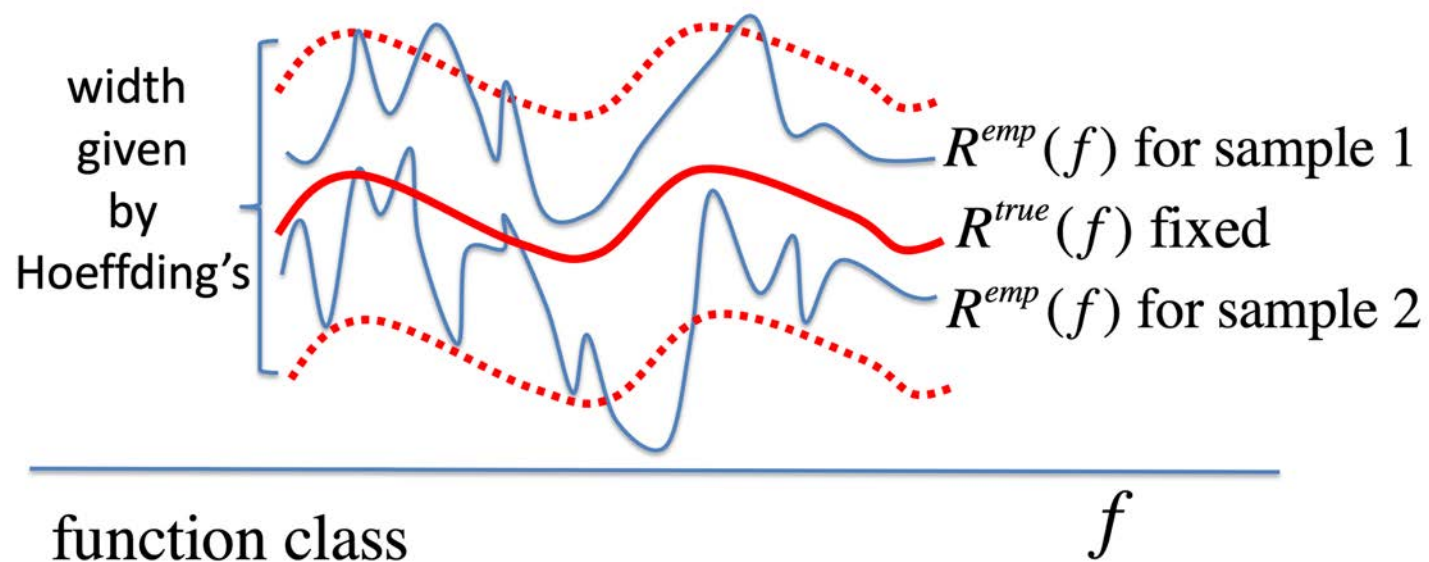
 $R^{true}(f)$ fixed

function class

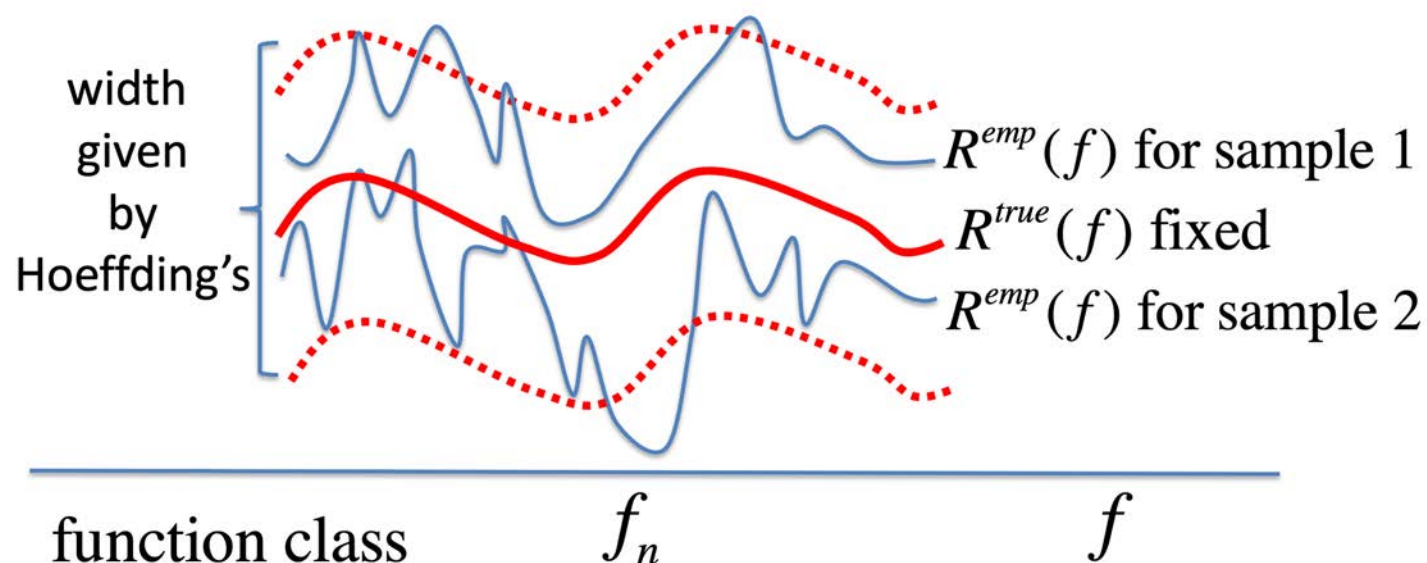
f







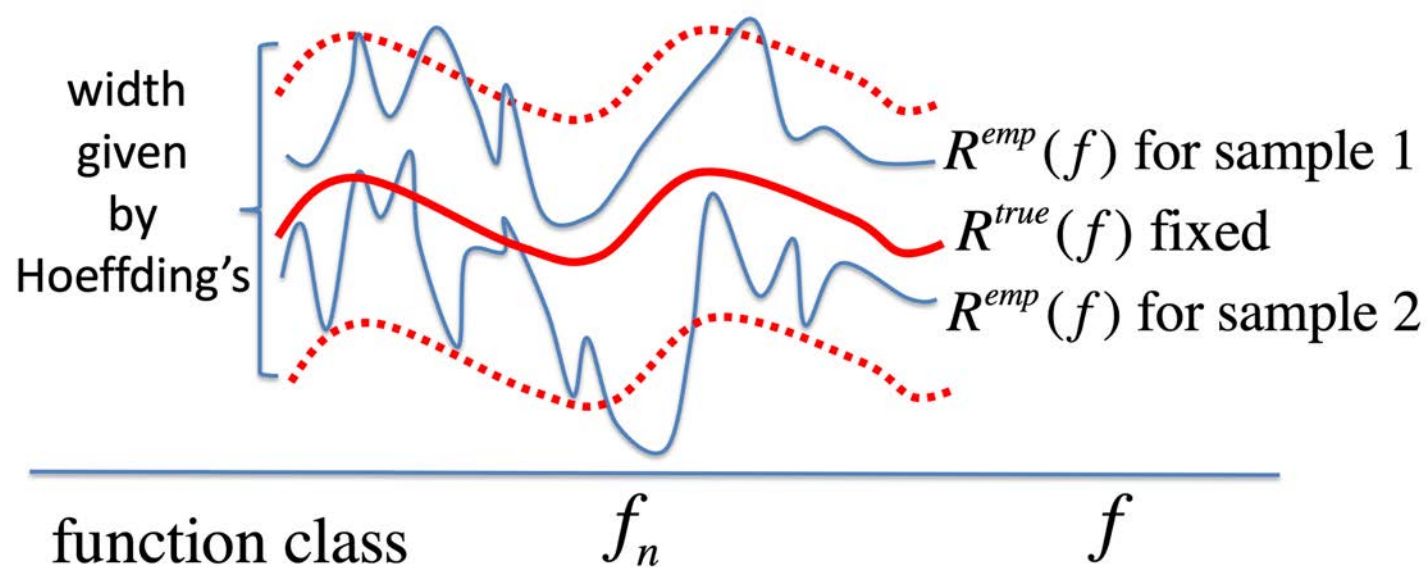
If \mathcal{F} is large, there are more opportunities to find an f where distance between $R^{\text{emp}}(f)$ and $R^{\text{true}}(f)$ is large.



We need to make sure this doesn't happen!

Solution: want that w.h.p., $R^{\text{emp}}(f)$ and $R^{\text{true}}(f)$ are close for all f in \mathcal{F} .

Uniform bounds



Solution: want that w.h.p., $R^{emp}(f)$ and $R^{true}(f)$ are close for all f in \mathcal{F} .

Statistical Learning Theory

Part 4: The Ockham's Razor Bound

Cynthia Rudin

Duke

Uniform deviations

$$R^{\text{true}}(f_n) - R^{\text{emp}}(f_n) \leq \sup_{f \in \mathcal{F}} (R^{\text{true}}(f) - R^{\text{emp}}(f))$$

\mathcal{F} (or equivalently \mathcal{G}) is finite

g_j is the j th function in \mathcal{G}

$$C_j^{(\text{bad})} = \{ \text{“bad” datasets } \mathbf{Z} \text{ for } g_j \text{ where } P^{\text{true}}g_j - P^{\text{emp}}g_j \geq \epsilon \}$$

From Hoeffding's Inequality

For all j , with prob at least $1 - \delta$, $\mathbb{P}_{\mathbf{Z} \sim D^n}[\mathbf{Z} \in C_j^{(\text{bad})}] \leq \delta$

$$\mathbb{P}_{\mathbf{Z} \sim D^n} [\mathbf{Z} \in C_1^{(\text{bad})} \cup C_2^{(\text{bad})}] \quad \text{Union Bound}$$

$$\leq \mathbb{P}_{\mathbf{Z} \sim D^n} [\mathbf{Z} \in C_1^{(\text{bad})}] + \mathbb{P}_{\mathbf{Z} \sim D^n} [\mathbf{Z} \in C_2^{(\text{bad})}] \leq 2\delta$$

the probability that we hit a bad dataset for either g_1 or g_2 is
