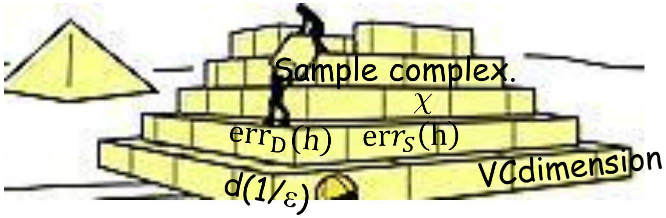
Machine Learning Theory II

Maria-Florina (Nina) Balcan

February 11th, 2015



Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

- E.g.: logistic regression, SVM, Adaboost, etc.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.

Today's focus: Sample Complexity for Supervised Classification (Function Approximation)

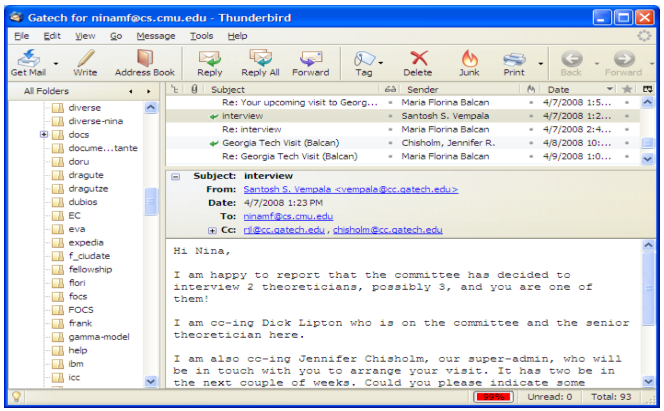
- Statistical Learning Theory (Vapnik)
- PAC (Valiant)

- Recommended reading: Mitchell: Ch. 7
  - Suggested exercises: 7.1, 7.2, 7.7
- Additional resources: my learning theory course!

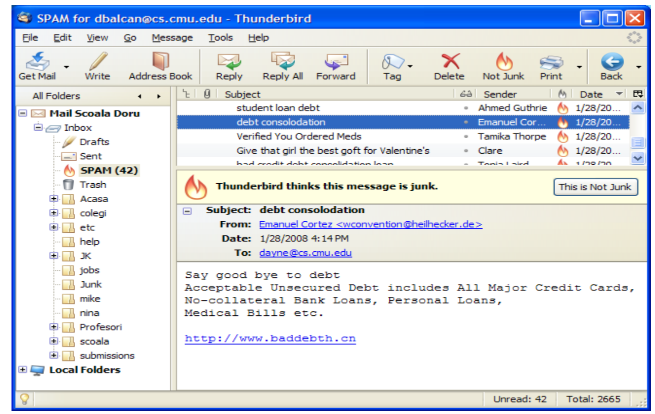
Supervised Learning

- E.g., which emails are spam and which are important.

