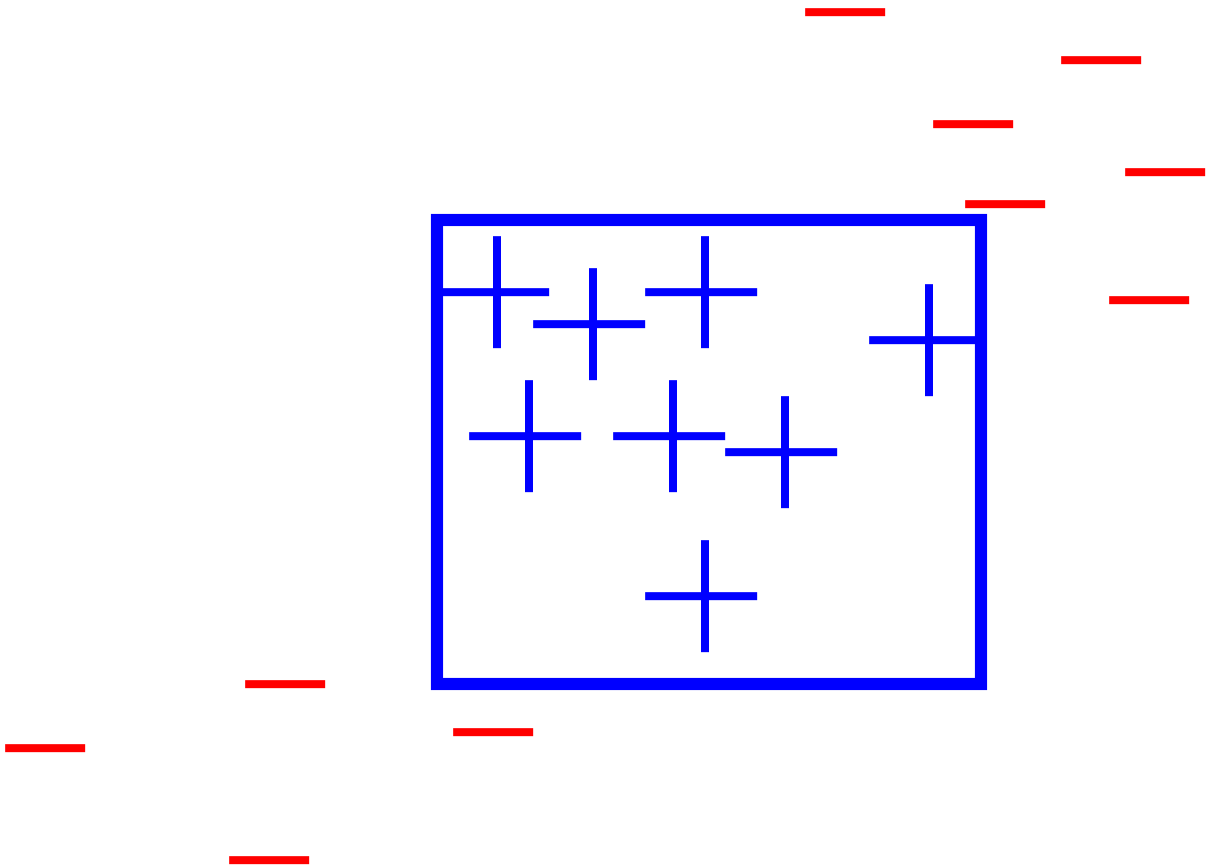
PAC Model and Generalization Bounds

Overview

- Probably Approximately Correct (PAC) model
- Basic generalization bounds
  - finite hypothesis class
  - infinite hypothesis class
    - Simple case
    - More next week

Motivating Example (PAC)

- Concept: Average body-size person
- Inputs: for each person:
  - height
  - weight
- Sample: labeled examples of persons
  - label + : average body-size
  - label - : not average body-size
- Two dimensional inputs



Motivating Example (PAC)

- Assumption: Target concept is a rectangle.
  - Realizable case
- Goal:
  - Find a rectangle that "approximate" the target.
- Formally:
  - With high probability
  - output a rectangle such that its error is low.

Example (Modeling)

- Assume:
  - Fixed distribution over persons.
- Goal:
  - Low error with respect to THIS distribution!!!
- How does the distribution look like?
  - Highly complex.
  - Each parameter is not uniform.
  - Highly correlated.

PAC approach

- Assume that the distribution is fixed.
  - But unknown
- Samples are drawn are i.i.d.
  - independent
  - identical
- Concentrate on the decision rule rather than distribution.

