

Computational Learning Theory

Machine Learning

Daniele Loiacono



POLITECNICO
MILANO 1863

Outline and References

□ Outline

- ▶ Basics [ML 7.1,7.2]
- ▶ PAC-Learning [ML 7.3]
- ▶ VC Dimension [ML 7.4]

□ References

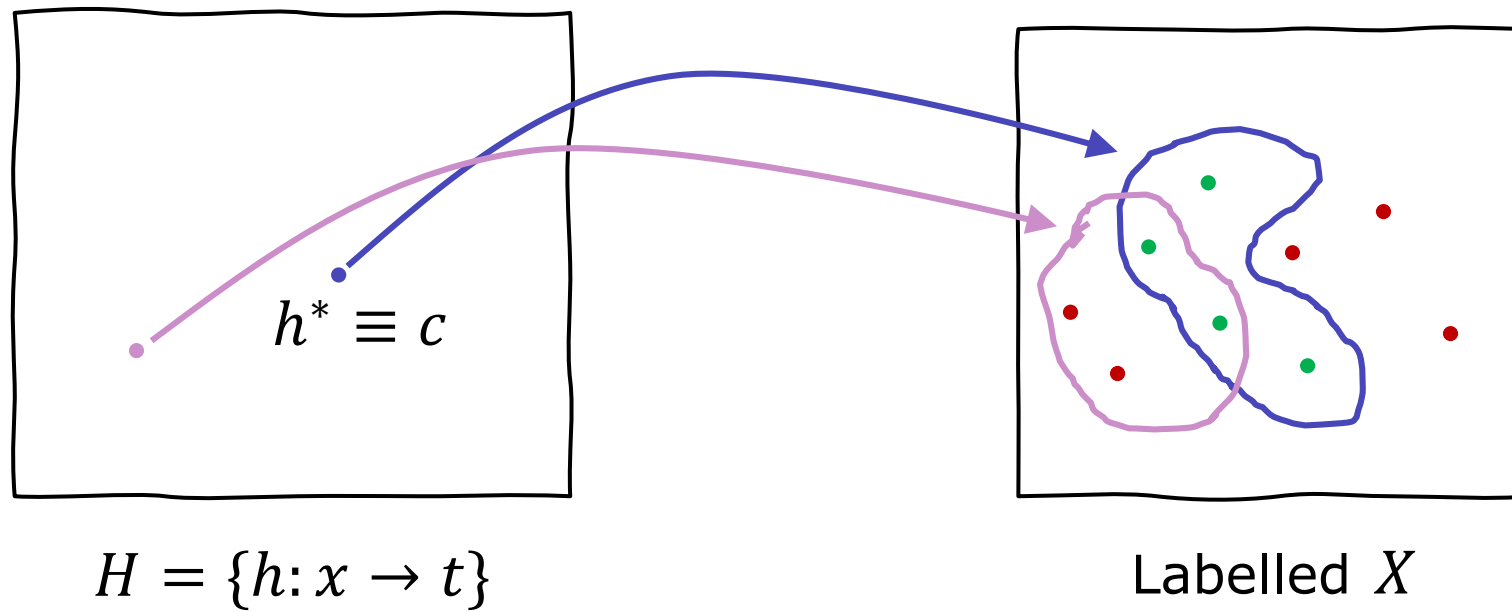
- ▶ Machine Learning, Mitchell [ML]



What is computational learning theory about?

- ❑ It aims at studying the general laws of inductive learning, by modeling:
 - ▶ Complexity of hypothesis space
 - ▶ Bound on training samples
 - ▶ Bound on accuracy
 - ▶ Probability of successful learning
 - ▶ ...
- ❑ This allows to answer questions like:
 - ▶ How many training samples does a learner need to **converge** (with some probability) to a **successful** (with some minimum accuracy) hypothesis?
 - ▶ What will be the performance of the learner on unseen data?
 - ▶ When a concept is learnable?
 - ▶ ...

(Let's Go Back to) The Big Picture



- ❑ A learner (L) wants to learn a *concept* (c) that maps the data in the input space (X) to a target (t)
- ❑ Let assume that L found an hypothesis h^* with no errors on the training data
- ❑ How many training samples of X are necessary to be sure that L actually learned the true concept, i.e., $h^* \equiv c$?

«No Free Lunch» Theorems

- Let $ACC_G(L)$ be the **generalization accuracy** of learner L , i.e., the accuracy of L on samples that are **not** in the training set
- Let \mathcal{F} be the set of all the possible concepts $y = f(\mathbf{x})$
- For any learner L and any possible training set:

$$\frac{1}{|\mathcal{F}|} \sum_{\mathcal{F}} Acc_G(L) = \frac{1}{2}$$

- ▶ **Proof Sketch:** for every concept f where $ACC_G(L) = 0.5 + \delta$, exists a concept f' where $ACC_G(L) = 0.5 - \delta$: $\forall \mathbf{x} \in \mathcal{D}, f'(\mathbf{x}) = f(\mathbf{x}); \forall \mathbf{x} \notin \mathcal{D}, f'(\mathbf{x}) \neq f(\mathbf{x})$
- ▶ **Corollary:** for any two learners, L_1 and L_2 , if $\exists f$ where $ACC_G(L_1) > ACC_G(L_2)$ then $\exists f'$ where $ACC_G(L_2) > ACC_G(L_1)$

«No Free Lunch» Theorems

□ Let $ACC_G(L)$ be the **generalization accuracy** of learner L , i.e., the accuracy of L on samples that are **not** in the training set

□ Let

□ For

What does this mean in practice?

