Lecture 28: 5/1/17

Two Types of Error

① True Error (aka. expected risk) (aka. Generalization Error)

$$R(h) = P_{x \sim p^*(x)}(c^*(x) \neq h(x))$$ ← always unknown.

② Train Error (aka. empirical risk)

$$\hat{R}(h) = P_{x \sim S}(c^*(x) \neq h(x))$$

$$S = \{x^{(1)}, \ldots, x^{(N)}\}$$

known, computable

$$= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(c^*(x^{(i)}) \neq h(x^{(i)}))$$

$$= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y^{(i)} \neq h(x^{(i)}))$$

Three Decisions Fns. of Interest

✓ ① True function (oracle), c^*

$$y^{(i)} = c^*(x^{(i)}) \quad \forall i$$

✓ ② Expected Risk Minimizer (lowest true error)

$$h^* = \underset{h \in H}{\text{argmin}}\ R(h)$$

hypothesis class.

③ Empirical Risk Minimizer (lowest training error)

$$\hat{h} = \underset{h \in H}{\text{argmin}}\ \hat{R}(h) = \underset{h \in H}{\text{argmin}}\ \frac{1}{N} \sum_{i} \mathbb{1}(y^{(i)} \neq h(x^{(i)}))$$

Q: Which of these are unknown?

PAC Learning

Q: Can we bound $R(h)$ in terms of $\hat{R}(h)$?

A: Yes!

PAC stands for Probably
Approximately
Correct

PAC learner yields hypothesis h, which is approximately correct with high probability

$R(h) \approx 0$

$\Pr(R(h) \approx 0) \approx 1$

Def: PAC Criterion

$$\Pr(\forall h, |R(h) - \hat{R}(h)| \le \underset{\text{small}}{\epsilon}) \ge 1 - \underset{\text{small}}{\delta}$$

Q: What is random?

A: $\hat{R}$ is measured on a random sample S of training examples

Def: Sample complexity is the minimum value N of training examples s.t. the PAC criterion holds for a given ϵ, δ

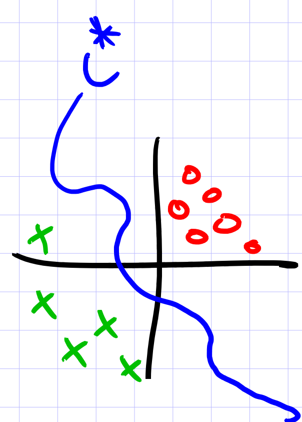
Def: Hypothesis space H is PAC learnable if sample complexity N is a polynomial function of $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ for some learning algorithm.

4 Bounds

Two cases for c^*

A) Realizable case: $c^* \in H$

B) Agnostic case: c^* is not necessarily in H



Two Cases for H:

A) Finite $|H| < +\infty$

B) Infinite $|H| = +\infty$

Thm 1: Sample Complexity (Realizable, Finite $|H|$)

$$N \geq \frac{1}{\epsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$ labeled examples

are sufficient to ensure that with prob. $\geq (1-\delta)$ all $h \in H$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

★ Bound is inversely linear in ϵ
(Halving the error requires double the examples)

★ Bound is logarithmic in $|H|$
(Doubling the hyp. space size requires only logarithmically more train exs. N)

Ex: Conjunctions

H = class of conjunctions over $\vec{x} \in \{0,1\}^M$

e.g. $h(\vec{x}) = x_1 \bar{x}_3 x_4 = x_1(1-x_3)x_4$

$h(\vec{x}) = x_1 \bar{x}_2 x_4 \bar{x}_6$

Q: Suppose $M = 10$, $\epsilon = 0.1$, $\delta = 0.01$, what is the samp. comp.?

~~$|H| = 2^M$~~ $|H| = 3^M$

Then $N \geq \frac{1}{0.1}\left[\ln(3^M) + \ln\left(\frac{1}{0.01}\right)\right]$

$= \frac{1}{0.1}\left[M \ln(3) + \ln\left(\frac{1}{0.01}\right)\right]$

$= \frac{1}{0.1}\left[10 \ln(3) + \ln\left(\frac{1}{0.01}\right)\right] \approx 156$ examples

Corollary 1:

With prob. $(1-\delta)$ for all $h \in H$ s.t. $\hat{R}(h) = 0$ we have that: $R(h) \leq \frac{1}{N}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$

$\boxed{\text{probably}}$ $\boxed{\text{approximately correct}}$ assuming a $\boxed{\text{consistent learner}}$