Not spam



spam

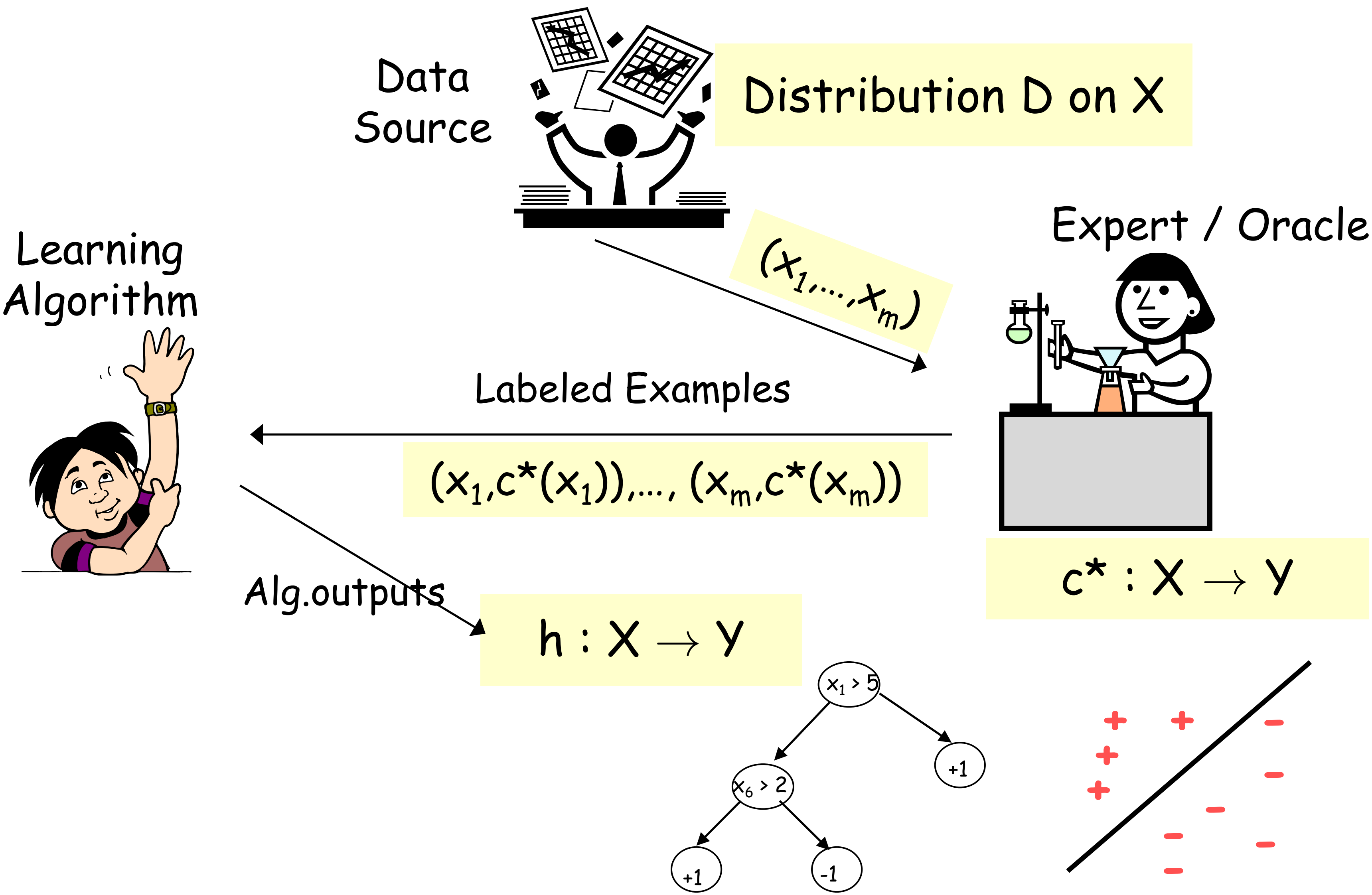


- E.g., classify images as man versus women.

Man

Women

PAC/SLT models for Supervised Learning

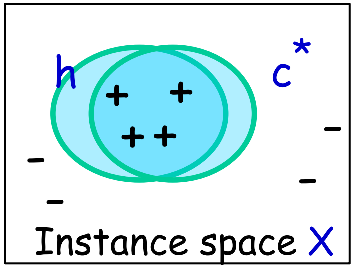


PAC/SLT models for Supervised Learning

- X – feature/instance space; distribution D over X

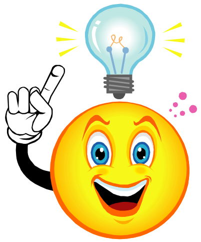
  e.g., $X = R^d$ or $X = \{0,1\}^d$

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
  - labeled examples - drawn i.i.d. from D and labeled by target c^*
  - labels $\in \{-1,1\}$ - binary classification



- Algo does optimization over S, find hypothesis h.
- Goal: h has small error over D.

$$err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$$



Bias: fix hypothesis space H [whose complexity is not too large]

- Realizable: $c^* \in H$.
- Agnostic: c^* "close to" H.