There is no such thing as a winner-takes-all in ML!

► **In ML we always operate under some assumptions!** except f'

► **Consider two learners L_1 and L_2 . If $ACC_G(L_1) > ACC_G(L_2)$ then $\exists f'$ where $ACC_G(L_2) > ACC_G(L_1)$**

Probably Learning an Approximately Correct Hypothesis

Basics

□ Problem setting

- ▶ Let X be the **instance space**
- ▶ Let $H = \{h: X \rightarrow \{0,1\}\}$ be the **hypothesis space** of L
- ▶ Let $C = \{c: X \rightarrow \{0,1\}\}$ be the **set of all the possible target functions (concepts)** we might want to learn
- ▶ Let \mathcal{D} be **training data** drawn from a **stationary distribution $P(X)$** and **labeled (without noise) according to a concept c we want to learn**

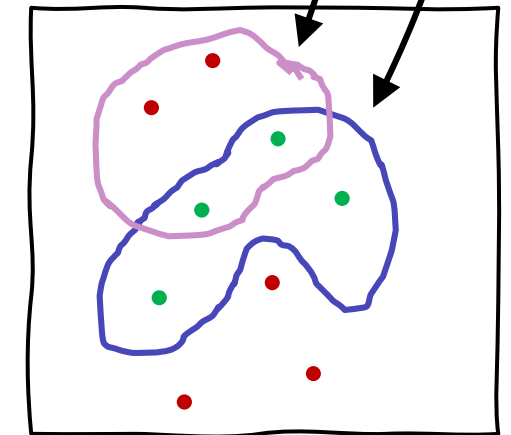
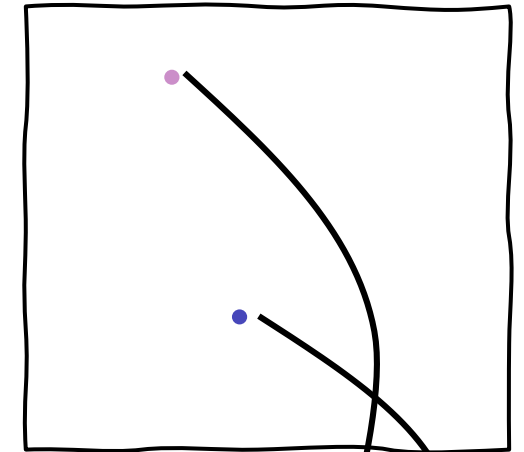
i.e. what a learner can learn...

we try to learn one c of C .

- **A learner L outputs a hypothesis $h \in H$ such that**

$$h^* = \arg \min_{h \in H} error_{train}(h)$$

$$H = \{h: X \rightarrow \{0,1\}\}$$



Labelled X

How do we compute the error?

- We compute the error of a hypothesis as the probability of misclassifying a sample:

$$error_{\mathcal{D}}(h) = \Pr_{x \in \mathcal{D}}[h(x) \neq c(x)] = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} I(h(x) \neq c(x))$$

\mathcal{D} is the training data



- This is the training error! We are interested in the true error of h :

$$error_{true}(h) = \Pr_{x \sim P(X)}[h(x) \neq c(x)]$$

$P(X)$ is the input space distribution

unfortunately we don't have access to $P(X)$...

Can we exploit the inductive bias?

- We say that h overfits the training data if $error_{true} > error_{\mathcal{D}}$...
- ? □ ... but can we bound $error_{true}$ given $error_{\mathcal{D}}$??
- Let assume...
 - ▶ $error_{true}$ is the probability of making a mistake on a sample
 - ▶ we can compute $error_{\mathcal{D}}$ that is the average error probability on \mathcal{D}
 - ▶ assuming a Bernoulli distribution for the error probability, the 95% CI is:

so using this
I could bound
the true error!

$$error_{true}(h) = error_{\mathcal{D}}(h) \pm \underbrace{1.96}_{95\%} \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

95% confidence
interval for a
Bernoulli R.V.

- Is this correct? No! Because \mathcal{D} is the training data and not independent of h
- So, we need to bound the error under more strict assumptions



number of observations we have.

