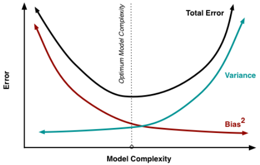
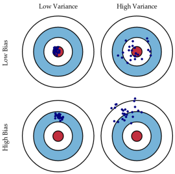
Machine Learning

PAC-Learning and VC-Dimension



Marcello Restelli

April 21, 2020



Outline

PAC-Learning

- Overfitting happens because training error is **bad estimate** of generalization error
  - Can we infer something about generalization error from training error?
- Overfitting happens when the learner **doesn't see "enough" examples**
  - Can we estimate how many examples are **enough**?

A Simple Setting...

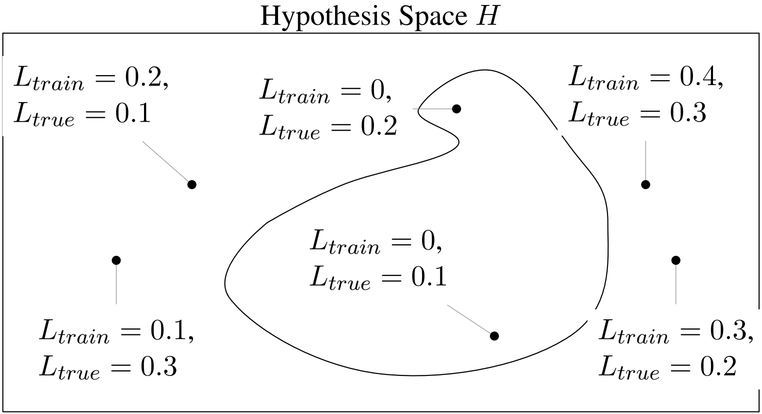
- Given
  - Set of instances $\mathbf{X}$
  - Set of hypotheses H
  - Set of possible target concepts C (Boolean functions)
  - Training instances generated by a fixed, unknown probability distribution $\mathcal{P}$ over $\mathbf{X}$
- Learner **observes** sequence $\mathcal{D}$ of training examples $\langle x, c(x) \rangle$, for some target concept $c \in C$
  - Instances x are drawn from distribution $\mathcal{P}$
  - Teacher provides **deterministic** target value $c(x)$ for each instance
- Learner **must output a hypothesis** h estimating c
  - h is **evaluated** by its performance on **subsequent instances** drawn according to $\mathcal{P}$

$$L_{true} = Pr_{x \in \mathcal{P}}[c(x) \neq h(x)]$$

  - We want to **bound** L_{true} given L_{train}

Version Spaces

- First consider when training error of h is **zero**
- Version Space: $VS_{H,\mathcal{D}}$: subset of hypotheses in H **consistent** with training data $\mathcal{D}$



- Can we **bound the error** in the version space?

How Likely is learner to Pick a Bad Hypothesis?

Theorem

*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 $0 \leq \epsilon \leq 1$, the probability that $VS_{H,\mathcal{D}}$ contains a hypothesis error greater then ϵ is less than $|H|e^{-\epsilon N}$:*

$$Pr(\exists h \in H : L_{train}(h) = 0 \land L_{true}(h) \geq \epsilon) \leq |H|e^{-\epsilon N}$$

How Likely is learner to Pick a Bad Hypothesis?

Proof.

$$
\begin{aligned}
&Pr((L_{train}(h_1) = 0 \wedge L_{true}(h_1) \geq \epsilon) \vee \cdots \vee (L_{train}(h_{|H|}) = 0 \wedge L_{true}(h_{|H|}) \geq \epsilon)) \\
&\quad \leq \sum_{h \in H} Pr(L_{train}(h) = 0 \wedge L_{true}(h) \geq \epsilon) && \text{(Union bound)} \\
&\quad \leq \sum_{h \in H} Pr(L_{train}(h) = 0 | L_{true}(h) \geq \epsilon) && \text{(Bound using Bayes' rule)} \\
&\quad \leq \sum_{h \in H_{bad}} (1 - \epsilon)^N && \text{(Bound on individual } h_i s\text{)} \\
&\quad \leq |H|(1 - \epsilon)^N && (|H_{bad}| \leq |H|) \\
&\quad \leq |H| e^{-\epsilon N} && (1 - \epsilon \leq e^{-\epsilon}, \text{ for } 0 \leq \epsilon \leq 1)
\end{aligned}
$$