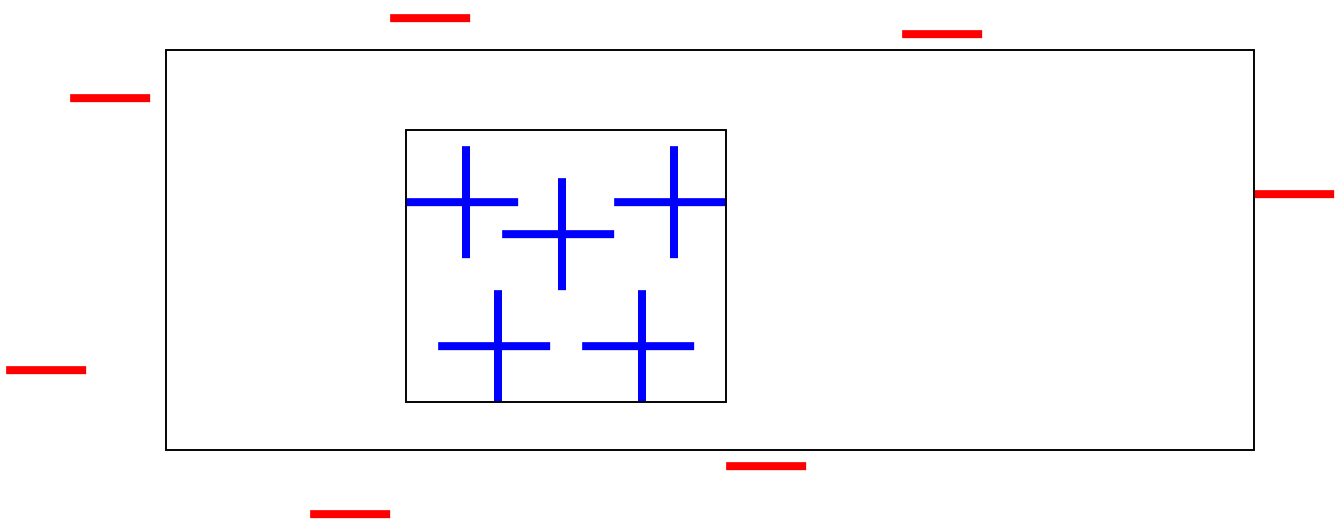
PAC Learning

- Task: learn a rectangle from examples.
- Input: points (x_1,x_2) and classification + or -
  - classifies by a rectangle R
- Goal:
  - With the fewest examples
  - compute R' efficiently
  - R' is a good approximation for R

PAC Learning: Accuracy

- Testing the accuracy of a hypothesis:
  - using the distribution D of examples.
- Error = R Δ R'
- Pr[Error] = D(Error) = D(R Δ R')
- We would like Pr[Error] to be controllable.
- Given a parameter ε:
  - Find R' such that Pr[Error] < ε.

PAC Learning: Hypothesis



- Which Rectangle should we choose?
- Later we show it is not that important.

PAC model: Setting

- A distribution: D (unknown)
- Target function: c_t from C
  - $c_t : X \rightarrow \{0,1\}$
- Hypothesis: h from H
  - $h: X \rightarrow \{0,1\}$
- Error probability:
  - $error(h) = Prob_D[h(x) \neq c_t(x)]$
- Oracle: $EX(c_t,D)$

PAC Learning: Definition

- C and H are concept classes over X.
- C is PAC learnable by H if
- There Exist an Algorithm A such that:
  - For any distribution D over X and c_t in C
  - for every input ε and δ:
  - outputs a hypothesis h in H,
    - while having access to $EX(c_t,D)$
  - with probability $1-\delta$ we have $error(h) < \varepsilon$
- Complexities: sample, running time

PAC: comments

- We only assumed that examples are i.i.d.
- We have two independent parameters:
  - Accuracy ε
  - Confidence δ
- Hypothesis is tested on the same distribution as the sample.
- No assumption about the likelihood of concepts.

Finite Concept class

- Assume $C=H$ and finite.
  - realizable case
- h is ε-bad if $error(h) > \varepsilon$.
- Algorithm:
  - Sample a set S of $m(\varepsilon,\delta)$ examples.
  - Find $\hbar$ in H which is consistent.
    - Basically ERM
- Algorithm fails if $\hbar$ is ε-bad.