NOT CORRECT

Version Space (1)

- A hypothesis h is **consistent** with a training dataset \mathcal{D} of the concept c if and only if $h(x) = c(x)$ for each training sample in \mathcal{D}

$$\text{Consistent}(h, \mathcal{D}) \stackrel{\text{def}}{=} \forall \langle x, c(x) \rangle \in \mathcal{D}, h(x) = c(x)$$

- The version space, $VS_{H, \mathcal{D}}$ with respect to hypothesis space H and labelled dataset \mathcal{D} , is the subset of hypotheses in H consistent with \mathcal{D}

$$VS_{H, \mathcal{D}} \stackrel{\text{def}}{=} \{h \in H \mid \text{Consistent}(h, \mathcal{D})\}$$

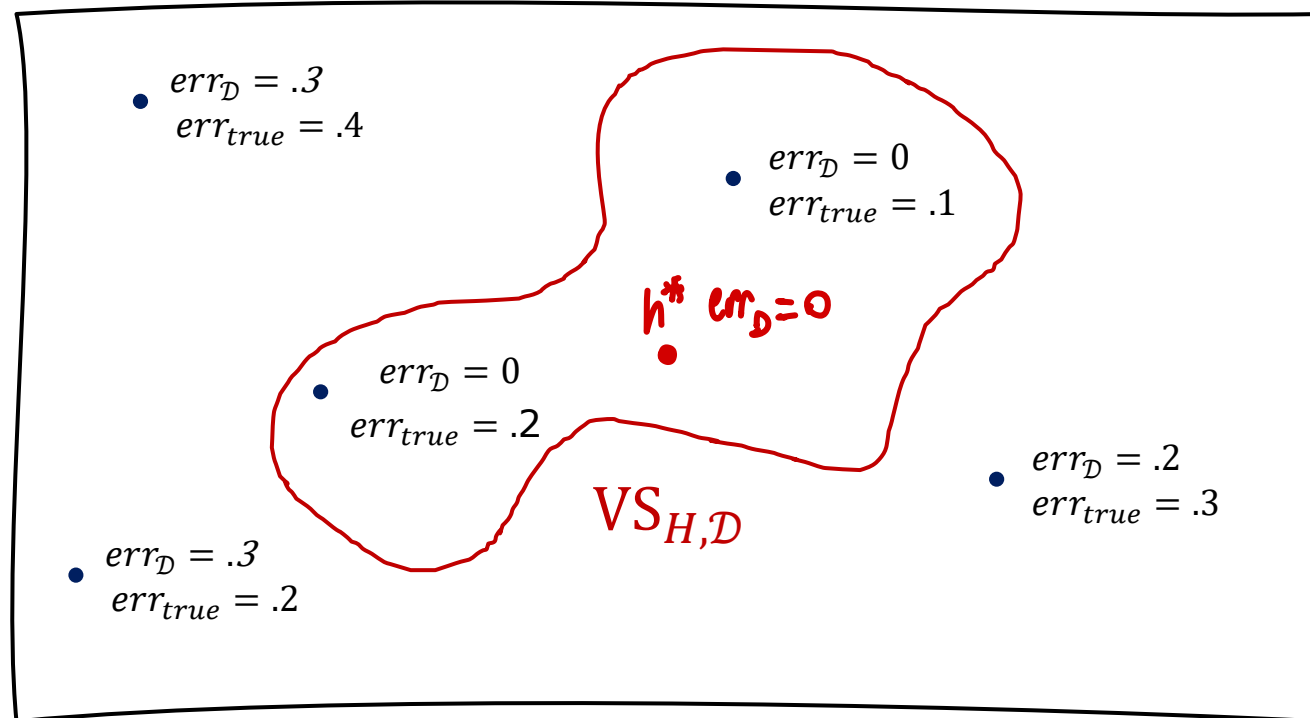
- From now on, we consider only **consistent learners**, that always output a **consistent hypothesis**, i.e., an hypothesis in $VS_{H, \mathcal{D}}$, assuming it is not empty
- Can we bound the $error_{true}$ of a consistent learner?

1) The learner makes no mistake on train data;
2) The learned hypothesis is one of the elts of $VS_{H, \mathcal{D}}$

Now?

Version Space (2)

Hypothesis space H



- If we wish to bound the $error_{true}$ of a consistent learner, we need to find a bound for all the hypotheses in $VS_{H,D}$

Bound for Consistent Learners ← solution to our problem.

If the hypothesis space H is finite and \mathcal{D} is a sequence of $N \geq 1$ independent random examples of some target concept c , then for any $0 \leq \varepsilon \leq 1$, the probability that $VS_{H,\mathcal{D}}$ contains a hypothesis error greater than ε is less than $|H|e^{-\varepsilon N}$

$$Pr(\underbrace{\exists h \in H : error_{\mathcal{D}}(h) = 0}_{\text{error on training set.}} \wedge \underbrace{error_{true}(h) \geq \varepsilon}_{\text{above a threshold.}}) \leq |H|e^{-\varepsilon N}$$

