23 Learning Theory

humans are good at generalizing

learning theory: tries to explain how ML algos. generalize
- as well as how much training data is needed

observation: If we want to generalize, we must constrain what hypotheses our learner can consider

Range Space: pair (P, H) where:
(Set System)
- P = set of all training/test points (can be infinite)
- H: hypothesis class/set of hypotheses/ranges
  - a hypothesis is a subset $h \subseteq P$ that specifies which points h predicts as being in class C
    - ↳ each h is a binary classifier
    - ↳ H is a set of sets of points

Some Examples:

- Power Set Classifier: P is set of k numbers, H is power set of P → all 2^k subsets of P
- linear classifier: $P = \mathbb{R}^d$, H is set of all halfspaces, each of the form $\{\vec{x} \mid \vec{w}^T \vec{x} \geq -\alpha\}$
  - P & H infinite
    - H contains every possible halfspace (every possible classifier in d dimensions)

* Powerset classifier powerful since it can learn every possible hypothesis
  - can't generalize at all, doesn't learn anything
* linear classifier can generalize from a few points
  - disadvantage if data not linearly separable

Now, how well does training error predict test error? How does this differ for each classifier?

- Suppose all training & test pts. drawn independently from same probability distribution D defined on P
  - D determines each pt.'s label
    - C & N might have overlapping distributions
- let h be a hypothesis. $x \in h \Rightarrow x$ predicted as class C

risk/generalization error $R(h)$: probability that h misclassifies random x from D
- probability $x \in C$ but $x \notin h$

* almost same as test error: Risk is avg. test error for test pts drawn randomly from D
  - infinite # test pts $\Rightarrow$ risk = test error

Empirical Risk: Let $X \subseteq P$ be set of n training pts. from D. $\hat{R}(h)$ = percent of X misclassified by h (training error)
  - Same as lecture 12 definition if 0-1 loss

h misclassifies a training pt w/ probability $R(h)$
  - total # misclassified pts. is a binomial distribution
    - $n \to \infty$ then $\hat{R}(h)$ better approximates $R(h)$

at $n=20$, we're lucky and have low training error

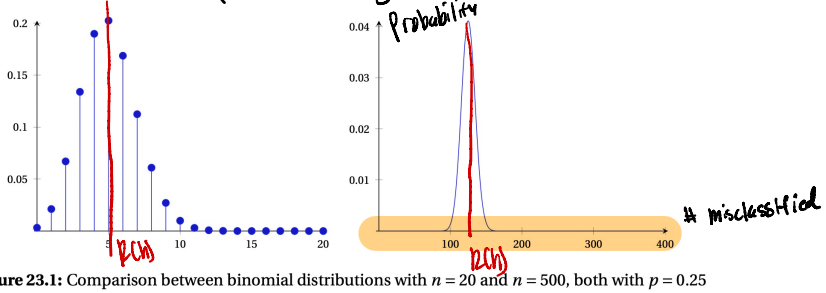
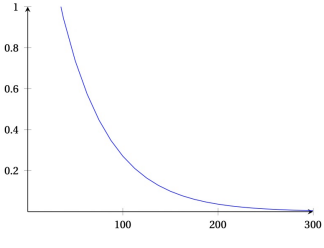


Figure 23.1: Comparison between binomial distributions with $n = 20$ and $n = 500$, both with $p = 0.25$

- if infinite data, then distribution would become infinitely narrow
  ↳ training error = risk

What's the probability of a bad estimate $\hat{R}(h)$?

Hoeffding's Inequality: $P(|\hat{R}(h) - R(h)| > \varepsilon) \leq 2e^{-2\varepsilon^2 n}$, n = # training pts.



← big $n \Rightarrow$ unlikely for binomial RV to be far from its mean

Figure 23.2: Hoeffding's bound for $\varepsilon = 0.1$, plotting how n affects $\mathbb{P}(|\hat{R}(h) - R(h)| > \varepsilon)$

↳ take $\geq n=200$ to have high confidence of achieving error bound $\varepsilon = 0.1$

Idea for learning algorithm: choose $\hat{h} \in H$ that minimizes training error $\hat{R}(\hat{h})$

↳ empirical risk minimization

*none of the algorithms we've studied do this since it's computationally infeasible to pick the best hypothesis
- e.g. SVMs when data not lin. seperable → finds classifier w/ low training error that's not necessarily the one w/ minimum training error (optimal)
- If we pretend that we have the compute to try every hypothesis, there's still a problem
↳ some h might be lucky & have high R(h) that gets lucky w/ low $\hat{R}(h)$

- issue isn't too many hypotheses, but too many dichotomies

Dichotomy: dichotomy of X: $X \wedge h$, $h \in H$
↳ training pts. that h predicts to be in class C
"fn. mapping x_i to C or N"

e.g. linear classifier: CCN & CCC are valid, but CNC isn't

for n training pts, there are 2^n possible dichotomies
- more dichotomies → more likely for 1 to get lucky and have low empirical risk

- In the extreme case that H allows all 2^n dichotomies, then some $\hat{h}$ exists s.t. $\hat{R}(\hat{h}) = 0$ even if every $h \in H$ has high risk
* power set imposes no structure → overfitting

given Π dichotomies, $P[\text{at least one dichotomy has } |\hat{R} - R| > \varepsilon] \leq 2\Pi e^{-2\varepsilon^2 n} = \delta$

If we fix δ & solve for ε, we have $P\left[|\hat{R} - R| \leq \sqrt{\frac{1}{2n} \ln \frac{2\Pi}{\delta}} = \varepsilon\right] \geq 1 - \delta$

Smaller Π, larger n → training error better estimate for R(h)
- Small Π: less likely to overfit (less V, more B)
- doesn't necessarily mean low risk. If h doesn't fit data well, both train & test error large
ideally h small but fits data well

Suppose $h^* \in H$ be the minimizer of $R(h^*)$. This is the "best classifier."