PAC/SLT models for Supervised Learning

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
- Does optimization over S, find hypothesis $h \in H$.
- Goal: h has small error over D.

True error: $\mathrm{err}_D(h) = \Pr_{x\sim D}(h(x) \neq c^*(x))$

How often $h(x) \neq c^*(x)$ over future instances drawn at random from D

- But, can only measure:

Training error: $\mathrm{err}_S(h) = \frac{1}{m}\sum_i I\big(h(x_i) \neq c^*(x_i)\big)$

How often $h(x) \neq c^*(x)$ over training instances

Sample complexity: bound $err_D(h)$ **in terms of** $err_S(h)$

Sample Complexity for Supervised Learning

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

Bound only logarithmic in |H|, linear in $1/\epsilon$

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Probability over different samples of m training examples

So, if $c^* \in H$ and can find consistent fns, then only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1-\delta$

Sample Complexity for Supervised Learning

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Example: H is the class of conjunctions over $X = \{0,1\}^n$. $|H| = 3^n$

E.g., $h = x_1 \overline{x_3} x_5$ or $h = x_1 \overline{x_2} x_4 x_9$

Then $m \geq \frac{1}{\epsilon}\left[n \ln 3 + \ln\left(\frac{1}{\delta}\right)\right]$ suffice

Sample Complexity: Finite Hypothesis Spaces Realizable Case

1) PAC: How many examples suffice to guarantee small error whp.

Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

2) Statistical Learning Way:

With probability at least $1-\delta$, for all $h \in H$ s.t. $err_S(h) = 0$ we have

$$err_D(h) \leq \frac{1}{m}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right).$$

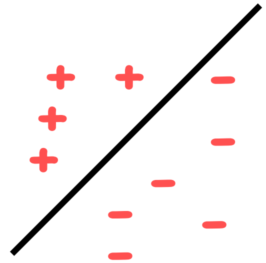
Supervised Learning: PAC model (Valiant)

- X - instance space, e.g., $X = \{0,1\}^n$ or $X = R^n$
- $S_l=\{(x_i, y_i)\}$ - labeled examples drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in$ {-1,1} - binary classification

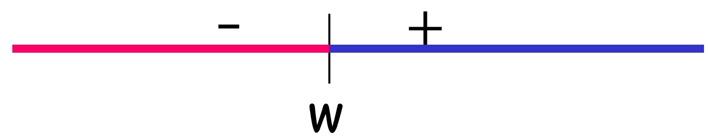
- Algorithm A PAC-learns concept class H if for any target c* in H, any distrib. D over X, any ε, δ > 0:
  - A uses at most poly(n,1/ε,1/δ,size(c*)) examples and running time.
  - With prob. $\geq 1 - \delta$, A produces h in H of error at $\leq \varepsilon$.

What if H is infinite?

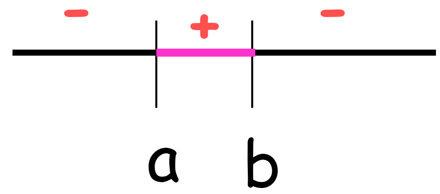
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



Sample Complexity: Infinite Hypothesis Spaces

- H[m] - maximum number of ways to split m points using concepts in H; i.e. $H[m] = \max_{|S|=m} |H[S]|$

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $\mathrm{H[m]} = \mathrm{O}\left(\mathrm{m}^{\mathrm{VCdim(H)}}\right)$

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Effective number of hypotheses

- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

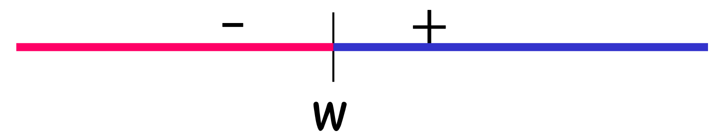
$$H[m] = \max_{|S|=m} |H[S]|$$

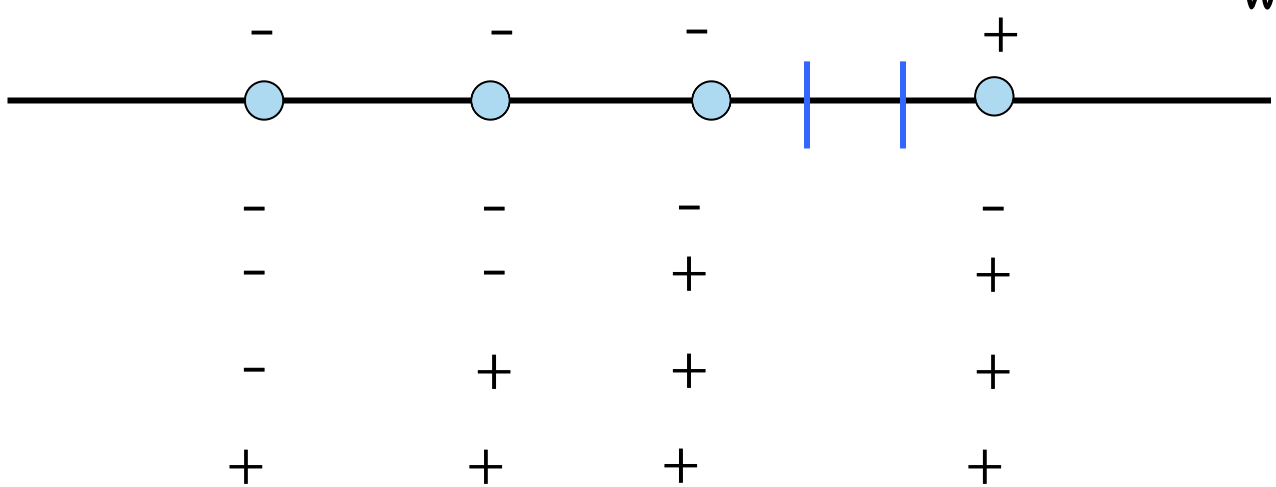
Effective number of hypotheses

- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

E.g., H= Thresholds on the real line





$|H[S]| = 5$

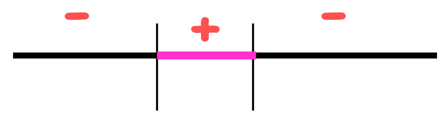
In general, if |S|=m (all distinct), $|H[S]| = m + 1 \ll 2^m$

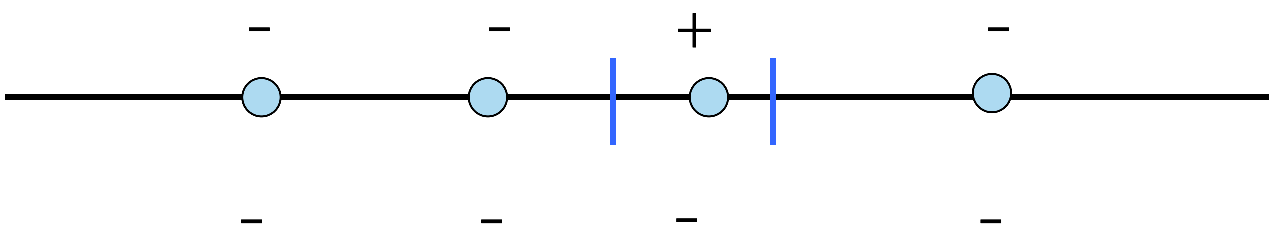
Effective number of hypotheses

- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

E.g., H= Intervals on the real line





In general, |S|=m (all distinct), $H[m] = \frac{m(m+1)}{2} + 1 = O(m^2) \ll 2^m$

There are m+1 possible options for the first part, m left for the second part, the order does not matter, so (m choose 2) + 1 (for empty interval).