Proof

$$Pr((error_{\mathcal{D}}(h_1) = 0 \wedge error_{true}(h_1) \geq \varepsilon) \vee \dots \vee (error_{\mathcal{D}}(h_{|VS_{H,\mathcal{D}}|}) = 0 \wedge error_{true}(h_{|VS_{H,\mathcal{D}}|}) \geq \varepsilon))$$

$$\leq \sum_{h \in VS_{H,\mathcal{D}}} Pr(error_{\mathcal{D}}(h) = 0 \wedge error_{true}(h) \geq \varepsilon) \quad (\text{Union bound})$$

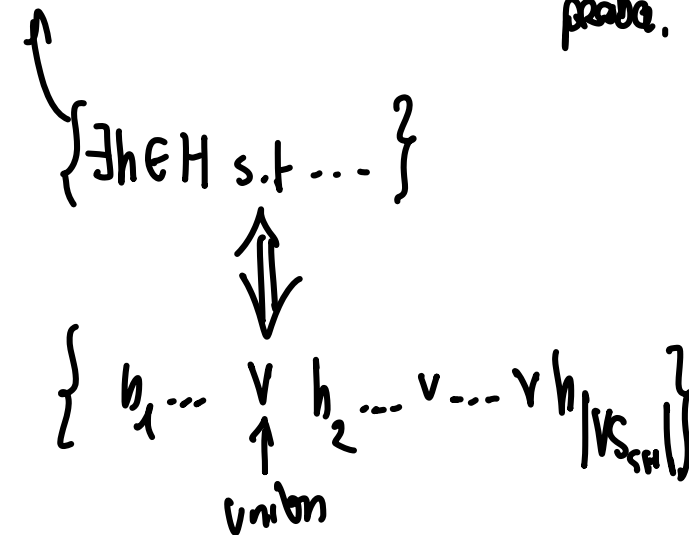
$$\leq \sum_{h \in VS_{H,\mathcal{D}}} Pr(error_{\mathcal{D}}(h) = 0 | error_{true}(h) \geq \varepsilon) \quad (\text{Bound using Bayes' rule})$$

$$\leq \sum_{h \in VS_{H,\mathcal{D}}} (1 - \varepsilon)^N \quad (\text{Bound on individual } h)$$

$$\leq |H|(1 - \varepsilon)^N \quad (|VS_{H,\mathcal{D}}| \leq |H|)$$

$$\leq |H|e^{-\varepsilon N} \quad (1 - \varepsilon \leq e^{-\varepsilon}, \text{ for } 0 \leq \varepsilon \leq 1)$$

the larger is N ,
the smaller the probab.



What does it mean in practice?

- Let say that δ is the probability to have $error_{true} > \varepsilon$ for a consistent hypothesis:

$$\mathbb{P}\left(\exists h \in H : \begin{cases} error_D(h) = 0 \\ error_{true}(h) \geq \varepsilon \end{cases}\right) \leq |H| e^{-\varepsilon N} \leq \delta$$

← let's choose a level δ ... we can then find N to respect δ and ε .

- We can bound N after setting ε and δ :

The nb of samples to have the proba bound δ :

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

- We can bound ε after setting N and δ :

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

if N is already fixed & we cannot change it.

What does it mean in practice?

- Let say that δ is the probability to have $error_{true} > \varepsilon$ for a consistent hypothesis:

$$|H|e^{-\varepsilon N} \leq \delta$$

- We can bound N after setting ε and δ :

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

- We can bound ε after setting N and δ :

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

Can be exponential in
#features

Example: Conjunction of up to N Boolean Literals

□ Consider a classification problem

- ▶ Instance space is $X = \langle x_1, x_2, x_3, x_4 \rangle$, where x_i is a boolean variable
- ▶ Each hypothesis h is a rule like this: *in each x_i : 1 or 0 or ? (*)*
 $\text{if } (x_1 = 1, x_2 = ?, x_3 = 0, x_4 = 1) \text{ then } y = 1, \text{ otherwise } y = 0$

□ How many samples N are necessary to guarantee that, with a probability at least of 0.99, the error of a consistent hypothesis is not greater than 0.05?

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

$\boxed{0.05}$ points to ε
 $\boxed{3^4}$ points to $|H|$ with a note $(*)$
 $\boxed{0.01}$ points to δ
 $\boxed{N \geq 180}$ is shown in a box
Handwritten notes: *Impossible here since $|H| = 3^4 = 81$.*
if we want 99% confidence $\rightarrow \delta = 0.01$.

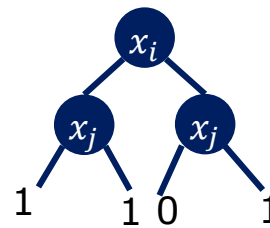
□ How does it scale with respect to the number of variables (M)?

- ▶ $M=10 \rightarrow N \geq 312$
- ▶ $M=100 \rightarrow N \geq 2290$ *\rightsquigarrow "only 2290".*

Example: Decision Tree (depth=2) *SIMILARLY*

□ Consider a classification problem

- ▶ Instance space is $X = \langle x_1, \dots, x_M \rangle$, where x_i is a boolean variable
- ▶ Each hypothesis h is a rule is a decision trees of depth 2 using only two variables:



□ How many samples N are necessary to guarantee that, with a probability at least of 0.99, the error of a consistent hypothesis is not greater than 0.05?