Analysis

- Consistent hypothesis:
  - classifies all examples correctly
- Fix an hypothesis g which is ε-bad.
- The probability that g is consistent:
  - $\Pr[g \text{ consistent}] = \prod_i \Pr[g(x_i)=y_i] \leq (1-\varepsilon)^m < e^{-\varepsilon m}$
- Need: Prob[∃ g: g consistent & ε-bad]

Analysis

- The probability that:
  - exists g which is ε-bad and consistent:
  - Use union bound:

    Pr[∃ g: g is consistent and ε-bad]

    ≤ $\sum_{\text{g is } \varepsilon\text{-bad}}$ Pr[g is consistent and ε-bad]

    ≤ |H| Pr[g consistent and ε-bad] ≤ $2\,|H|\,e^{-2\varepsilon m}$
- Sample size: $m > (1/2\varepsilon)\ ln\ (2|H|/\delta)$
  - OR: $\Pr\left[\hbar\ is\ \frac{\log\frac{2|H|}{\delta}}{m} - bad\right] \leq \delta$

PAC: non-realizable case

- What happens if c_t not in H
- Needs to redefine the goal.
- Let h^* in H minimize the error $\beta = error(h^*)$
- Goal: find h in H such that
  - $error(h) \leq error(h^*) + \varepsilon = \beta + \varepsilon$
- Algorithm ERM
  - Empirical Risk Minimization

Analysis

- For each h in H:
  - let $\widehat{error}(h)$ be the error on the sample S.
- Compute the probability that:
  - $|\widehat{error}(h) - error(h)| < \varepsilon/2$
  - Chernoff bound: $2exp(-2(\varepsilon/2)^2 m)$
- Consider entire H :
  - Want to holds for any *h*
  - With prob $1-2|H| exp(-2(\varepsilon/2)^2 m) = 1-\delta$
- Sample size $\boldsymbol{m > (2/\varepsilon^2) \ln (2|H|/\delta)}$

Correctness

- Assume that for all h in H:
  - $|\widehat{error}(h) - error(h)| < \varepsilon/2$
- In particular:
  - $\widehat{error}(h^*) < error(h^*) + \varepsilon/2$
  - $error(h) - \varepsilon/2 < \widehat{error}(h)$
- For the output h_{ERM}:
  - $\widehat{error}(h_{ERM}) < \widehat{error}(h^*)$
- Conclusion: $error(h_{ERM}) < error(h^*) + \varepsilon$

ERM: Finite Hypothesis Class

- Theroem:
  - For any finite class H
  - For any target function c_t
  - Using sample size m
  - ERM will output ħ s.t.

$$\Pr\left[\text{error}(ħ) > \min_{\text{h in H}} \text{error}(h) + \sqrt{\frac{2\log^{2|H|}/_{\delta}}{m}}\right] \leq \delta$$

Example: Learning OR of literals

- Inputs: $x_1, \ldots, x_n$
- Literals : $x_1, \bar{x}_1$
- OR functions: $x_1 \vee \bar{x}_4 \vee x_7$
- Number of functions? $\mathbf{3^n}$

ELIM: Algorithm for learning OR

- Keep a list of all literals
- For every example whose classification is 0:
  - Erase all the literals that are 1.
- Example
- Correctness:
  - Our hypothesis h: OR of our set of remaining literals.
  - Our set of literals always includes the target OR literals.
  - Every time h predicts zero: we are correct.
- Sample size: $m > (1/\varepsilon) \ln (3^n/\delta)$

Learning parity

- Functions: $x_1 \oplus x_7 \oplus x_9$
- Number of functions:
  - 2^n
- Algorithm:
  - Sample set of examples
  - Solve linear equations
- Sample size: $m > (1/\varepsilon) \ln (2^n/\delta)$

Lower Bounds

- No free lunch Theorems
- Impossibility results
- Too few examples → any algorithm fails
- General structure:
  - Show an H and D
  - Show that if m too small,
    - For any output, expected error is high

Lower bounds: Hypothesis class

- Fix a finite domain X
- Let H be any Boolean function over X
  - $|H| = 2^{|X|}$
- Let D be uniform over X
  - $D(x) = 1/|X|$
- Target function c_t at random from H
  - For each x, $Pr[c_t(x)=1]= ½$
    - Realizable case

Lower Bounds: Hypothesis class

- Assume we sample only |X|/2 examples
  - At least |X|/2 points not observed
- For each such point: $\Pr[c_t(x)=1]= ½$
  - Regardless of the sample
- Error rate:
  - For each unseen point: expected error ½
  - Expected error at least ¼