□

Using a Probably Approximately Correct (PAC) Bound

- If we want this probability to be at most δ

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

- Pick ϵ and δ, compute N

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

- Pick N and δ, compute ϵ

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

- **Note**: the number of M-ary boolean functions is 2^{2^M}. So the bounds have an **exponential** dependency on the number of features M

Example: Learning Conjunctions

- Suppose H contains **conjunctions of constraints** on up to M Boolean attributes (i.e., M literals)
- $|H| = 3^M$
- How many examples are sufficient to ensure with probability at least $(1 - \delta)$ that every h in $VS_{H,\mathcal{D}}$ satisfies $L_{true}(h) \leq \epsilon$?

$$
\begin{aligned}
N &\geq \frac{1}{\epsilon}\left(\ln 3^M + \ln\left(\frac{1}{\delta}\right)\right) \\
&\geq \frac{1}{\epsilon}\left(M \ln 3 + \ln\left(\frac{1}{\delta}\right)\right)
\end{aligned}
$$

PAC Learning

Consider a class C of possible target concepts defined over a set of instances X of length n, and a Learner L using hypothesis space H.

Definition

C is **PAC-learnable** if there exists an algorithm L such that for every $f \in C$, for any distribution $\mathcal{P}$, for any ϵ such that $0 \leq \epsilon < 1/2$, and δ such that $0 \leq \delta < 1/2$, algorithm L, with probability at least $1 - \delta$, outputs a concept h such that $L_{true}(h) \leq \epsilon$ using a number of samples that is polynomial of $1/\epsilon$ and $1/\delta$

Definition

C is **efficiently PAC-learnable** by L using H iff for all $c \in C$, distributions $\mathcal{P}$ over X, ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $L_{true}(h) \leq \epsilon$, in time that is **polynomial** in $1/\epsilon$, $1/\delta$, M and $size(c)$

Agnostic Learning

- Usually the train error is **not equal to zero**: the Version Space is **empty**!
- What Happens with Inconsistent Hypotheses?
- We need to bound the **gap** between training and true errors

$$L_{true}(h) \leq L_{train}(h) + \epsilon$$

- Using the **Hoeffding bound**: for N i.i.d. coin flips $X_1, \dots, X_N$, where $X_i \in \{0, 1\}$ and $0 < \epsilon < 1$, we define the empirical mean $\overline{X} = \frac{1}{N}(X_1 + \cdots + X_N)$, obtaining the following bound:

$$Pr(\mathbb{E}[\overline{X}] - \overline{X} > \epsilon) \leq e^{-2N\epsilon^2}$$

Theorem

Hypothesis space H finite, dataset $\mathcal{D}$ with N i.i.d. samples, $0 < \epsilon < 1$: for any learned hypothesis h:

$$Pr(\exists h \in H | L_{true}(h) - L_{train}(h) > \epsilon) \leq |H| e^{-2N\epsilon^2}$$

PAC Bound and Bias-Variance Tradeoff

$$L_{true}(h) \leq \underbrace{L_{train}(h)}_{\text{Bias}} + \underbrace{\sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}}_{Variance}$$

- For large $|H|$
  - Low bias (assuming we can find a good h)
  - High variance (because bound is loser)
- For small $|H|$
  - High bias (is there a good h?)
  - Low variance (tighter bound)
- Given δ, ϵ how large should N be?