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

Annotations:

- ε is 0.05
- $|H|$ is $\frac{M(M-1)}{2} 16$
- δ is 0.01

Probably Learning an Approximately Correct Hypothesis

- Considering a class C of possible target concepts defined over an instance space X with an encoding length M , and a learner L using a hypothesis space H we define:

C is PAC-learnable by L using H if for all $c \in C$, for any distribution $P(X)$, ε (such that $0 < \varepsilon < 1/2$), and δ (such that $0 < \delta < 1/2$), learner L will with a probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_{true}(h) \leq \varepsilon$, in time that is polynomial in $1/\varepsilon$, $1/\delta$, M , and $size(c)$.


- So, PAC-learnability is only about computational complexity? What about the complexity with respect to the number of training samples N ?
- A sufficient condition to prove PAC-learnability is proving that a learner L requires only a polynomial number of training examples, and processing per example is polynomial

↪ computational complexity ↔ bound on nb of training samples.

Agnostic Learning PB of previous bound: it requires **CONSISTENT** learners (no error on training set).

- ❑ So far, we assumed that $c \in H$, or at least that $VS_{H,\mathcal{D}}$ is not empty, and the learner L will always output a hypothesis h such that $error_{\mathcal{D}}(h) = 0$
- ❑ But in general (**agnostic**) learner will output a hypothesis h such that $error_{\mathcal{D}}(h) > 0$
- ❑ Can we bound $error_{true}(h)$ given $error_{\mathcal{D}}(h)$?

If the hypothesis space H is **finite** and \mathcal{D} is a sequence of $N \geq 1$ i.i.d. examples of some target concept c , then for any $0 \leq \varepsilon \leq 1$, and for any learned hypothesis h , the probability that $error_{true}(h) - error_{\mathcal{D}}(h) > \varepsilon$ is less than $|H|e^{-2N\varepsilon^2}$

$$Pr(\exists h \in H : error_{true}(h) > \underbrace{error_{\mathcal{D}}(h) + \varepsilon}) \leq |H|e^{-2N\varepsilon^2}$$


❑ Can we bound $error_{true}(h)$ given $error_{\mathcal{D}}(h)$?

Agnostic Learning - Proof

- **Additive Hoeffding Bound:** let $\hat{\theta}$ be the empirical mean of N i.i.d. Bernoulli random variables with mean θ :

$$Pr(\theta > \hat{\theta} + \varepsilon) \leq e^{-2N\varepsilon^2} \quad N \text{ iid } \mathcal{B}(\theta) \text{ RV}$$

True mean \nearrow θ \nwarrow *Empirical mean* $\hat{\theta}$

- So for any single hypothesis h :

$$Pr(error_{true}(h) > error_{\mathcal{D}}(h) + \varepsilon) \leq e^{-2N\varepsilon^2}$$

- As we want this to be true for all the hypothesis in H :

$$Pr(\exists h \in H : error_{true}(h) > error_{\mathcal{D}}(h) + \varepsilon) \leq \underbrace{|H|}_{\text{union bound}} e^{-2N\varepsilon^2}$$

Bounds for Agnostic Learning

- Similarly to what done before, we can bound the sample complexity:

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

- We can also bound the true error of the hypothesis as:

$$error_{true}(h) \leq error_{\mathcal{D}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

what we wanted to do @ the beginning!

- We found the bias and variance decomposition we previously saw in the course!

UNFORTUNATELY, for usual problems, $|H| = +\infty$ } Example for linear regression.
↳ we need a new set of bounds, not using $|H|$ (size of hypothesis space).