prob to hit a bad dataset in $C_1^{(\text{bad})} \leq$ prob to hit a bad dataset in $C_2^{(\text{bad})}$.

From Hoeffding's Inequality

For all j , with prob at least $1 - \delta$, $\mathbb{P}_{\mathbf{Z} \sim D^n} [\mathbf{Z} \in C_j^{(\text{bad})}] \leq \delta$.

Union Bound

$$\mathbb{P}[C_1^{(\text{bad})} \cup \dots \cup C_M^{(\text{bad})}] \leq \sum_{j=1}^M \mathbb{P}[C_j^{(\text{bad})}] \leq M\delta$$

Prob that our (random) dataset is bad for any of the functions g_1, \dots, g_M

\leq

Sum of probs that our (random) dataset is bad for each of the functions

From Hoeffding's Inequality

For all j , with prob at least $1 - \delta$, $\mathbb{P}_{\mathbf{Z} \sim D^n}[\mathbf{Z} \in C_j^{(\text{bad})}] \leq \delta$

$$\mathbb{P}_{\mathbf{Z} \sim D^n} [\exists g \in \{g_1, \dots, g_M\} : P^{\text{true}} g - P^{\text{emp}} g \geq \epsilon]$$

(prob there's a bad data for some function)

$$\leq \sum_{j=1}^M \mathbb{P}_{\mathbf{Z} \sim D^n} [P^{\text{true}} g_j - P^{\text{emp}} g_j \geq \epsilon]$$

(from Union Bound)

$$\leq \sum_{j=1}^M \exp(-2n\epsilon^2) \quad \text{(from Hoeffding's Inequality)}$$

$$= M \exp(-2n\epsilon^2).$$

$$\mathbb{P}_{\mathbf{Z} \sim D^n} [\exists g \in \{g_1, \dots, g_M\} : P^{\text{true}} g - P^{\text{emp}} g \geq \epsilon]$$

$$\leq M \exp(-2n\epsilon^2) =: \delta$$

$$\epsilon = \sqrt{\frac{\log M + \log \frac{1}{\delta}}{2n}}$$

Invert:



with probability at least $1 - \delta$,

$$\forall g \in \{g_1, \dots, g_M\} : P^{\text{true}} g - P^{\text{emp}} g \leq \sqrt{\frac{\log M + \log \frac{1}{\delta}}{2n}}.$$



Replacing g with f , we have the main result.