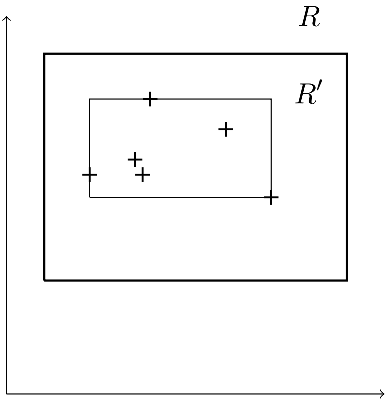
$$N \geq \frac{1}{2\epsilon^2}\left(\ln |H| + \ln \frac{1}{\delta}\right)$$

What about Continuous Hypothesis Spaces?

- **Continuous** hypothesis space
  - $|H| = \infty$
  - Infinite variance???

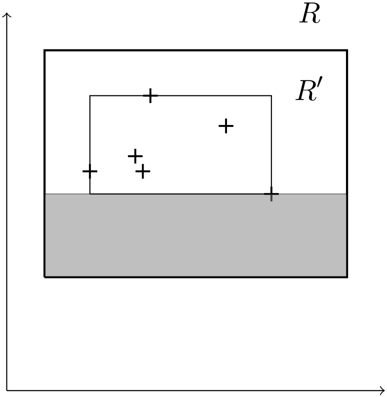
Example: Learning Axis Aligned Rectangles

- We want to learn an **unknown** target axis-aligned rectangle: R
- We have **randomly drawn samples** with a label that indicate whether the point is **contained or not** in R
- Consider the hypothesis corresponding to the **tightest rectangle** R' around positive samples
- The **error region** is the difference between R and R', that can be seen as the **union** of four rectangular regions



Example: Learning Axis Aligned Rectangles

- In **each** of these regions we want an error **less than** $\epsilon/4$
- When N samples are drawn, a bad event is when the probability of all the N samples of being **outside** this region is **at most** $(1-\epsilon/4)^N$
- The same holds for the other three regions, ans so by **union bound** we get $4(1-\epsilon/4)^N$



- We want that the probability of a bad event is less than δ:

$$4(1-\epsilon/4)^N \leq \delta$$

- By exploiting the inequality $(1-x) \leq e^{-x}$, we get:

$$N \geq (4/\epsilon)\ln(4/\delta)$$

What about Continuous Hypothesis Spaces?

- **Continuous** hypothesis space
  - $|H| = \infty$
  - Infinite variance???
- It is important the **number of points that can be classified exactly!**
- **Question**: Can we get a bound error as a function of the number of points that can be completely labeled?

Shattering a Set of Instances

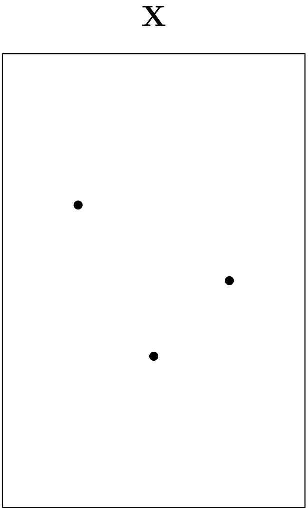
Definition (Dichotomy)

A **dichotomy** of a set S is a partition of S into two disjoint subsets

Definition (Shattering)

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

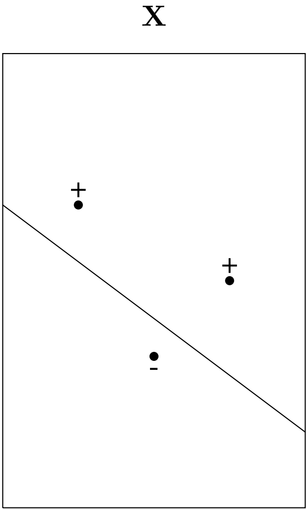
Example: Three Instances Shattered



(a)

(b)