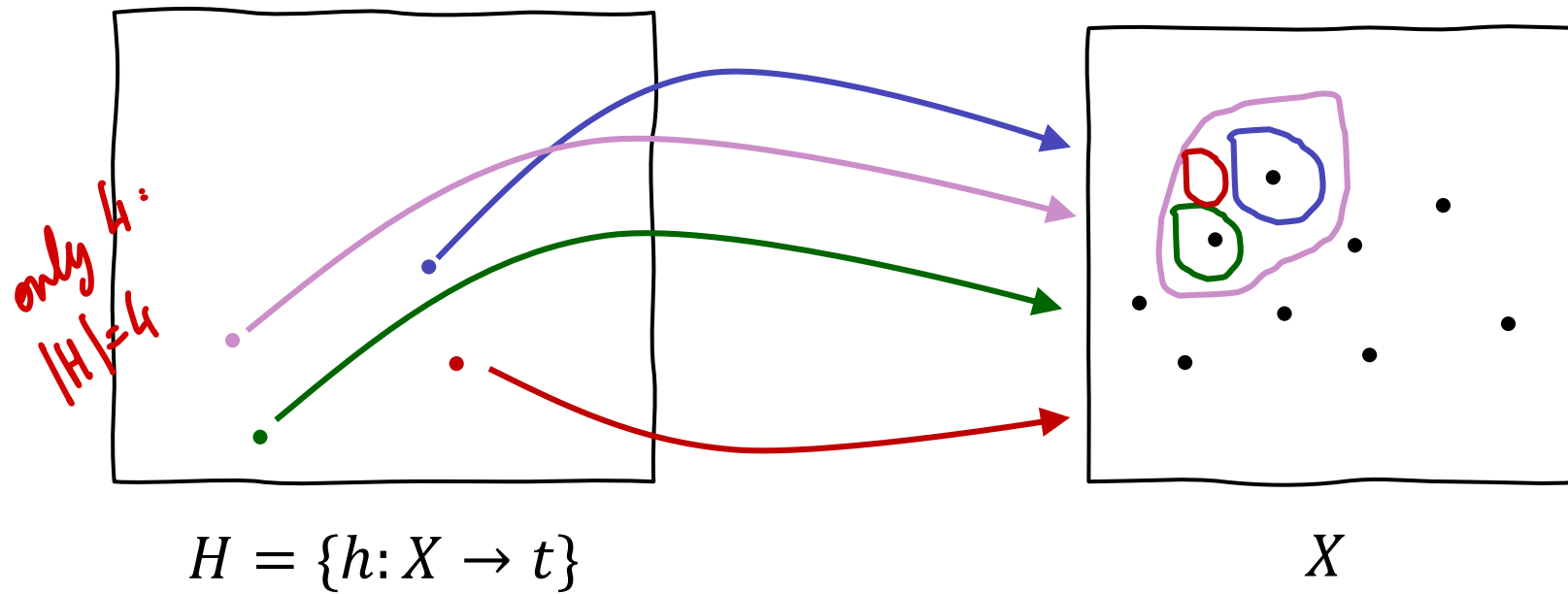
PAC-Learning with Infinite Hypotheses Spaces

- Previously we found this PAC-Learning bound for the number of samples:

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

- If $|H|$ is infinite, what does this mean? What can we use instead of $|H|$?
- The answer is the largest subset of X for which $|H|$ can guarantee a zero training error (regardless of the target function c)
- We call **VC dimension** the size of this subset

Intuition Behind Using VC Dimension



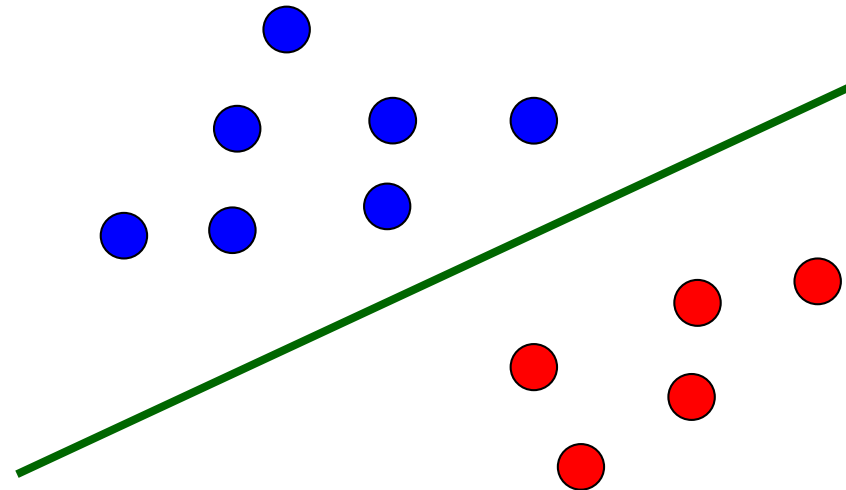
- ❑ Let assume that $|X|=N$, how big is $|C|$?
- ❑ Assuming $|H| = 2^N$, we can always find $h \in H$ with $error_D(h)=0$
- ❑ Does $error_D(h)$ tells something more on the error on other samples in X ?
- ❑ What happens instead if with H we can classify correctly no more than 2 training samples?

VC Dimension

- We define a **dichotomy** of a set S of instances as a partition of S into two disjoint subsets, i.e., labeling each instance in S as positive or negative
- We say that a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy
- The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H over instance space X , is the largest finite subset of X shattered by H

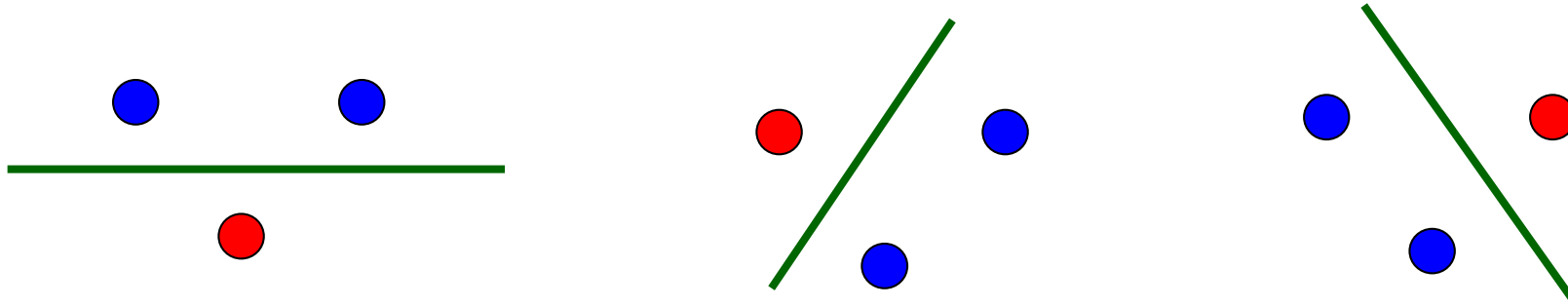
Example: VC dimension of linear classifier