- $\hat{h}$ minimizes empirical risk $\hat{R}$, but we want minimizer of actual risk
  - can't know what h^* is

↳ Small Π and large $n \Rightarrow \hat{h}$ probably close to h^*

↳ $P(\hat{h}$ nearly optimal risk$) \geq 1-\delta$

best possible risk

$$R(\hat{h}) \leq \hat{R}(\hat{h}) + \varepsilon \leq \hat{R}(h^*) + \varepsilon \leq R(h^*) + 2\varepsilon$$

risk on our classifier

↳ only off by 2ε

$\Rightarrow$ high n and small $\Pi \Rightarrow$ empirical risk minimization chooses $\hat{h}$ close to best $h \in H$

- fixed δ and ε, the Sample Complexity is # training pts. needed to achieve this w/ probability $\geq 1-\delta$

$$n \geq \frac{1}{2\varepsilon^2} \ln \frac{2\Pi}{\delta}$$

→ Small Π → don't need too many training points

if $\Pi = 2^n$, then $n = n$ → powerset classifier can't learn much nor generalize

↳ reduce Π by using linear classifier

Shatter fn. & linear classifiers

let $\Pi_H(X)$ be # dichotomies in X: $\Pi_H(X) = |\{X \cap h \mid h \in H\}| \in [1, 2^n]$

ways pts. can be labeled

Shatter fn: $\Pi_H(n) = \max_{|X|=n, X \subseteq P} \Pi_H(X)$ (most dichotomies out of any point set of size n) pick n that gives biggest $\Pi_H(X)$

e.g. linear classifier, H = set of all possible half spaces

↳ $\Pi_H(3) = 2^3 = 8$

$\Pi_H(4) = 14 \neq 16$ $\binom{N}{N}$

thm: for all range spaces, either $\Pi_H(n)$ is a polynomial in n or $\Pi_H(n) = 2^n$
- Surprising fact
- if m pts, some training & some test, then either range space permits every possible dichotomy, or it only permits a polynomial subset of 2^m possible dichotomies
  - training pts. don't help
  - Once training pts. labeled, # ways to classify test pts. reduced dramatically

* no shatter fn. is neither polynomial nor 2^m

We actually know exact # dichotomies for linear classifiers:

Cover's thm: linear classifiers in $\mathbb{R}^d$ allow up to $\Pi_H(n) = 2\sum_{i=0}^{d} \binom{n-1}{i}$ dichotomies

- for $n \leq d+1$, $\Pi_H(n) = 2^n$
- for $n \geq d+1$, $\Pi_H(n) \leq 2\left(\frac{e^{(n-1)}}{d}\right)^d$
  - polynomial in d, exponent d

$\Rightarrow$ Sample complexity needed to achieve $R(\hat{h}) \leq \hat{R}(\hat{h}) + \varepsilon \leq R(h^*) + 2\varepsilon$ w/ prob $\geq 1-\delta$ is

$$n \geq \frac{1}{2\varepsilon^2}\left(d \ln \frac{n-1}{d} + d + \ln \frac{4}{\delta}\right)$$

linear in d

corollary: linear classifier needs only $n \in O(d)$ training pts. for training error to accurately predict risk or test error

On the other hand, classifier w/ 2^n possible dichotomies, no amount of training data will ever guarantee $\hat{R}(h)$ accurately predicts $R(h)$

VC dimension of (P, H) is $VC(H) = \max\{n \mid \Pi_H(n) = 2^n\}$ (can be ∞)

H shatters a set of n pts. if $\Pi_H(n) = 2^n$

↳ VC(H) is size of largest X that H can shatter

↳ Such an X is a pt. set for which all 2^n dichotomies are possible

VC based on observation that sometimes makes it easy to bound polynomial if shatter fn. isn't 2^n for all n

theorem: $\Pi_H(n) \leq \sum_{i=0}^{VC(H)} \binom{n}{i}$ => if $n \geq VC(H)$, then $\Pi_H(n) \leq \left(\frac{e^n}{VC(H)}\right)^{VC(H)}$

→ VC dim. is upper bound on exponent of the polynomial

Corollary: $O(VC(H))$ training pts. suffice for accuracy, though hidden constant big

e.g. linear classifier in a plane, $\Pi_H(3) = 8$ but $\Pi_H(4) = 14$

=> $VC(H) = 3$, $\Pi_H(n) \leq \frac{e^3 n^3}{27}$ and $O(1)$ sample complexity

* VC(H) not always tightest bound

Important: there is a polynomial bound on the shatter fn.

→ to get generalization, need to limit expressiveness of H to limit the number of possible dichotomies

↳ may or may not increase B, but not limiting $\Pi_H(n)$ results in bad overfitting.