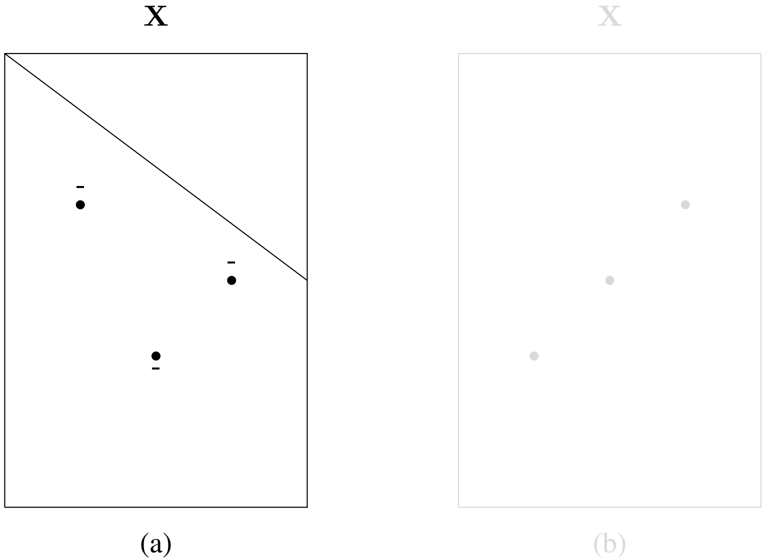
Example: Three Instances Shattered



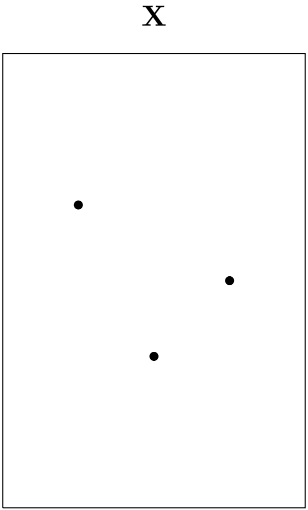
(a)

(b)

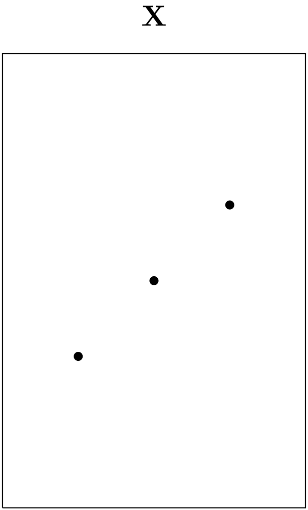
Example: Three Instances Shattered



Example: Three Instances Shattered

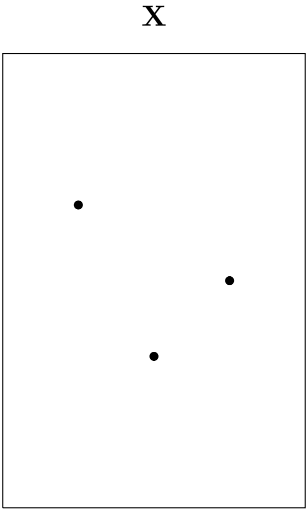


(a)

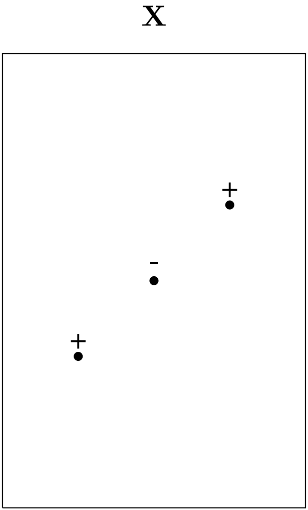


(b)

Example: Three Instances Shattered

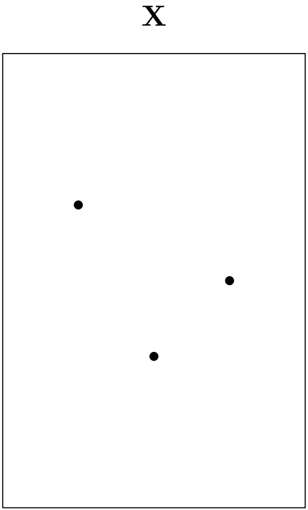


(a)

