Generalization and Overfitting

Sample Complexity Results for Supervised Classification

Maria-Florina (Nina) Balcan

March 1st, 2019

Admin

Midterm: in class, March 4th.

Closed book.

Allowed to bring one sheet of notes (front and back).

Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

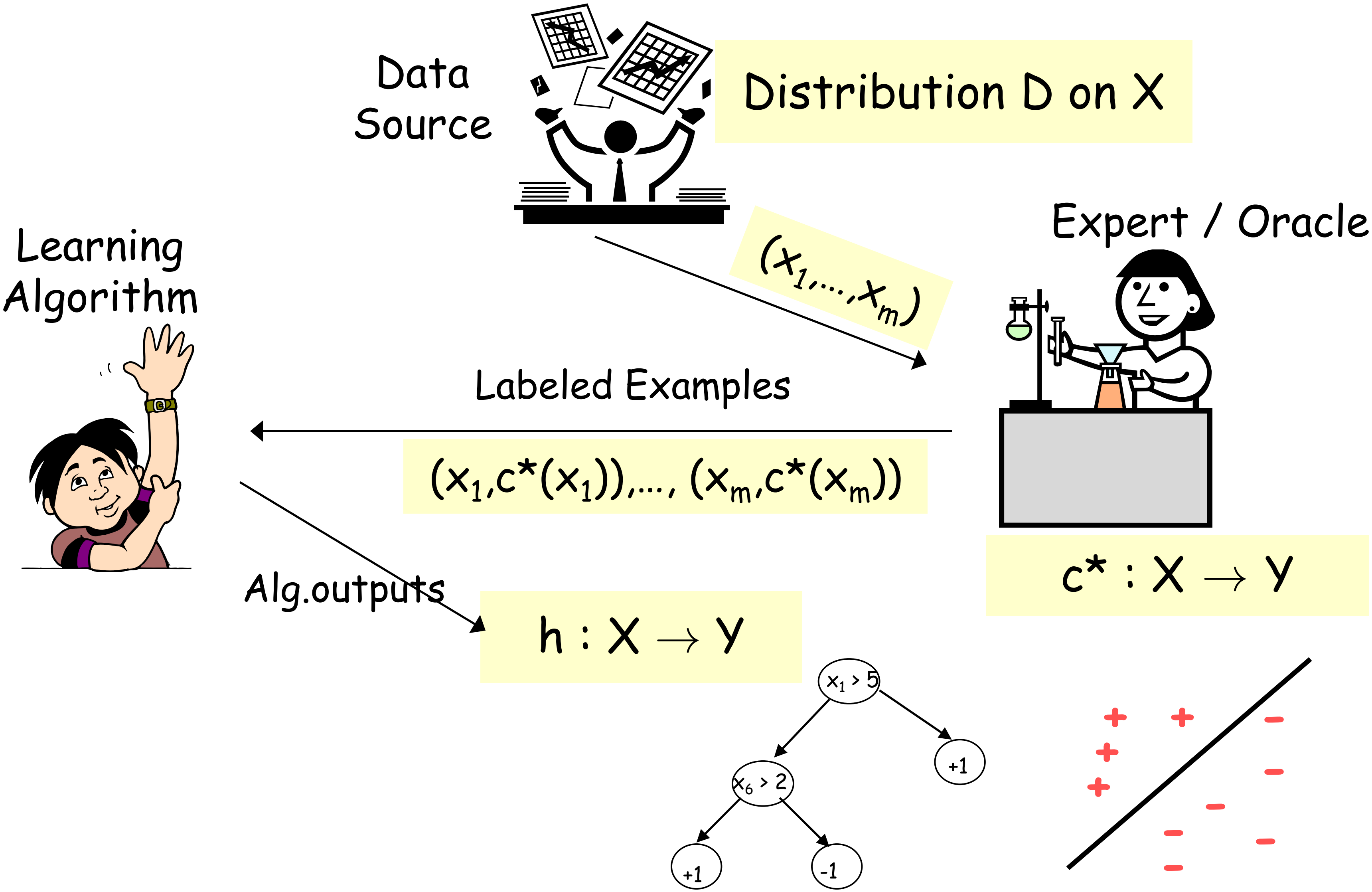
- E.g.: logistic regression, SVM, Adaboost, etc.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.

