Generalization theory

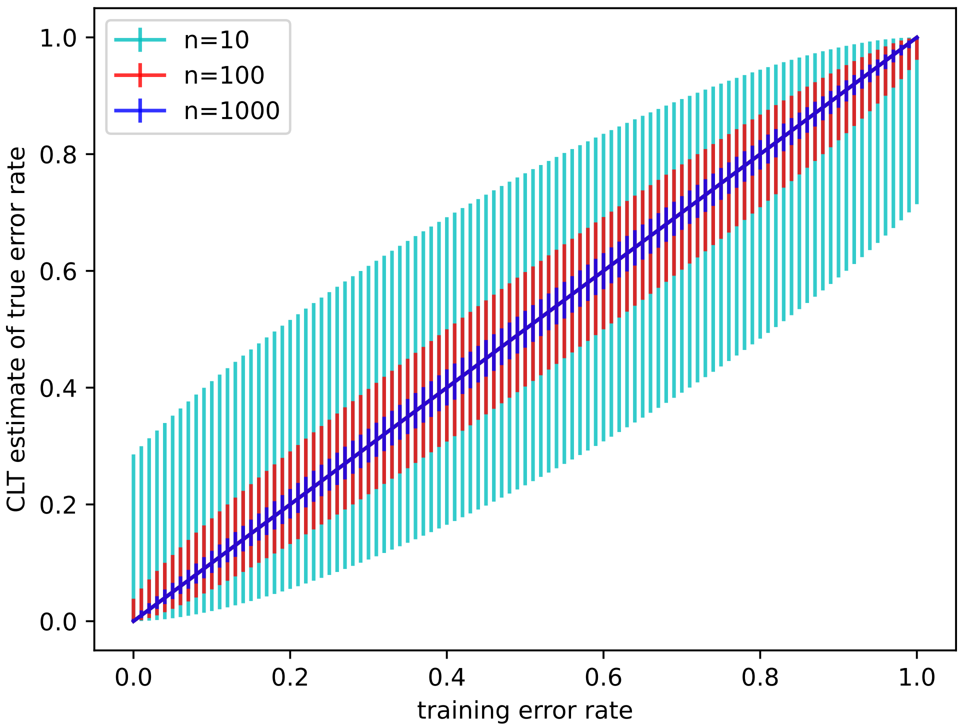
COMS 4771 Fall 2023

In-sample vs. out-of-sample performance

- Basic premise: training data is sample from population (or distribution)
- In-sample: what happens on training data
- Out-of-sample: what happens in overall population
- Learning algorithm: find classifier f with low training error rate $\widehat{\text{err}}[f]$
  - Will this classifier f also have low (true) error rate $\text{err}[f]$?
  - Basic answer from statistical learning theory: Yes, if classifier is chosen from a "simple" function class $\mathcal{F}$

Training error rate of a fixed classifier

Suppose you chose classifier f before even looking at the training data $\mathcal{S} = ((X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})) \overset{\text{i.i.d.}}{\sim} (X, Y)$



Training error rate of learned classifier

Usually, we choose a classifier $\hat{f}$ based on the training data $\mathcal{S}$

Why can't previous analysis apply?

Two different random variables, $\widehat{\mathrm{err}}[\hat{f}]$ and $\mathrm{err}[\hat{f}]$:

$$\widehat{\mathrm{err}}[\hat{f}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{f}(X^{(i)}) \neq Y^{(i)}\}, \qquad \mathrm{err}[\hat{f}] = \Pr(\hat{f}(X) \neq Y \mid \hat{f})$$

Typically how different are they?

Conservative answer: if $\hat{f}$ is chosen from $\mathcal{F}$, then

$$\Pr(|\widehat{\mathrm{err}}[\hat{f}] - \mathrm{err}[\hat{f}]| > \epsilon) \leq \Pr(\text{there exists } f \in \mathcal{F} \text{ s.t. } |\widehat{\mathrm{err}}[f] - \mathrm{err}[f]| > \epsilon)$$