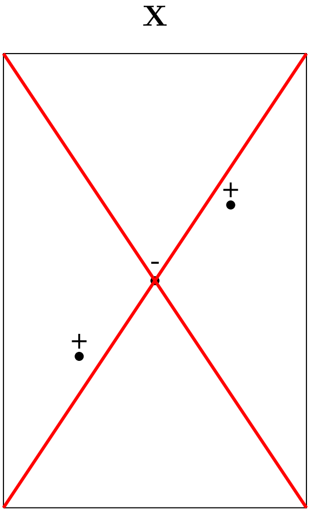


(b)

Example: Three Instances Shattered

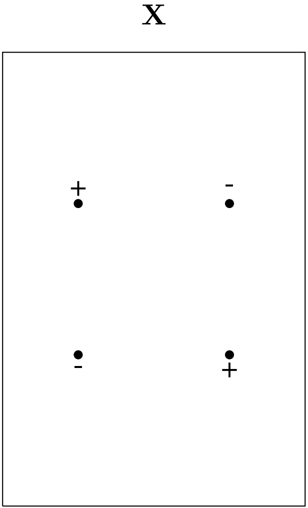


(a)

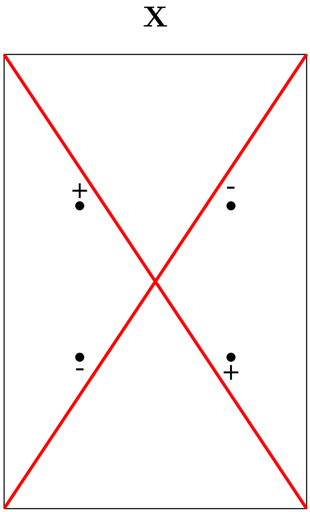


(b)

Example: Four Instances Shattered



Example: Four Instances Shattered



VC Dimension

Definition

The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the **size of the largest finite subset** of X shattered by H. If arbitrarily large finite sets of X can be **shattered** by H, then $VC(H) \equiv \infty$

VC Dimension of Linear Decision Surfaces

- How many points can a linear boundary classify exactly in 1-D?
  - 2
- How many points can a linear boundary classify exactly in 2-D?
  - 3
- How many points can a linear boundary classify exactly in M-D?
  - $M + 1$
- **Rule of thumb**: number of parameters in model often matches max number of points
- But in general it can be completely **different**!
- There are problem where the **number of parameters is infinte** (e.g., SVMs) and the VC dimension is **finite**!
- There can also be a hypothesis space with **1 parameter** and **infinite VC-dimension**!

VC-Dimension Examples

- Examples:
  - Linear classifier
    - $VC(H) = M + 1$, for M features plus constant term
  - Neural networks
    - $VC(H) =$ number of parameters
    - Local minima means NNs will probably not find best parameters
  - 1-Nearest neighbor
    - $VC(H) = \infty$
  - SVM with Gaussian Kernel
    - $VC(H) = \infty$

Sample Complexity from VC Dimension

How many randomly drawn examples suffice to guarantee error of at most ϵ with probability at least $(1-\delta)$?

$$N \geq \frac{1}{\epsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8VC(H)\log_2\left(\frac{13}{\epsilon}\right)\right)$$

PAC Bound using VC dimension

$$L_{true}(h) \leq L_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2N}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{N}}$$

- Same bias/variance tradeoff as always
- Now, just a function of $VC(H)$
- **Structural Risk Minimization**: choose the hypothesis space H to **minimize** the above bound on expected true error!

VC Dimension Properties

Theorem

The VC dimension of a hypothesis space $|H| < \infty$ is bounded from above:

$$VC(H) \leq \log_2(|H|)$$

Proof.

If $VC(H) = d$ then there exist at least 2^d functions in H, since there are at least 2^d possible labelings: $|H| \geq 2^d$ $\square$

Theorem

Concept class C with $VC(C) = \infty$ is not PAC-learnable.