Effective number of hypotheses

- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

Definition: H shatters S if $|H[S]| = 2^{|S|}$.

Sample Complexity: Infinite Hypothesis Spaces

- H[m] - max number of ways to split m points using concepts in H

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Very Very Rough Idea:

S= $\{x_1, x_2, \ldots, x_m\}$ i.i.d. from D

B: $\exists\ h \in H$ with $err_S(h) = 0$ but $err_D(h) \geq \epsilon$.

S′ =$\{x'_1, \ldots, x'_m\}$ another i.i.d. "*ghost sample*" from D

B′: $\exists\ h \in H$ with $err_S(h) = 0$ but $err_{S'}(h) \geq \epsilon$.

Claim: To bound P(B), sufficient to bound P(B′)

Over $S \cup S'$ only H[2m] effective hypotheses left… but, no randomness left.

Need randomness to bound the probability of a bad event, another symmetrization trick….

Sample Complexity: Infinite Hypothesis Spaces Realizable Case

H[m] - max number of ways to split m points using concepts in H

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

- Not too easy to interpret sometimes hard to calculate exactly, but can get a good bound using "VC-dimension

If $H[m] = 2^m$, then $m \geq \frac{m}{\epsilon}(\ldots.)$ ☹

- VC-dimension is roughly the point at which H stops looking like it contains all functions, so hope for solving for m.

Sample Complexity: Infinite Hypothesis Spaces

H[m] - max number of ways to split m points using concepts in H

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O\left(m^{VCdim(H)}\right)$

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Shattering, VC-dimension

Definition: H shatters S if $|\mathrm{H}[\mathrm{S}]| = 2^{|S|}$.

A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways, all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(\mathrm{H}) = \infty$

Shattering, VC-dimension

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(\mathrm{H}) = \infty$

To show that VC-dimension is d:

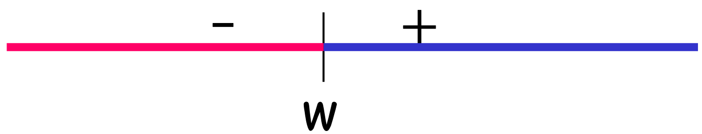
- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

Fact: If H is finite, then $\mathrm{VCdim}(\mathrm{H}) \leq \log(|\mathrm{H}|)$.

Shattering, VC-dimension

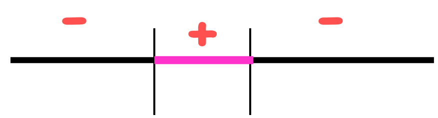
If the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that can be shattered.

E.g., H= Thresholds on the real line

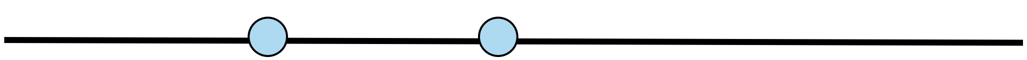


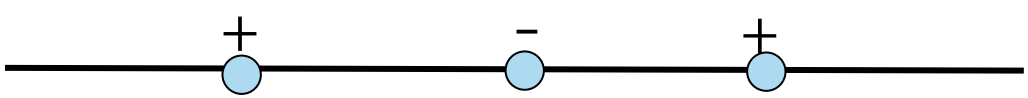
$\mathrm{VCdim}(\mathrm{H}) = 1$

E.g., H= Intervals on the real line



$\mathrm{VCdim}(\mathrm{H}) = 2$

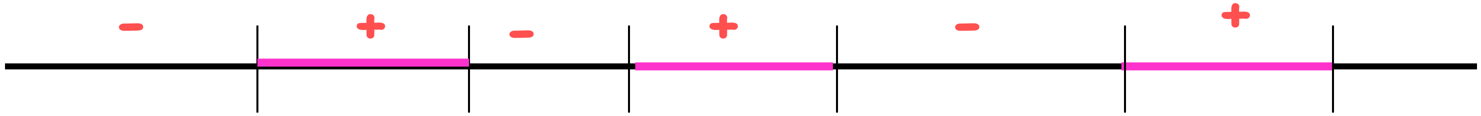




Shattering, VC-dimension

If the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that can be shattered.