Union bound: For any events A and B,

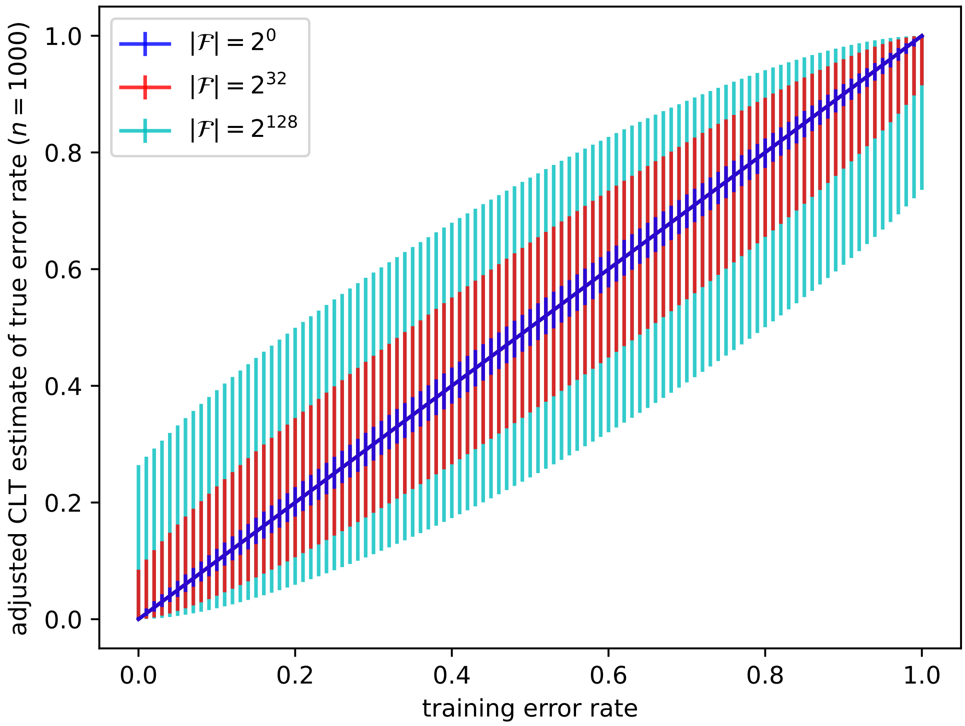
$$\Pr(A \text{ or } B) = \Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$

Chernoff bound: for any fixed $f \colon \mathcal{X} \to \mathcal{Y}$,

$$\Pr(|\widehat{\mathrm{err}}[f] - \mathrm{err}[f]| > \epsilon) \leq 2\exp\bigl(-2n\epsilon^2\bigr)$$

Comparison to bound for a single f based on CLT:

- Doesn't have factor of $\sqrt{\mathrm{err}[f](1 - \mathrm{err}[f])}$ from single f CLT bound
  - Can get this using advanced version of "Chernoff bound"
- Scary/weird constants
  - But inside the logarithm (and maybe can be improved)
- Bound grows with $\sqrt{\ln|\mathcal{F}|}$
  - Roughly like reducing n by a factor of # bits needed to represent a classifier $f \in \mathcal{F}$



Counting number of behaviors

The cardinality of $\mathcal{F}$ is a crude measure of its "complexity"

- Example: $\mathcal{F}$ is all "threshold functions on $\mathbb{R}$"

$$f_t(x) = \mathbb{1}\{x > t\}$$

  - There are uncountably-many such classifiers, one per $t \in \mathbb{R}$
  - But can only label a dataset of size n in $n + 1$ different ways

Better measure: number of behaviors on the unlabeled data $x^{(1)}, \dots, x^{(n)}$

$$S(\mathcal{F}; (x^{(i)})_{i=1}^n) = |\{(f(x^{(1)}), \dots, f(x^{(n)})) : f \in \mathcal{F}\}|$$

Examples:

- If $\mathcal{F}$ = all threshold functions on $\mathbb{R}$,

$$S(\mathcal{F}; (x^{(i)})_{i=1}^n) \leq n + 1$$

- If $\mathcal{F}$ = all linear classifiers in $\mathbb{R}^d$,

$$S(\mathcal{F}; (x^{(i)})_{i=1}^n) \leq O(n^d)$$

Number of behaviors of large margin linear classifiers:

- Consider unlabeled data $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ satisfying $\|x^{(i)}\| \leq 1$
- Let $\mathcal{F}$ = homogeneous linear classifiers with margin $\gamma > 0$ on these n data points (i.e., distance from $x^{(i)}$ to decision boundary is $\geq \gamma$)
- What is the number of behaviors of $\mathcal{F}$ on $(x^{(i)})_{i=1}^n$?