Thm 2: Sample Complexity (Agnostic, Finite $|H|$)

$$N \geq \frac{1}{2\epsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right] \text{ labeled examples}$$

are sufficient so that with prob. $\geq (1-\delta)$ all $h \in H$ have $|R(h) - \hat{R}(h)| \leq \epsilon$

PAC Criterion: $P(\forall h, |R(h)-\hat{R}(h)| \leq \epsilon) \geq (1-\delta)$

★ Bound is inversely quadratic in ϵ "quadratically more"
(halving the error requires 4x examples)

Corollary 4:

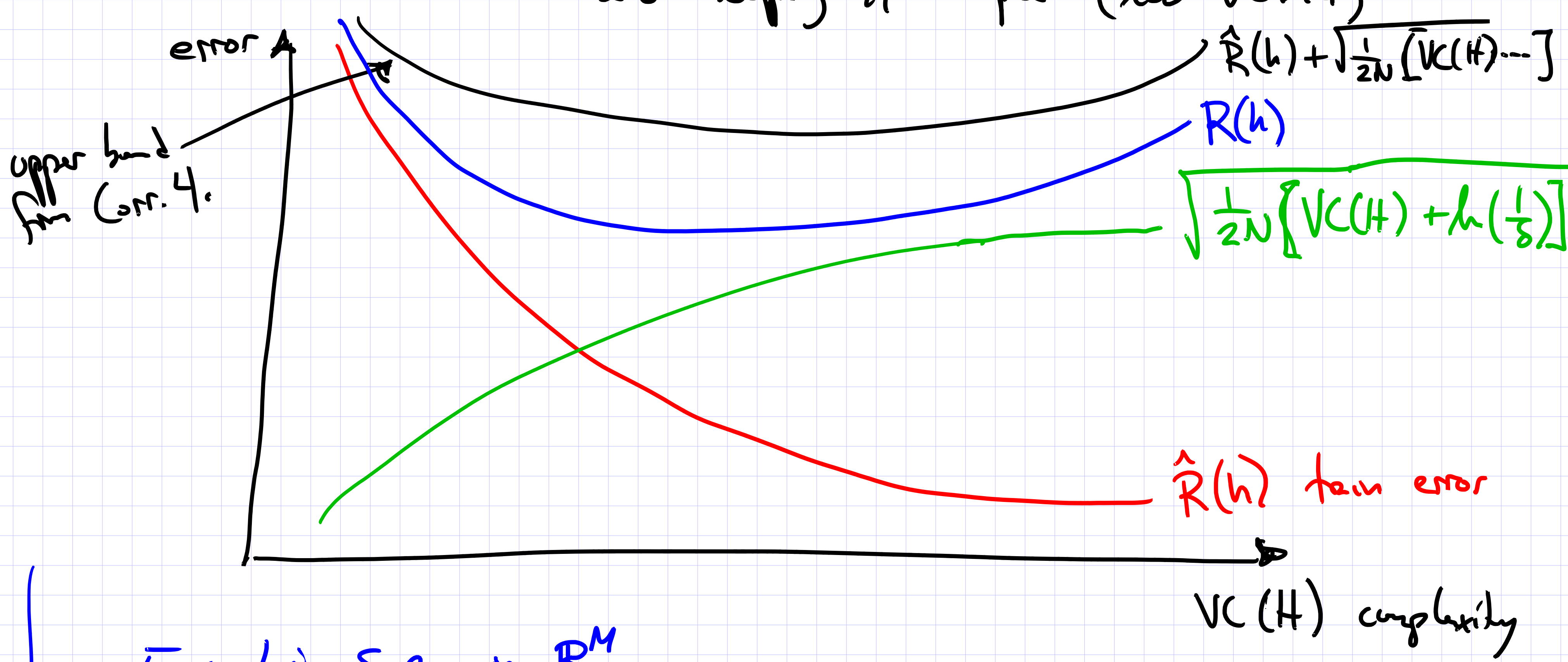
with prob. $(1-\delta)$ for all $h \in H$

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left[VC(H) + \ln\left(\frac{1}{\delta}\right)\right]}$$

Structural Risk Minimization

Q: Should Corr. 4 inform how do we model selection?
A: Yes!

Model Selection: tradeoff between low train error and keeping H simple (low VCDim)



Ex: Lin Sep in $\mathbb{R}^M$

$VC(H) = M+1$

How to tradeoff?

Train w/ L1 Regularizer ← prefers 0s in $\vec{\theta}$