- Holds no matter which function f our algorithm chooses.
- Says that as long as our hypothesis space isn't too big, we obtain knowledge about the true risk.
- *logarithmic* in M
- Applies to finite hypothesis spaces
 - E.g., decision trees over binary/categorical variables
 - E.g., linear models with integer coefficients
 - E.g., neural networks with integer weights
- Infinite hypothesis spaces coming soon!

Theorem. (Ockham's Razor = Hoeffding + Union Bound)

This bound applies for finite \mathcal{F} , so $\mathcal{F} = \{f_1 \dots f_M\}$. For all $\delta > 0$ with probability at least $1 - \delta$,

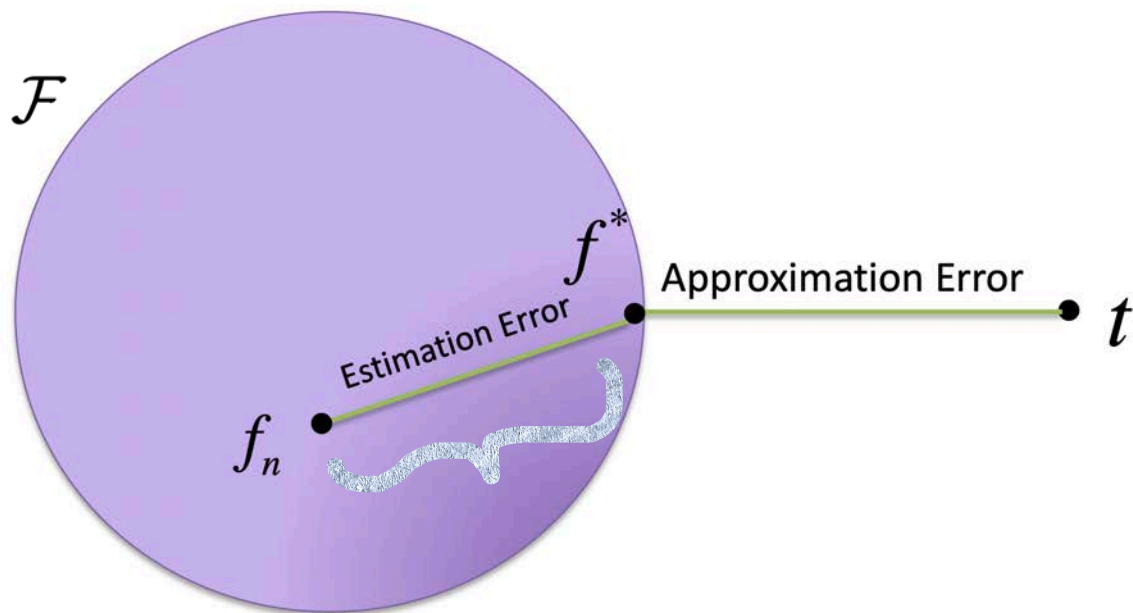
$$\forall f \in \mathcal{F}, \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log M + \log \frac{1}{\delta}}{2n}}.$$

Statistical Learning Theory

Part 5: Back to Estimation Error

Cynthia Rudin

Duke




Assuming $f_n \in \arg \min_{f \in \mathcal{F}} R^{\text{emp}}(f)$, we can use Ockham's Razor to bound the estimation error.

$$R^{\text{emp}}(f^*) - R^{\text{emp}}(f_n) \stackrel{\star}{\geq} 0$$

$$R^{\text{true}}(f_n) = R^{\text{true}}(f_n) - \overbrace{R^{\text{true}}(f^*)} + \overbrace{R^{\text{true}}(f^*)}$$

$$\begin{aligned}
R^{\text{true}}(f_n) &\leq \overbrace{[R^{\text{emp}}(f^*) - R^{\text{emp}}(f_n)]}^{\star} + R^{\text{true}}(f_n) - \overbrace{R^{\text{true}}(f^*)} + \overbrace{R^{\text{true}}(f^*)} \\
&= R^{\text{emp}}(f^*) - \overbrace{R^{\text{true}}(f^*)} - R^{\text{emp}}(f_n) + R^{\text{true}}(f_n) + \overbrace{R^{\text{true}}(f^*)} \\
&\leq \overbrace{|R^{\text{true}}(f^*) - R^{\text{emp}}(f^*)|} + \overbrace{|R^{\text{true}}(f_n) - R^{\text{emp}}(f_n)|} + \overbrace{R^{\text{true}}(f^*)} \\
&\leq 2 \sup_{f \in \mathcal{F}} |R^{\text{true}}(f) - R^{\text{emp}}(f)| + \overbrace{R^{\text{true}}(f^*)}.
\end{aligned}$$

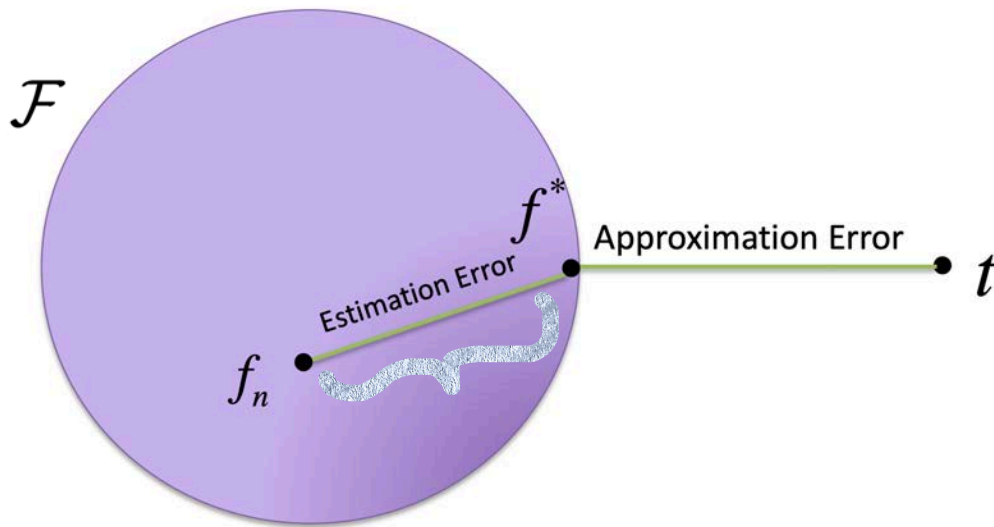
with probability at least $1 - \delta$:

$$R^{\text{true}}(f_n) \leq 2\sqrt{\frac{\log M + \log \frac{2}{\delta}}{2n}} + R^{\text{true}}(f^*)$$


$$R^{\text{true}}(f_n) \leq 2 \sup_{f \in \mathcal{F}} |R^{\text{true}}(f) - R^{\text{emp}}(f)| + \overbrace{R^{\text{true}}(f^*)}.$$

with probability at least $1 - \delta$:

$$R^{\text{true}}(f_n) \leq 2\sqrt{\frac{\log M + \log \frac{2}{\delta}}{2n}} + R^{\text{true}}(f^*)$$



Statistical Learning Theory

Part 6: Summary & perspective so far

Cynthia Rudin

Duke

- Generalization = data + knowledge
(like restricting to a class of functions \mathcal{F}).
- For a fixed function f , with high probability,

$$R^{\text{true}}(f) - R^{\text{emp}}(f) \approx 1/\sqrt{n}.$$

- Generalization = data + knowledge
(like restricting to a class of functions \mathcal{F}).
- For a fixed function f , with high probability,

$$R^{\text{true}}(f) - R^{\text{emp}}(f) \approx 1/\sqrt{n}.$$

- For if the function class is finite, $|\mathcal{F}| = M$, with high probability,

$$\sup_{f \in \mathcal{F}} [R^{\text{true}}(f) - R^{\text{emp}}(f)] \approx \sqrt{\log M/n}.$$

the extra term is because we want to choose f_n in a way that depends on data.

- The union bound is in general loose, because it is as bad as if all the $f_j(Z)$'s are independent
- The bound is vacuous when there are an infinite number of functions in \mathcal{F} .

$$\sup_{f \in \mathcal{F}} [R^{\text{true}}(f) - R^{\text{emp}}(f)] \approx \sqrt{\log M/n}.$$

Statistical Learning Theory

Part 7: The case of countably infinite functions

Cynthia Rudin

Duke

- Let's extend the Ockham's Razor bound to the countably infinite case.
- Recall the 1-sided Hoeffding's Inequality. For any g ,

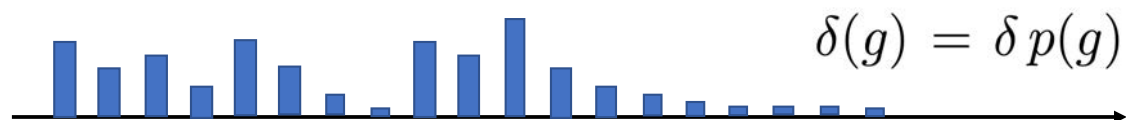
$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[P^{\text{true}}_g - P^{\text{emp}}_g \geq \sqrt{\frac{\log \frac{1}{\delta(g)}}{2n}} \right] \leq \delta(g)$$

This time, make δ depend on g .

If we have a countably infinite \mathcal{G} , the union bound gives:

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[\exists g \in \mathcal{G} : P^{\text{true}}_g - P^{\text{emp}}_g \geq \sqrt{\frac{\log \frac{1}{\delta(g)}}{2n}} \right] \leq \sum_{g \in \mathcal{G}} \delta(g)$$

- This creates a probability distribution over g 's.



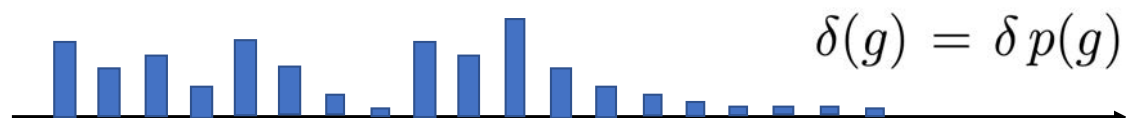
After doing inversion,

with probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, P^{\text{true}} g \leq P^{\text{emp}} g + \sqrt{\frac{\log \frac{1}{p(g)} + \log \frac{1}{\delta}}{2n}}$$

$$\mathbb{P}_{\mathbf{Z} \sim D^n} \left[\exists g \in \mathcal{G} : P^{\text{true}} g - P^{\text{emp}} g \geq \sqrt{\frac{\log \frac{1}{\delta(g)}}{2n}} \right] \leq \sum_{g \in \mathcal{G}} \delta(g) =: \delta$$

- This creates a probability distribution over g 's.



After doing inversion,

with probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, P^{\text{true}} g \leq P^{\text{emp}} g + \sqrt{\frac{\log \frac{1}{p(g)} + \log \frac{1}{\delta}}{2n}}$$

- Note: if \mathcal{G} is finite of size M , and $p(g) = \frac{1}{M}$, we get back to the $\log(M)$ term and the Ockham's Razor Bound.
- Does not work if any of the $p(g)$'s are 0.