PAC/SLT models for Supervised Learning



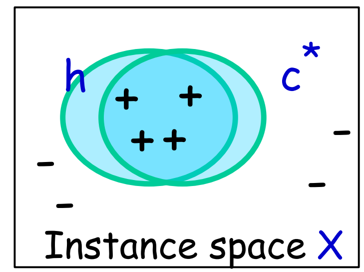
PAC/SLT models for Supervised Learning

- 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: $(x_1,c^*(x_1)),\dots,(x_m,c^*(x_m))$, 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: $\mathrm{err}_S(h) = \frac{1}{m}\sum_i I\left(h(x_i) \neq c^*(x_i)\right)$

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

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

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 $err_S(h) > 0$.

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

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 $err_S(h) > 0$.

Proof Assume k bad hypotheses $h_1, h_2, \ldots, h_k$ with $err_D(h_i) \geq \epsilon$

1) Fix h_i. Prob. h_i consistent with first training example is $\leq 1-\epsilon$.

   Prob. h_i consistent with first m training examples is $\leq (1-\epsilon)^m$.

2) Prob. that at least one h_i consistent with first m training examples is $\leq k\,(1-\epsilon)^m \leq |H|(1-\epsilon)^m$.

3) Calculate value of m so that $|H|(1-\epsilon)^m \leq \delta$

3) Use the fact that $1-x \leq e^{-x}$, sufficient to set $|H|(1-\epsilon)^m \leq |H|\, e^{-\epsilon m} \leq \delta$

What if $c^* \notin H$?

Sample Complexity: Uniform Convergence

Agnostic Case

Empirical Risk Minimization (ERM)

- Input: S: $(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$
- Output: Find h in H with smallest $\mathrm{err}_S(h)$

Theorem

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

labeled examples are sufficient s.t. with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

$1/\epsilon^2$ dependence [as opposed to $1/\epsilon$ for realizable]

Sample Complexity: Uniform Convergence
Agnostic Case

Empirical Risk Minimization (ERM)

- Input: S: $(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$
- Output: Find h in H with smallest $\mathrm{err}_S(h)$

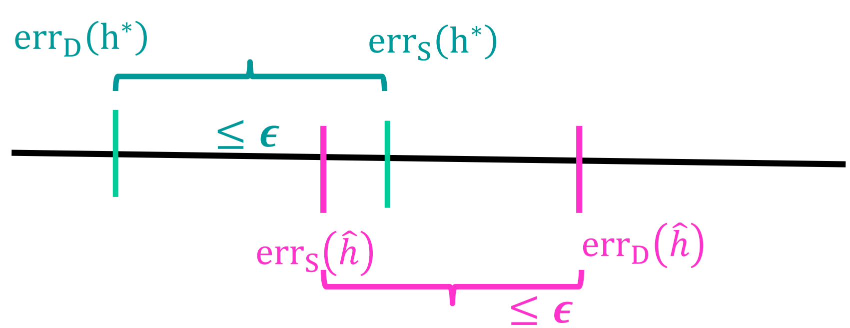
Theorem

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

labeled examples are sufficient s.t. with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

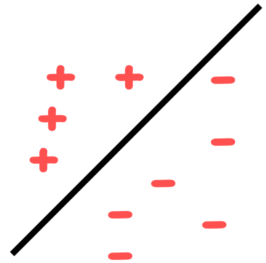
Fact:

W.h.p. $\geq 1-\delta$, $\mathrm{err}_D(\hat{h}) \leq \mathrm{err}_D(h^*) + 2\epsilon$, $\hat{h}$ is ERM output, h^* is hyp. of smallest true error rate.

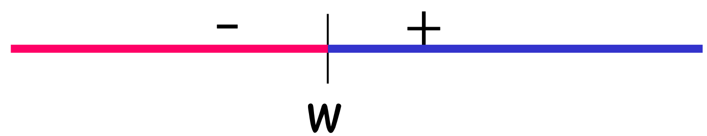


What if H is infinite?