- What about a linear classifier in 2D input space?




Example: VC dimension of linear classifier

□ What about a linear classifier in 2D input space?

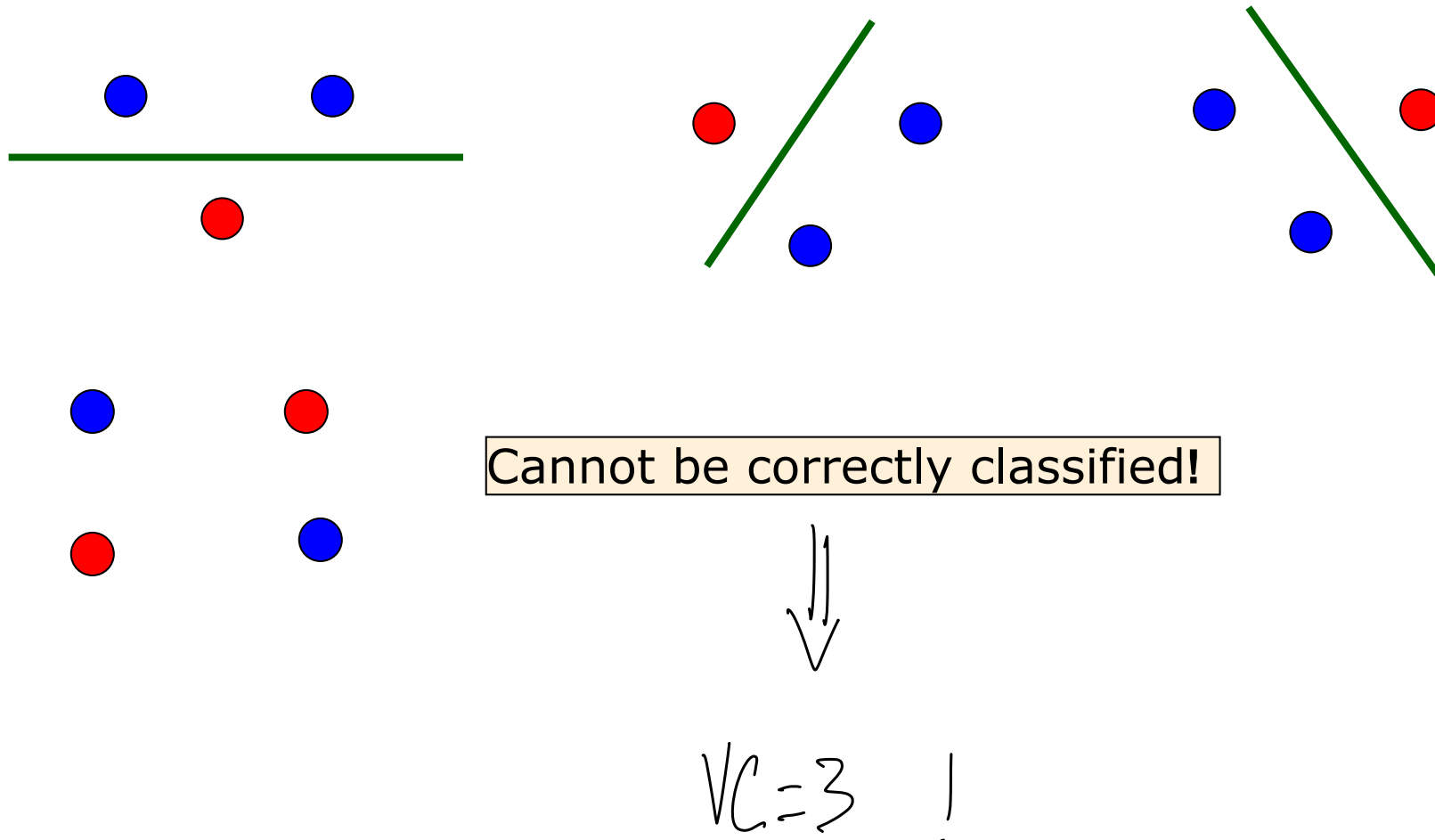


↳ VC 7, 3

OK but what about :  } They are on a line
so Dim = 1.