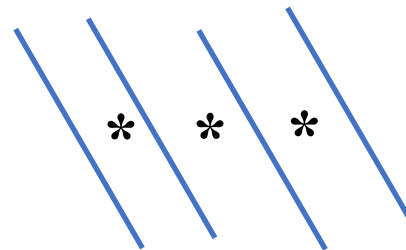
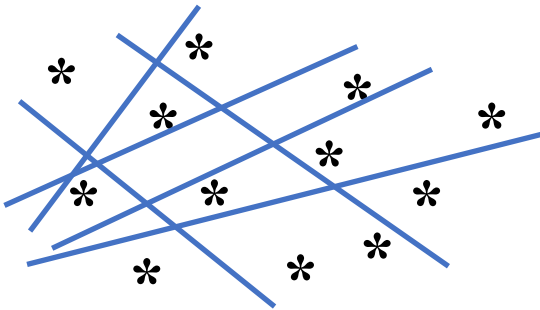
Statistical Learning Theory

Part 8: The growth function

Cynthia Rudin

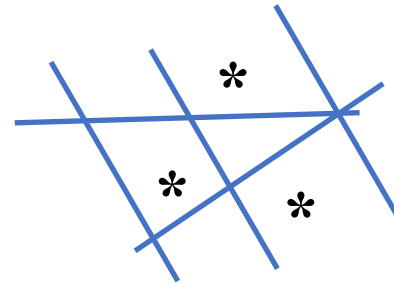
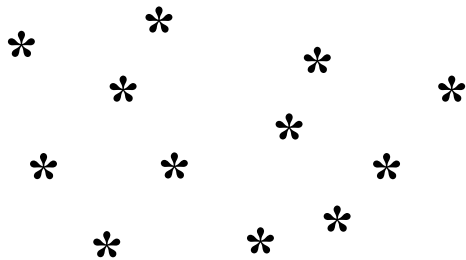
Duke

- Q. How do we reduce an infinite number of functions into a finite number of classifiers?
- A. Look at how the functions classify the data.



- Under some configurations of the data, we can classify in more ways.

- Q. How do we reduce an infinite number of functions into a finite number of classifiers?
- A. Look at how the functions classify the data.



- Under some configurations of the data, we can classify in more ways.

Given z_1, \dots, z_n , we consider

$$\mathcal{F}_{z_1, \dots, z_n} = \{f(z_1), \dots, f(z_n) : f \in \mathcal{F}\}.$$

↪ set of ways the data z_1, \dots, z_n can be classified by functions from \mathcal{F} .

Definition: The *growth function* of function class \mathcal{F} is the maximum number of ways into which n points can be classified by the function class.



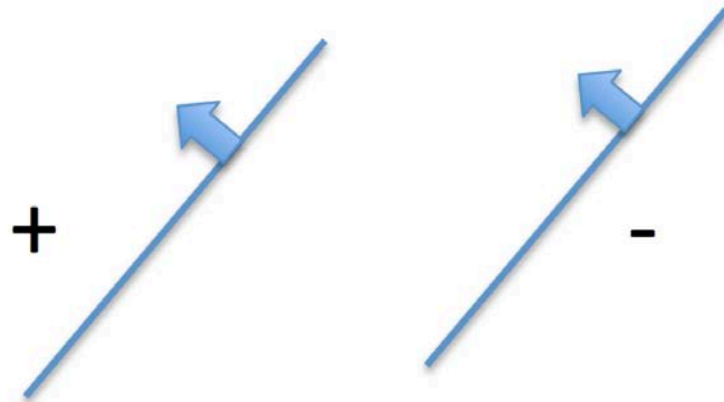
$$S_{\mathcal{F}}(n) = \sup_{(z_1, \dots, z_n)} |\mathcal{F}_{z_1, \dots, z_n}|$$

Examples

Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

$$S_{\mathcal{F}}(1) = 2 = 2^1$$



Examples

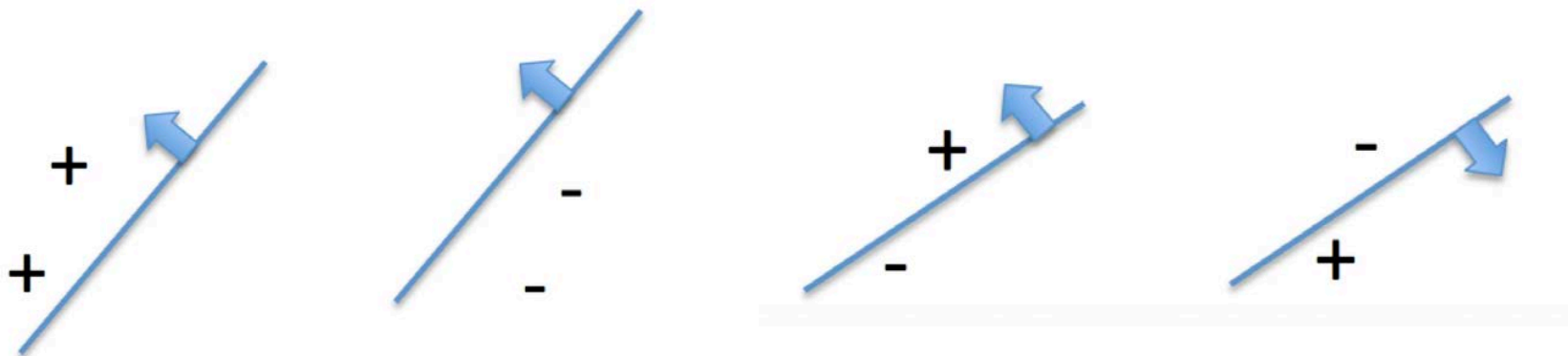
Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

$$S_{\mathcal{F}}(1) = 2 = 2^1$$

The growth function for two points is

$$S_{\mathcal{F}}(2) = 4 = 2^2$$



Examples

Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

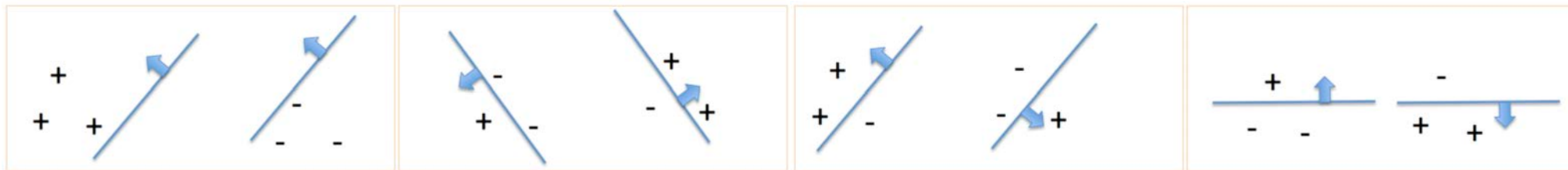
$$S_{\mathcal{F}}(1) = 2 = 2^1$$

The growth function for two points is

$$S_{\mathcal{F}}(2) = 4 = 2^2$$

The growth function for three points is

$$S_{\mathcal{F}}(3) = 8 = 2^3$$



Examples

Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

$$S_{\mathcal{F}}(1) = 2 = 2^1 \leftarrow$$

The growth function for two points is

$$S_{\mathcal{F}}(2) = 4 = 2^2 \leftarrow$$

The growth function for three points is

$$S_{\mathcal{F}}(3) = 8 = 2^3 \leftarrow$$

The growth function for four points is

???