Lower Bounds: accuracy & confidence

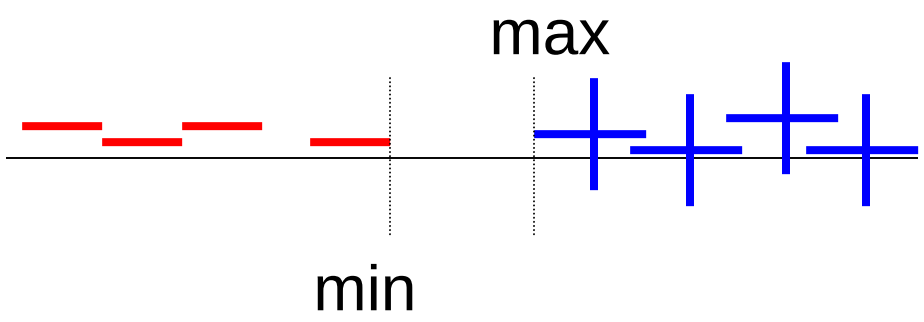
- Let $H=\{c_1,c_2\}$, D s.t. $Pr[c_1 \neq c_2]=4\varepsilon$
  - No function is ε-good for both c_1 and c_2
- Select c_t at random from H
- If for every x in S: $c_1(x)=c_2(x)$
  - we cannot distinguish.
  - $Pr[\forall i: c_1(x_i)=c_2(x_i)] = (1-4\varepsilon)^m$
  - $\delta \geq (1-4\varepsilon)^m \rightarrow m \geq \frac{\log \delta}{\log(1-4\epsilon)} \approx \frac{1}{4\epsilon}\log\frac{1}{\delta}$

Remark on finite precision

- Reducing infinite to finite class
  - assuming finite precision
- Real value on computers is 64 bits
  - Any real value parameter has only 2^{64} values
  - Class size H with d parameters 2^{64d}
    - log |H| = 64d
- Example: hyperplane

Infinite Concept class

- X=[0,1] and H={c_θ | θ in [0,1]}
- $c_\theta(x) = 0$ iff $x < \theta$
- Assume C=H:



- Which c_θ should we choose in [min,max]?

Threshold on a line

- General idea:
  - Take a large sample S of m examples
  - Return some consistent hypothesis c_θ
- Correctness:
  - Show that with probability 1-δ
  - The interval [min,max] has prob. less than ε
    - D([min,max]) $\leq \varepsilon$

Threshold on a line: Proof

- Need to show that
  - $Pr[\ D([min,max]) > \varepsilon\ ] < \delta$
- Proof: By Contradiction.
  - Assume the probability that *x* in [*min*,*max*] is at least ε
    - $D([min,max]) > \varepsilon$
  - Compute the probability over the sample S
    - Each example has prob at least ε to be in [min,max]
    - Probability that no x is in [*min*,*max*] is $(1-\varepsilon)^m$
  - *Need* $\delta \geq (1-\varepsilon)^m \rightarrow m > (1/\varepsilon)\ ln\ (1/\delta)$

Threshold on a line: Proof

- Need to show that
  - $Pr[\ D([min,max]) > \varepsilon\ ] < \delta$
- Proof: By Contradiction
  - Assume the probability that x in [*min*,*max*] at least ε
    - $D([min,max]) > \varepsilon$
  - Compute the probability over the sample S
    - Each example has prob at least ε to be in [min,max]
    - Probability that no x is in [*min*,*max*] is $(1-\varepsilon)^m$
  - *Need* $\delta \geq (1-\varepsilon)^m \rightarrow m > (1/\varepsilon)\ ln\ (1/\delta)$

What is WRONG ?!

Threshold on a line: why wrong?!

- Sample S is a random variable.
- min and max are a function of S.
  - Let's write min(S) and max(S)
    - Define only after we have the sample S
      - First sample S, then set min/max, no more x to sample
- Alternatively:
  - $\Pr_S$ [x in [min(S),max(S)] and x in S]
  - Actually, by definition, this is zero!