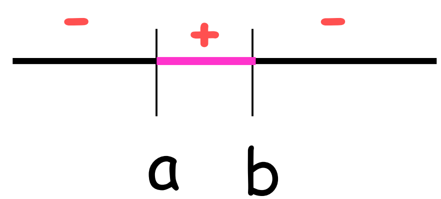
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



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

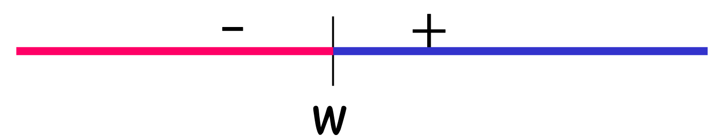
$$H[m] = \max_{|S|=m} |H[S]|$$

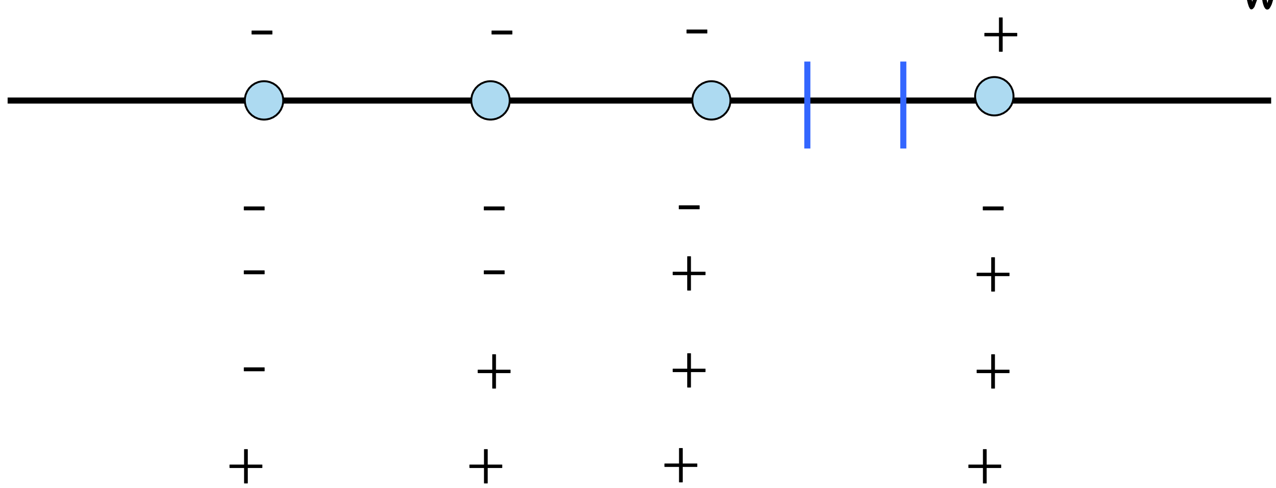
Effective number of hypotheses

- 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

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

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





$|H[S]| = 5$

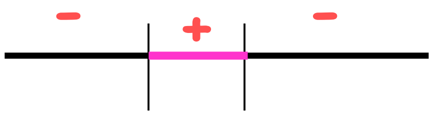
In general, if |S|=m (all distinct), $|H[S]| = m + 1 \ll 2^m$

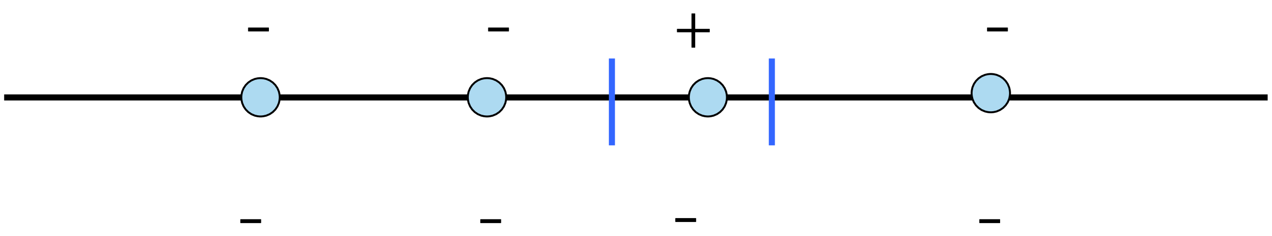
Effective number of hypotheses

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

- 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: $\mathrm{H[m]} = \mathrm{O}\left(\mathrm{m}^{\mathrm{VCdim(H)}}\right)$

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 $err_S(h) > 0$.

Shattering, VC-dimension

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

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

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

To show that VC-dimension is d:

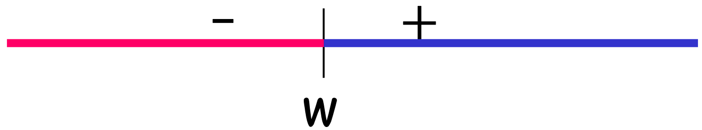
- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

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

Shattering, VC-dimension

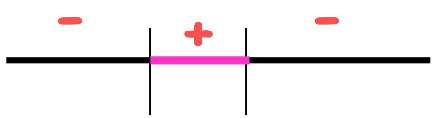
If the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that can be shattered.