E.g., H= Union of k intervals on the real line $\mathrm{VCdim(H)} = 2\mathrm{k}$



$\mathrm{VCdim(H)} \geq 2\mathrm{k}$

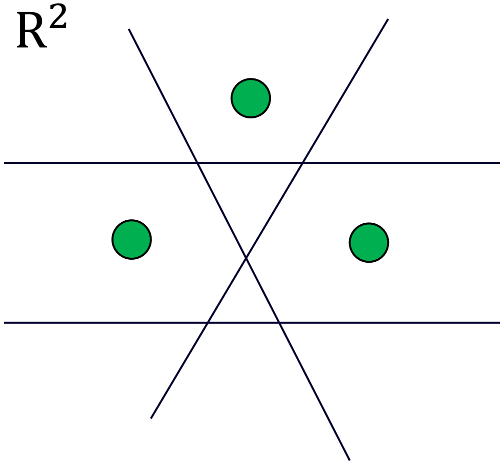
A sample of size 2k shatters (treat each pair of points as a separate case of intervals)

$\mathrm{VCdim(H)} < 2\mathrm{k} + 1$

Shattering, VC-dimension

E.g., **H=** linear separators in R^2

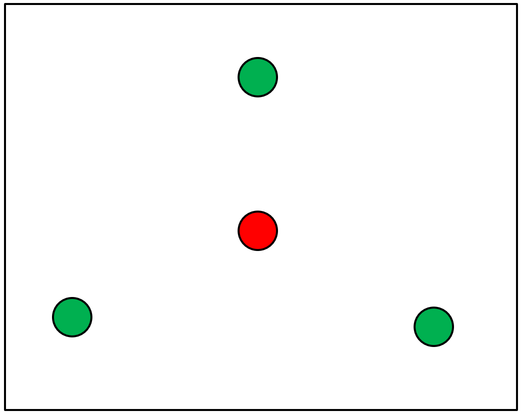
$\mathrm{VCdim(H)} \geq 3$



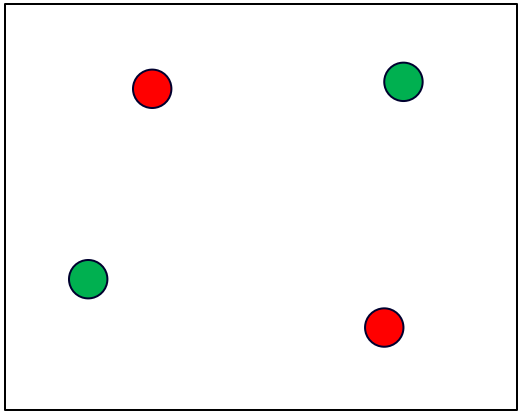
Shattering, VC-dimension

E.g., **H=** linear separators in $\mathbb{R}^2$

$\mathrm{VCdim}(H) < 4$



Case 1: one point inside the triangle formed by the others. Cannot label inside point as positive and outside points as negative.



Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.

Fact: VCdim of linear separators in $\mathbb{R}^d$ is d+1

Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

- $m \leq d$, then $H[m] = 2^m$
- m>d, then $H[m] = O\left(m^d\right)$

Proof: induction on m and d. Cool combinatorial argument!