Threshold on a line: corrected

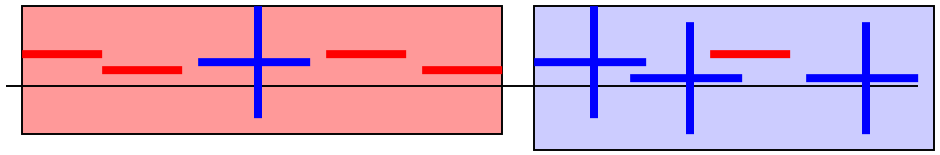
- Define events which do not depend on the sample S
  - Let *min'* be : $D([min', \theta]) = \varepsilon/2$
  - Let *max'* be : $D([\theta, max']) = \varepsilon/2$
- Goal: Show that with high probability
  - *min* in $[min', \theta]$ and *max* in $[\theta, max']$
  - Then any value in [min,max] is good.

Threshold on a line: corrected

- For any sample x:
  - Prob x in [*min*',θ] is ε/2
  - Prob x in [*θ*,*max*'] is ε/2
- Probabilities over S
  - Probability *min* not in [*min*',θ] is $(1-\varepsilon/2)^m$
  - Probability *max* not in [θ,*max*'] is $(1-\varepsilon/2)^m$
- Probability that [min,max] is bad
  - At most $2(1-\varepsilon/2)^m \leq \delta \rightarrow m \geq (2/\varepsilon)\log(2/\delta)$

Threshold: non-Realizable case

- Suppose we sample:



- ERM Algorithm:
  - Find the function h with lowest error!

Threshold: non-realizable

- Try to reduce to a finite class
  - Build an ε-net
- Given a class H define a class G
  - For every h in H
  - There exist a g in G such that
  - $D(g \Delta h) < \varepsilon/4$
- Algorithm: Find the best g in G.

Threshold: non-realizable

- Define: z_i as a $\varepsilon/4$ - net (w.r.t. D)
  - $D([z_i, z_{i+1}]) = \varepsilon/4$
  - $G = \{ c_{z\_i} \}$
    - $|G| = 4/\varepsilon$
- For any c_θ there is a g in G:
  - $|error(g) - error(c_\theta)| \leq \epsilon/4$
- Learn using G
- What is the problem?!

Threshold: proof non-realizable

- Goal: show that ERM works
  - use ε-net only in the proof.
- Sampling errors:
  - For g in G: $|\widehat{error}(g) - error(g)| \leq \frac{\epsilon}{16}$
  - For any c_θ in H there is a g in G
    - $|error(g) - error(c_\theta)| \leq \frac{\epsilon}{4}$
    - For any sample S: $|\widehat{error}(g) - \widehat{error}(c_\theta)| \leq \frac{3\epsilon}{8}$

Threshold: proof non-realizable

- Assume all three conditions hold
- For the best hypothesis h^*
  - Assume g^* is its approx. in G
- $error(h^*) \geq error(g^*) - \frac{\epsilon}{4}$

$$\geq \widehat{error}(g^*) - \frac{\epsilon}{4} - \frac{\epsilon}{16} = \widehat{error}(g^*) - \frac{5\epsilon}{16}$$

Threshold: proof non-realizable

- For the hypothesis selected by ERM h_{erm}
  - Assume g_{erm} is its approx. in G
- $error(h_{erm}) \leq error(g_{erm}) + \frac{\epsilon}{4} \leq$

  $\widehat{error}(g_{erm}) + \frac{5\epsilon}{16} \leq \widehat{error}(h_{erm}) + \frac{11\epsilon}{16}$
- $ERM: \widehat{error}(h_{erm}) \leq \widehat{error}\ (g^*)$

Threshold: proof non-realizable

- $error(h^*) \geq \widehat{error}(g^*) - \frac{5\epsilon}{16}$
- $\widehat{error}(g^*) \geq \widehat{error}(h_{erm})$
- $\widehat{error}(h_{erm}) \geq error(h_{erm}) - \frac{11\epsilon}{16}$
- $error(h^*) \geq error(h_{erm}) - \epsilon$

Threshold: proof non-realizable

- Completing the proof
- Computing probability the conditions hold
  - For g in G: $|\widehat{error}(g) - error(g)| \leq \frac{\epsilon}{16}$
    - G is finite with $\frac{\epsilon}{4}$ functions.
  - For any sample S: $|\widehat{error}(g) - \widehat{error}(c_\theta)| \leq \frac{3\epsilon}{8}$
    - Sufficient that for any $[z_i, z_{i+1}]$ at most $\frac{3\epsilon m}{8}$ samples
      - Expectation $\frac{2\epsilon m}{8}$

Summary

- PAC model
- Generalization bounds
  - Empirical Risk Minimization
  - Finite classes
  - Infinite classes
    - Threshold on interval
- Next class
  - Infinite classes: general methodology