-	+
+	-

It is less than 2^4 .

Examples

Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

$$S_{\mathcal{F}}(1) = 2 = 2^1 \leftarrow$$

The growth function for two points is

$$S_{\mathcal{F}}(2) = 4 = 2^2 \leftarrow$$

The growth function for three points is

$$S_{\mathcal{F}}(3) = 8 = 2^3 \leftarrow$$

The growth function for four points is

It is less than 2^4 .

The growth function can be used to measure the “capacity” of a set of functions.



Theorem-GrowthFunction (Vapnik-Chervonenkis) *For any $\delta > 0$, with probability at least $1 - \delta$ with respect to a random draw of the data,*

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}.$$

This theorem is non-vacuous for lines in the plane.

This theorem is non-vacuous for infinite function spaces.

In the finite case, strictly better than Ockham's Razor.

(except for constants)

$$S_{\mathcal{F}}(n) \leq M$$

Theorem-GrowthFunction (Vapnik-Chervonenkis) *For any $\delta > 0$, with probability at least $1 - \delta$ with respect to a random draw of the data,*

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}.$$

Proof ingredients:

- Symmetrization Lemma
 - Create a “ghost” sample.
 - Bound the difference between the behavior on one dataset versus another.
 - This gives us a bound on the behavior of a dataset with respect to the true risk.
- Hoeffding’s Inequality
- Union Bound
- Chebyshev’s Inequality

Statistical Learning Theory

Part 9: Definition of the VC dimension

Cynthia Rudin

Duke

Have you noticed that this bound is not easy to compute?!

Theorem-GrowthFunction (Vapnik-Chervonenkis) *For any $\delta > 0$, with probability at least $1 - \delta$ with respect to a random draw of the data,*

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}.$$

Halfspaces in 2D - binary functions whose decision boundary is a line.

The growth function for one point is

$$S_{\mathcal{F}}(1) = 2 = 2^1$$

The growth function for two points is

$$S_{\mathcal{F}}(2) = 4 = 2^2$$

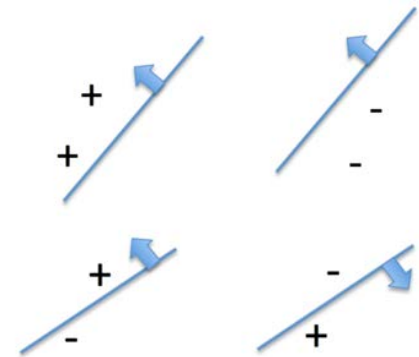
The growth function for three points is

$$S_{\mathcal{F}}(3) = 8 = 2^3$$

The growth function for four points is

It is less than 2^4 .

$$S_{\mathcal{F}}(n) \leq 2^n$$

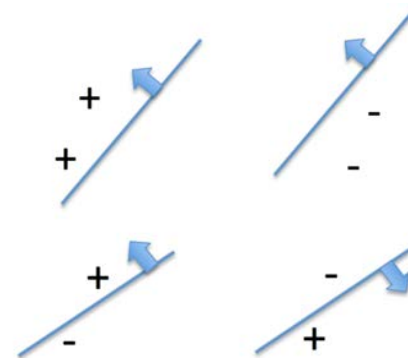


If $S_{\mathcal{F}}(n) = 2^n$, there is a dataset of n points where \mathcal{F} can perfectly classify them, no matter what the labels are.

\mathcal{F} *shatters* the set.

$$S_{\mathcal{F}}(n) \leq 2^n$$

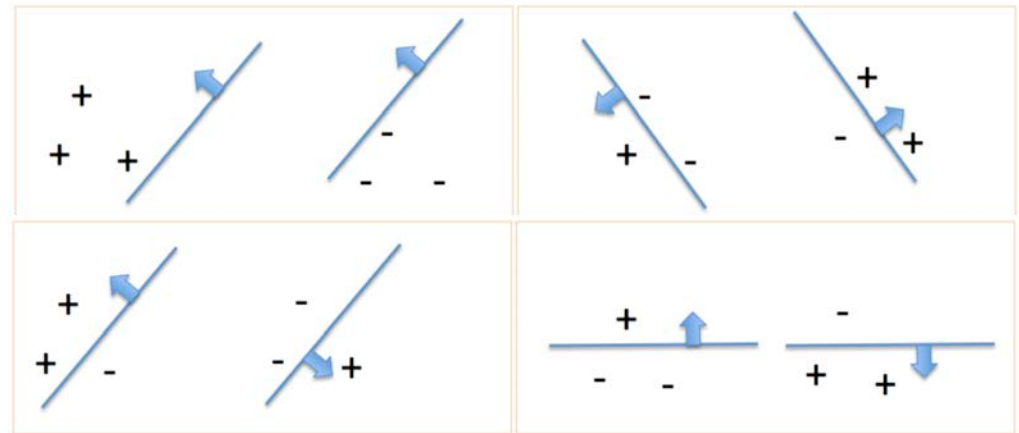
{Lines in the plane} shatters
2 data points in 2D



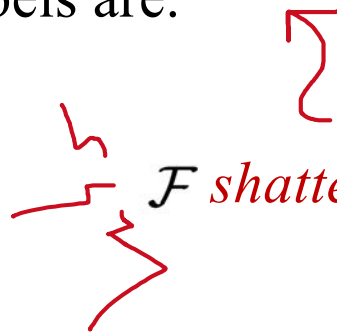
If $S_{\mathcal{F}}(n) = 2^n$, there is a dataset of n points where \mathcal{F} can perfectly classify them, no matter what the labels are.