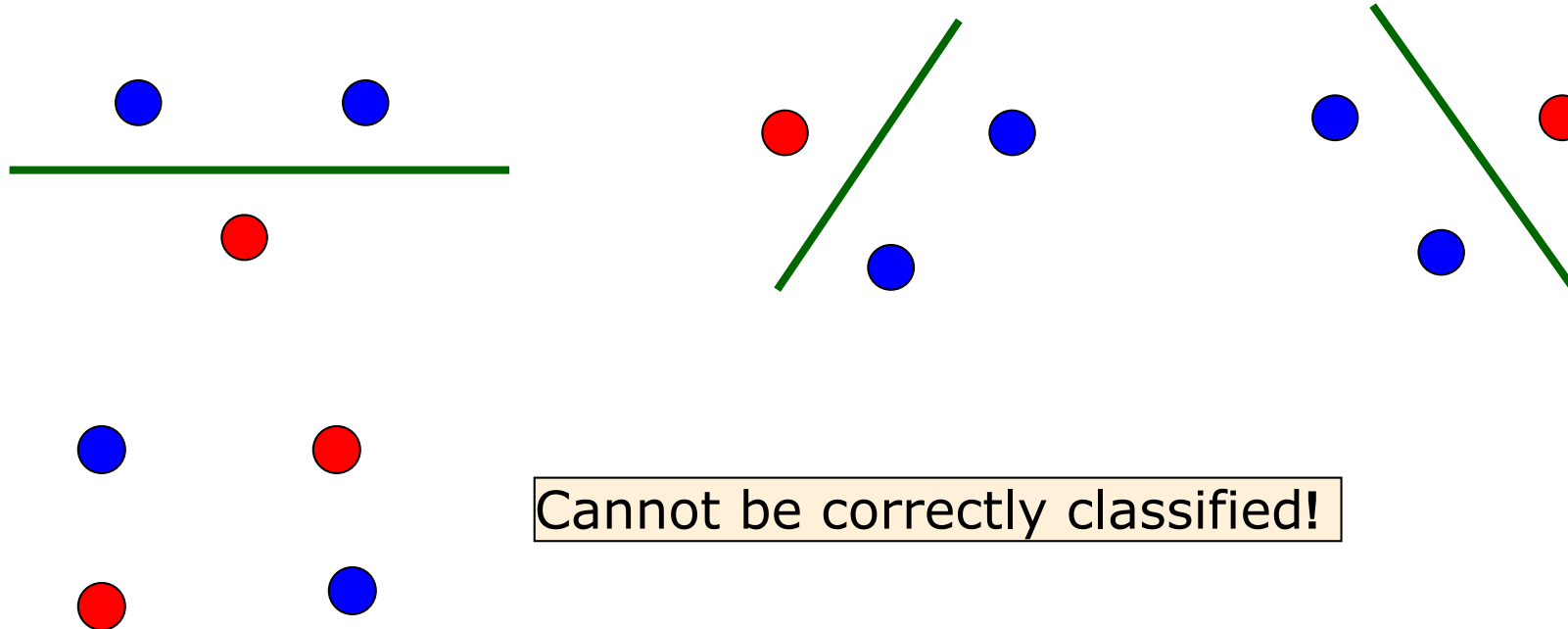
Example: VC dimension of linear classifier

- What about a linear classifier in 2D input space?



Example: VC dimension of linear classifier

- What about a linear classifier in 2D input space?



- A linear classifier in a 2D input space has $VC(H)=3$

Prop.: □ We can prove that a linear classifier in M -D_{in} input space has $VC(H)=M+1$

VC Dimension

Already
seen

- We define a **dichotomy** of a set S of instances as a partition of S into two disjoint subsets, i.e., labeling each instance in S as positive or negative
- We say that a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy
- The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H over instance space X , is the largest finite subset of X shattered by H
- ✓ □ If an arbitrarily large set of X can be shattered by H , $VC(H) = \infty$
- ✓ □ If $|H| < \infty$ then $VC(H) \leq \log_2(|H|)$
 - ▶ If $VC(H) = d$ it means there are in H at least 2^d hypotheses to label d instances
 - ▶ Thus, $|H| \geq 2^d$

Sample Complexity based on VC-Dimension

- How many randomly drawn examples suffice to guarantee that any hypothesis that perfectly fits the training data is probably $(1 - \delta)$ approximately (ε) correct ?

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



New bound: $\left\{ N \geq \frac{1}{\varepsilon} (8VC(H) \log_2(13/\varepsilon) + 4 \log_2(2/\delta)) \right.$

still " $\frac{1}{\varepsilon}$ ".

still " $\frac{1}{\delta}$ ".

Agnostic Learning: VC Bounds

□ With probability at least $(1 - \delta)$ every $h \in H$ satisfies the following inequality:

$$error_{true}(h) \leq error_{\mathcal{D}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$



$$error_{true}(h) \leq error_{\mathcal{D}}(h) + \sqrt{\frac{VC(H)(\ln \frac{2N}{VC(H)} + 1) + \ln \frac{4}{\delta}}{N}}$$