Hint: try proving it for intervals…

Sample Complexity: Infinite Hypothesis Spaces
Realizable Case

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O\left(m^{VCdim(H)}\right)$

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sample Complexity: Infinite Hypothesis Spaces
Realizable Case

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

E.g., H= linear separators in R^d

$$m = O\left(\frac{1}{\varepsilon}\left[d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

Sample complexity linear in d

So, if double the number of features, then I only need roughly twice the number of samples to do well.

Nearly Matching Bounds

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

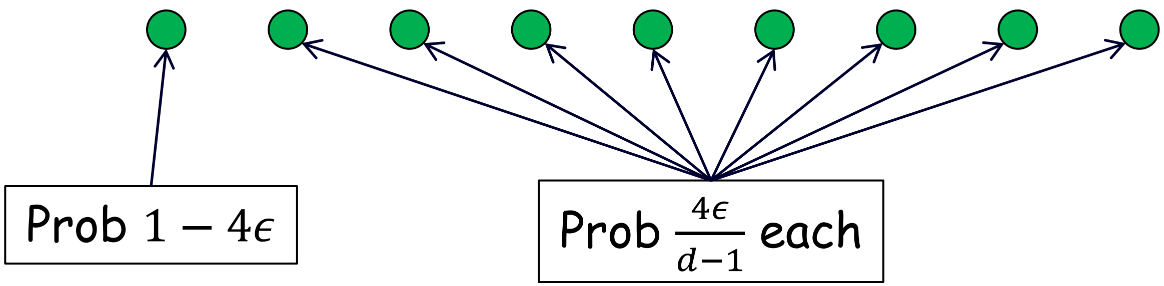
labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Theorem (lower bound):

For any H, any algo A, any $0 < \epsilon < 1/8$, any $\delta < .01$, $\exists$ distr. D and target $c^* \in H$ s.t. if A sees fewer than $\max\left[\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right), \frac{VCdim(H)-1}{32\epsilon}\right]$ examples, then with prob. $\geq \delta$, A produces h with $err_D(h) > \epsilon$.

Lower Bound (simpler form)

- Theorem: For any H there exists D such that any algorithm needs $\Omega\left(\frac{VCdim(H)}{\epsilon}\right)$ examples to reach error ϵ with prob $\geq \frac{3}{4}$.
- Proof: consider $d = VCdim(H)$ shattered points:



- Consider target c^* that labels these points randomly.
- Unless I see roughly $\frac{1}{2}$ of the rare points, have error $\geq \epsilon$
- Each example has only prob 4ϵ of being one of the rare points, and need to see $\frac{d-1}{2}$ rare points, so need to see $\Omega\left(\frac{d}{\epsilon}\right)$ total.

What if $c^* \notin H$?

Uniform Convergence

Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

- This basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in H$ (agnostic case)?

- What can we say if $c^* \notin H$?
- Can we say that whp all $h \in H$ satisfy $|err_D(h) - err_S(h)| \leq \varepsilon$?
  - Called "uniform convergence".
  - Motivates optimizing over S, even if we can't find a perfect function.

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Agnostic Case

What if there is no perfect h?

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

To prove bounds like this, need some good tail inequalities.

Hoeffding bounds

Consider coin of bias p flipped m times.
Let N be the observed # heads. Let $\varepsilon \in [0,1]$.

Hoeffding bounds:

- $\Pr[N/m > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, and
- $\Pr[N/m < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.

Exponentially decreasing tails

- Tail inequality: bound probability mass in tail of distribution (how concentrated is a random variable around its expectation).

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

- Proof: Just apply Hoeffding.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.
- So, whp, best on sample is ε-best over D.
  - Note: this is worse than previous bound ($1/\varepsilon$ has become $1/\varepsilon^2$), because we are asking for something stronger.
  - Can also get bounds "between" these two.

What you should know

- Notion of sample complexity.
- Understand reasoning behind the simple sample complexity bound for finite H.
- Shattering, VC dimension as measure of complexity, Sauer's lemma, form of the VC bounds (upper and lower bounds).