\mathcal{F} *shatters* the set.

{Lines in the plane} shatters
3 data points in 2D.



If $S_{\mathcal{F}}(n) = 2^n$, there is a dataset of n points where \mathcal{F} can perfectly classify them, no matter what the labels are.

 \mathcal{F} *shatters* the set.

{Lines in the plane} does not
shatter 4 data points in 2D. 

-	+
+	-

The **VC dimension** of \mathcal{F} is the largest number of points it can shatter.



Definition: The *VC dimension* of \mathcal{F} is the largest number of points n such that:

$$S_{\mathcal{F}}(n) = 2^n.$$



What is the VC dimension of halfspaces in 2 dimensions? 3

Can you guess the VC dimension of halfspaces in p dimensions? $p+1$

The **VC dimension** of \mathcal{F} is the largest number of points it can shatter.

Note: VC dimension is the largest number of points n such that *there exists* some configuration of them that can be shattered.

+ - + $\frac{-}{+ \quad +}$

The **VC dimension** of \mathcal{F} is the largest number of points it can shatter.

Note: VC dimension is the largest number of points n such that *there exists* some configuration of them that can be shattered.

To prove that VC dimension of \mathcal{F} is h ,

- show *there exists* a configuration of h points that can be shattered
- show *no* configuration of $h+1$ points exist that can be shattered

The **VC dimension** of \mathcal{F} is the largest number of points it can shatter.



Definition: The *VC dimension* of \mathcal{F} is the largest number of points n such that:

$$S_{\mathcal{F}}(n) = 2^n.$$



What is the VC dimension of halfspaces in 2 dimensions? 3

Can you guess the VC dimension of halfspaces in p dimensions? $p+1$

Is the VC dimension always related to the number of parameters? No.

Statistical Learning Theory

Part 10: VC dimension \neq number of parameters

Cynthia Rudin

Duke

There exists a 1-parameter family of functions with infinite VC dimension.

i	y	$x = 2\pi 10^{-i}$	$\text{sign}(\sin(tx))$	y
1	1	0.628318530717959	1	1
2	-1	0.062831853071796	-1	-1
3	1	0.006283185307180	1	1
4	1	0.000628318530718	1	1
5	1	0.000062831853072	1	1
6	-1	0.000006283185307	-1	-1
7	1	0.000000628318531	1	1
8	-1	0.000000062831853	-1	-1
9	-1	0.000000006283185	-1	-1
10	1	0.000000000628319	1	1
11	-1	0.000000000062832	-1	-1
12	1	0.000000000006283	1	1
13	-1	0.000000000000628	-1	-1

$$t = \frac{1}{4} \sum_{i=1}^n [(1 - y_i) 10^i + 1]$$

$$t = (1/4) * (\text{sum}(((1-y).*(10.^[i]))+1));$$

$$t = 5.050550500053250\text{e}+12$$

There exists a 1-parameter family of functions with infinite VC dimension.

VC dimension \neq number of parameters

Statistical Learning Theory

Part 11: VC bound

Cynthia Rudin

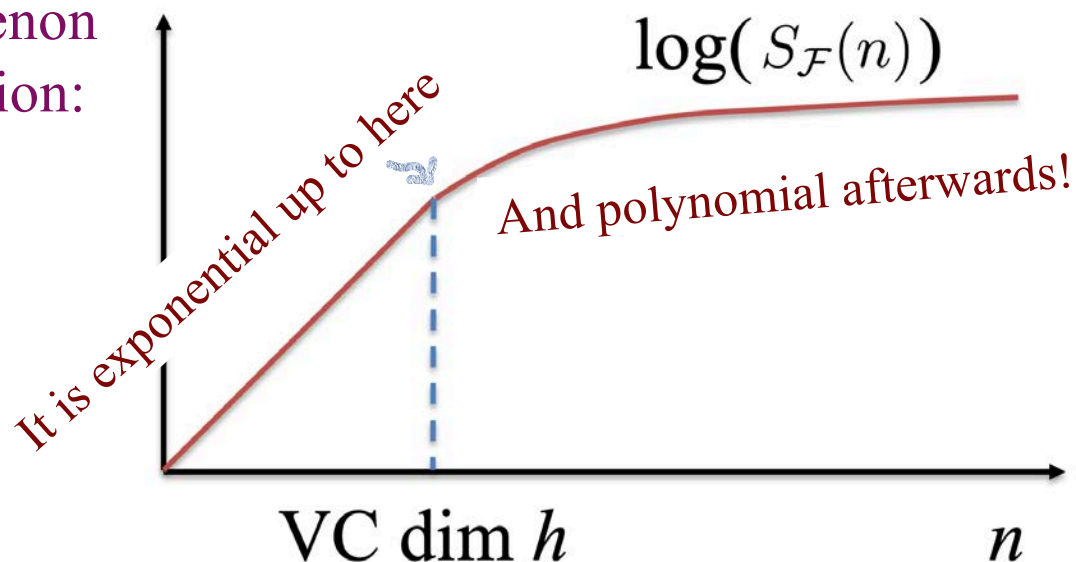
Duke

Recall that in Theorem GrowthFunction, we ran into a problem... Perhaps VC dim will solve it.

If a class of functions has VC dim h , we know we can shatter n observations when $n \leq h$ and $S_{\mathcal{F}}(n) = 2^n$.

When $n > h$, we can't shatter, and $S_{\mathcal{F}}(n) < 2^n$.