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

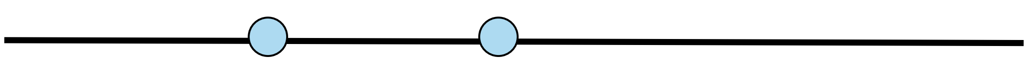


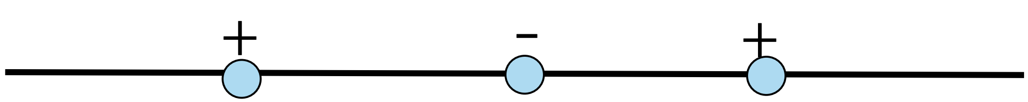
$\mathrm{VCdim}(\mathrm{H}) = 1$

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



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

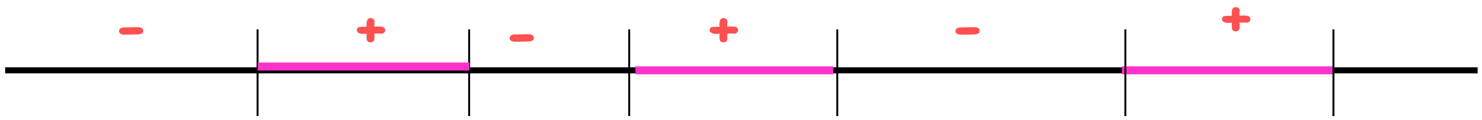




Shattering, VC-dimension

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



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

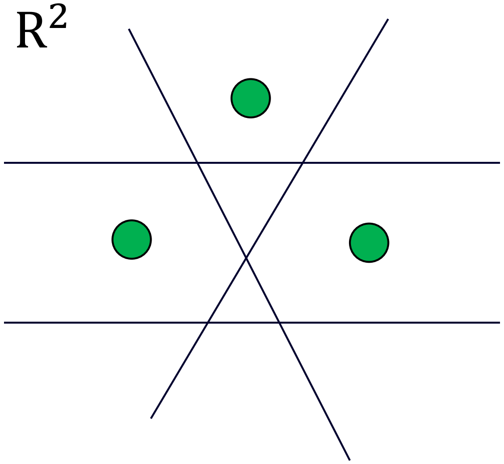
A sample of size 2k shatters (treat each pair of points as a separate case of intervals)

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

Shattering, VC-dimension

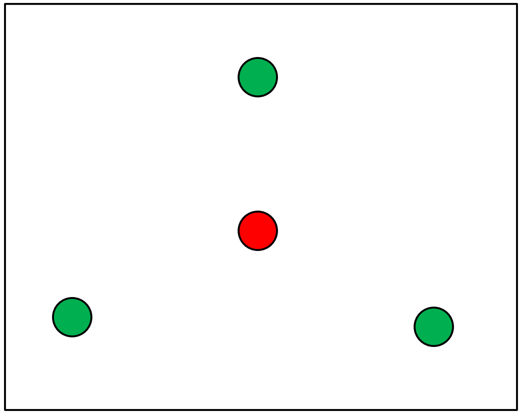
E.g., **H=** linear separators in $\mathbb{R}^2$

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



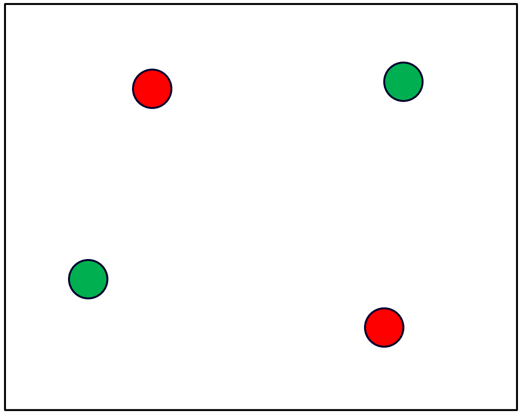
Shattering, VC-dimension

E.g., **H=** linear separators in 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 R^d is d+1

Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

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

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

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 $err_S(h) > 0$.

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.

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 $err_S(h) > 0$.

Statistical Learning Theory Style

$$\mathrm{err_D(h)} \leq \mathrm{err_S(h)} + \sqrt{\frac{1}{2\mathrm{m}}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right)}.$$

What you should know

- Notion of sample complexity.
- Shattering, VC dimension as measure of complexity, Sauer's lemma, form of the VC bounds.