CS 121.5 New Frontiers in ML Theory

October 4th, 2019

What is learning?

Supervised learning

- $X \to Y$
  input labels
- given labeled ex. (x_i, y_i)

- correct on examples close to those that were given
- Assume samples are iid from uniform D
  $(x, y) \sim D$

what you're hoping to learn

Given n iid samples $\{(x_i, y_i)\} \sim D$, $f \in F$

Goal Output $\hat{f}: X \to Y$ w/ low test error $\mathcal{L}_D(\hat{f}) = \Pr_{(x,y)\sim D}[\hat{f}(x) \neq y] \leq \varepsilon$

what you actually learn

$D_f = \{(x, f(x))\}$ ← label y guaranteed to be a function of x

PAC Learning (Probably Approximately Correct)

Family F is PAC-learnable if $\exists$ alg A s.t. $\forall D_f \in D_F$ $\forall \varepsilon, \delta$ and $n = |S| = \text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta})$

$\hat{f} \leftarrow A(S)$, $S = \{(x_i, y_i)\} \sim D^n$

$\Pr[\mathcal{L}_D(\hat{f}) \leq \varepsilon] \geq 1 - \delta \sim 0.999$

accuracy

measurement of "good" dataset sample

no matter what dist of input, you'll get a good quality classifier (low value of loss function).

Claim. $|F|$ finite $\to$ F is PAC-learnable w/

$$n = \frac{\log(|F|/\delta)}{\varepsilon^2} \text{ samples}$$

most important term

w/o regards to efficiency

$\uparrow n$ means you can afford to shrink ε and δ

Empirical Risk Minimization (ERM)

- want f s.t. $\mathcal{L}_D(f) = \mathbb{E}_{x,y\sim D}[\underbrace{1\{f(x)\neq y\}}_{\ell(f(x),y)}]$ small

population loss

Try: $\operatorname{argmin}_{f\in F} \mathcal{L}_S(f) = \frac{1}{n}\sum_{(x_i,y_i)\in S} \ell(f(x_i), y_i)$

empirical loss

works w/ uniform convergence (ε) (of F, D, n) = will hold over $S\sim D^n$: $\{f\in F : |\mathcal{L}_D(f) - \hat{\mathcal{L}}_S(f)| \leq \varepsilon\}$

Corollary: if family has uniform convergence w/ ε, $\mathcal{L}_D(\hat{f}_{erm}) \leq \min_{f\in F} \mathcal{L}_D(f) + \varepsilon_{unif}$

best classifier

Lemma: $\Pr_{S\sim D^n}[\exists f\in F : |\mathcal{L}_D(f) - \hat{\mathcal{L}}_S(f)| > \varepsilon] \leq \underbrace{|F| e^{-\varepsilon^2 n}}_{\delta}$

- 1D random variable variation from mean: Chernoff Bounds (exp. small in t^2)

- union bound over all functions in family: probability that any function deviates ≤ num functions · probability one deviates

Ex: Binary Linear Classifier $f_w(x) = I\{\langle w, x\rangle \geq 0\}$

$w \in \mathbb{R}^d$, $|F| \sim 2^{O(d)}$, $n \sim O(d)$

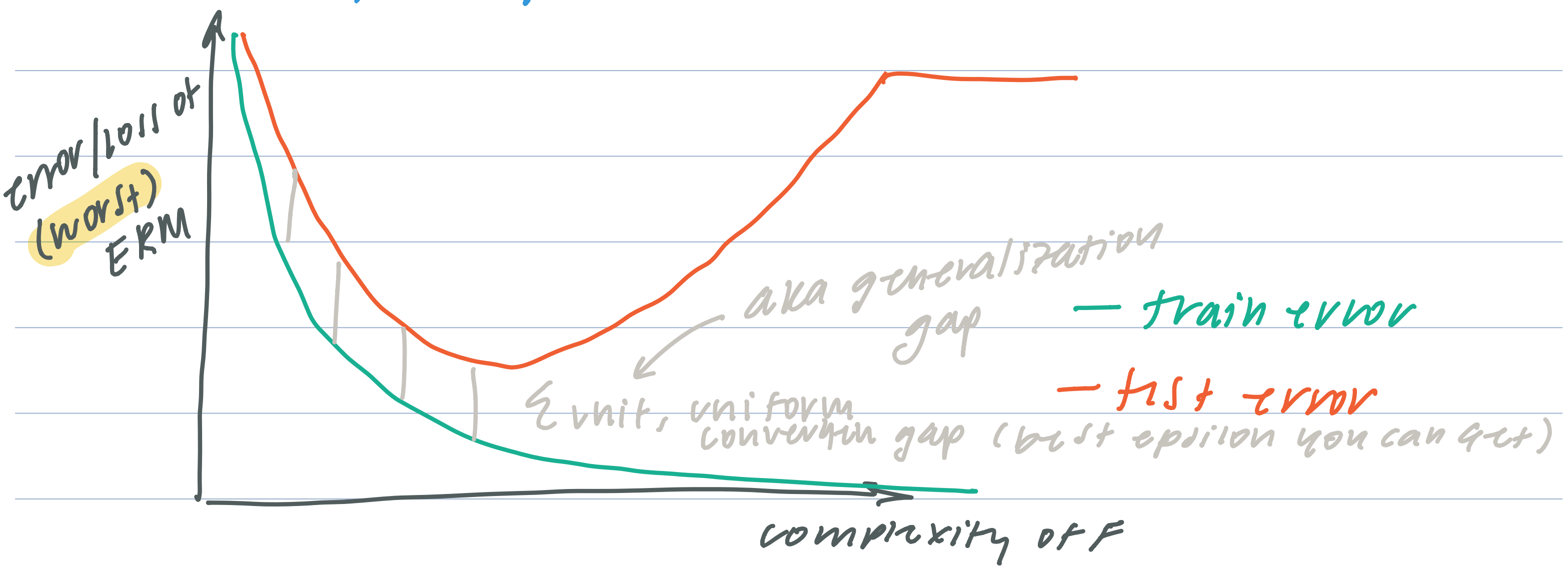
High Error:

- too simple classifier ← underfitting
- minimum pop error of family is no good
- classifier fits highly to train set ← overfitting

When does uniform convergence hold?

- not hold if function family is too big
- holds if small

Bias-Complexity / Variance Tradeoff



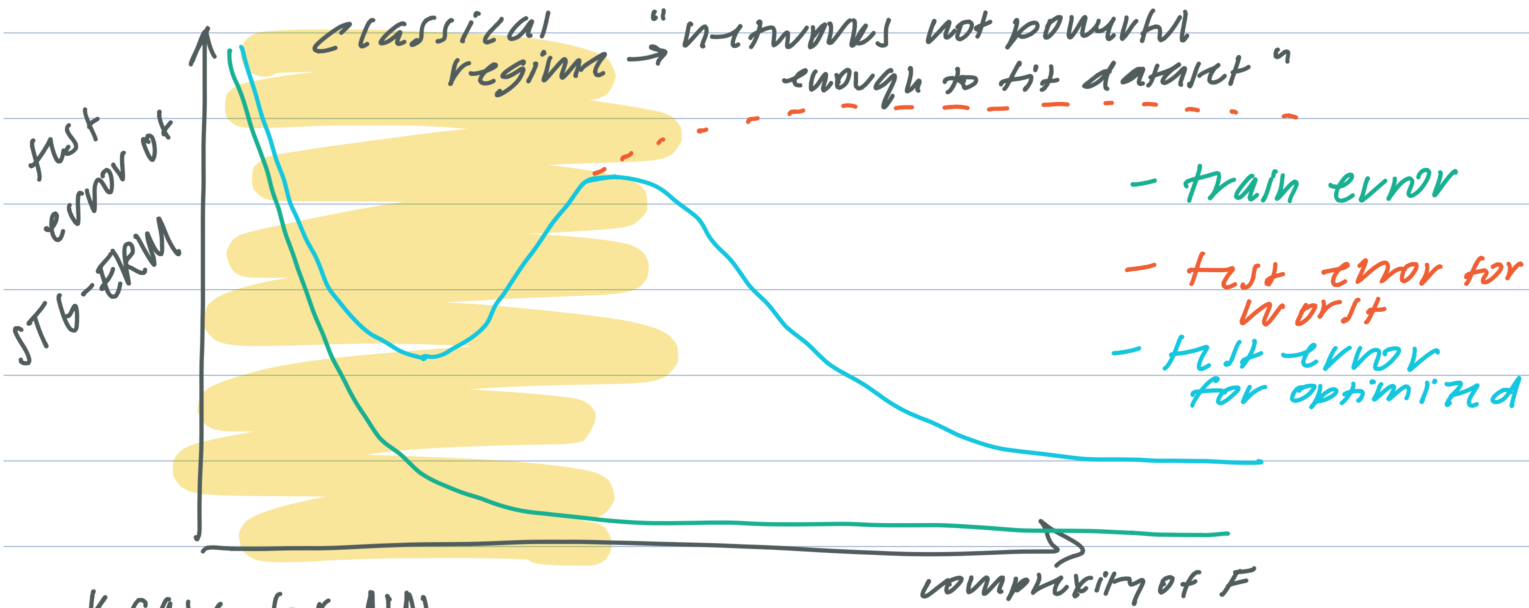
*n and D constant

random maker complexity: how well F can fit ← family of functions

n randomly samples from distribution

$R_{D,n}(F)$

- can fit, family is too complex



* case for NN

NN factors: dist D, model F, optimization algo, number samples n

<u>Training Algorithm A</u>: input (x_i, y_i) → output model M

Model Complexity$_D(A) := \max n$ s.t. Train Error$(A(S))$ ≈ 0 for $S \sim D^n$

For n train samples:

under-parametrized Model complexity$(A) \ll n$

critically-parametrized Model complexity$(A) = n$

over-parametrized Model complexity$(A) \gg n$