An intriguing phenomenon about the growth function:

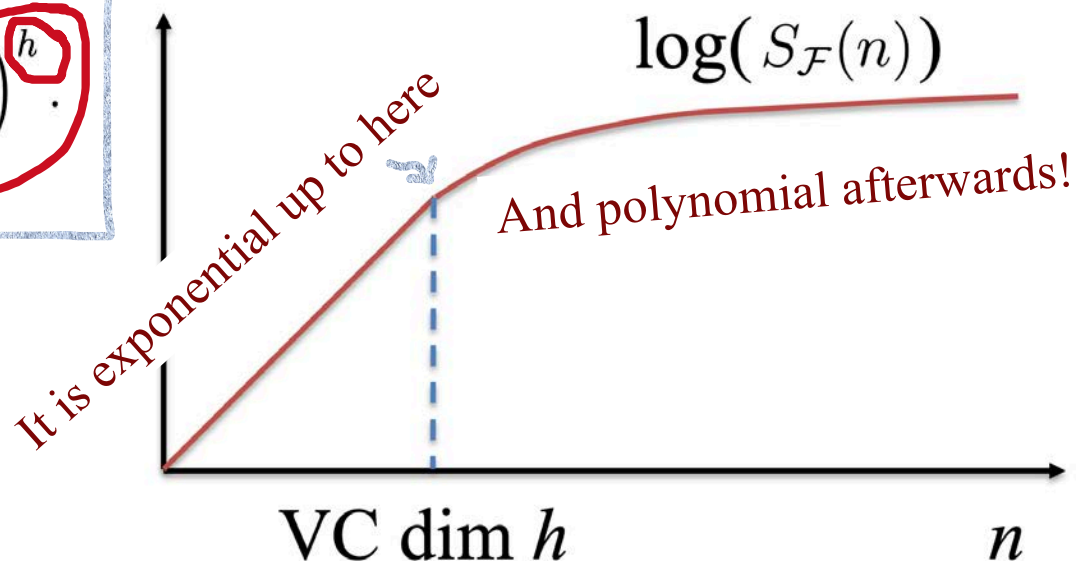


Lemma. (Vapnik and Chervonenkis, Sauer, Shelah) *Let \mathcal{F} be a class of functions with finite VC dimension h . Then for all $n \in \mathbb{N}$,*

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^h \binom{n}{i}$$

and for all $n \geq h$,

$$S_{\mathcal{F}}(n) \leq \left(\frac{en}{h} \right)^h.$$



Combining this lemma with
Theorem GrowthFunction:

Theorem VC-Bound. If \mathcal{F} has VC dim h , and for $n \geq h$,
with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{h(\ln(2n) + 1) + \log \frac{4}{\delta}}{n}}.$$

Difference between true and empirical risks is at most

$$\sqrt{\frac{h \log n}{n}}$$

This is sooo much better than infinite!

Theorem VC-Bound. If \mathcal{F} has VC dim h , and for $n \geq h$,
with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{\frac{h(\ln(2n) + 1) + \log \frac{4}{\delta}}{n}}.$$

Why is the VC bound important?

It's a generalization bound that is non-vacuous even for infinite function classes.

It's a finite sample bound.

VC dimension can be computed or bounded in many cases.

Beautiful combinatorial quantity.

Tells you what quantities are important for the learning process.

Caveat:

Too loose to be directly useful in practice. You can't minimize it and expect to keep the true risk low.

But, it tells you what quantities are important for the learning process.

with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{h(\ln(2n) + 1) + \log \frac{4}{\delta}}{n}}$$

Statistical Learning Theory

Part 12: The margin theory

Cynthia Rudin

Duke

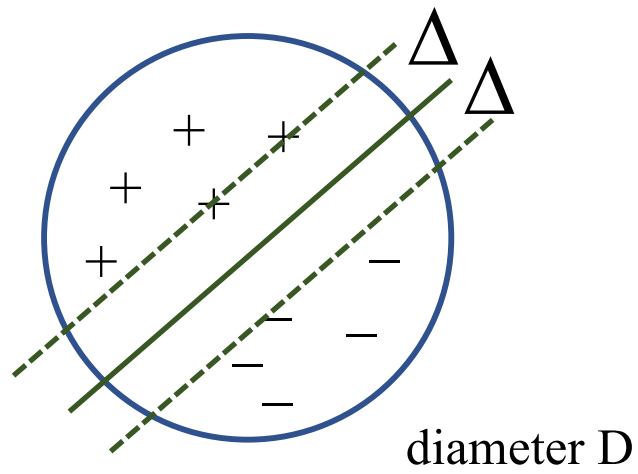
Caveat:

Too loose to be directly useful in practice. You can't minimize it and expect to keep the true risk low.

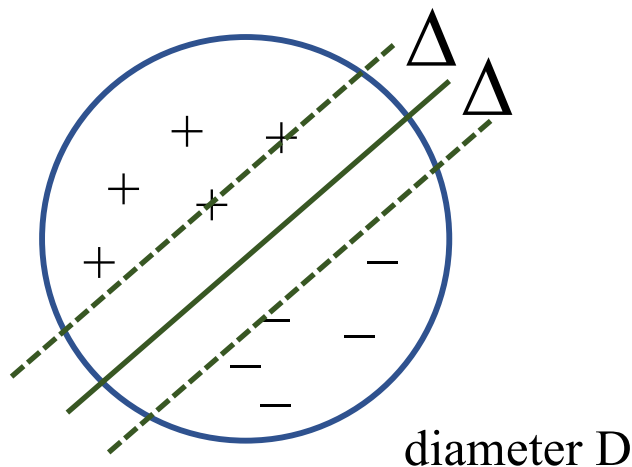
But, it tells you what quantities are important for the learning process.

with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{h(\ln(2n) + 1) + \log \frac{4}{\delta}}{n}}$$



“Gap-tolerant” classifiers




“Theorem” VC-Margin. (Vapnik) *For data in \mathbf{R}^p , the VC dimension h of (linear) gap-tolerant classifiers with gap Δ belong to a sphere of diameter D is bounded by the inequality:*

$$h \leq \min \left(\left\lceil \frac{D^2}{\Delta^2} \right\rceil, p \right) + 1. \quad \xrightarrow{\Delta \rightarrow 0} \quad p+1$$

An upper bound on the **True Risk** by:

- **Empirical Risk**
- **Margin**


$$\forall f \in \mathcal{F} \quad R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2 \frac{h(\ln(2n) + 1) + \log \frac{4}{\delta}}{n}}$$
$$h \leq \min \left(\left\lceil \frac{D^2}{\Delta^2} \right\rceil, p \right) + 1.$$

