Learning Theory

Lecture 2

Ruth Urner
York University, Toronto, Canada

SUMMER OF MACHINE LEARNING AT SKOLTECH

SIMLES online summer school

August 18, 2020

Course on Machine Learning Theory

Lecture 1, Monday Introduction to learning theory, notation, warm-up exercises

Lecture 2, Tuesday Small historical tour of ML theory, basic concepts and techniques in statistical learning theory: VC-theory and and No Free Lunch

Lecture 3, Wednesday Research topics

Lecture 4, Thursday Fireside chat

A Historical Tour of Learning Theory

Basic question:

How well does the error of predictor h that we observe on the data S represent the error it will make on unseen data/on the underlying process P?

We know:

$$\mathcal{L}_P(h) \ \leq \ \mathcal{L}_S(h) + |\mathcal{L}_P(h) - \mathcal{L}_S(h)|$$

We want:

$$|\mathcal{L}_S(h) - \mathcal{L}_P(h)| \leq \epsilon(|S|)$$

70's: Vapnik and Chervonenkis

If (and only if) a class H of binary predictors $h : X \to \{0, 1\}$ has finite VC-dimension, then we get uniform convergence of empirical to true losses:

$$|\mathcal{L}_S(h) - \mathcal{L}_P(h)| \leq \sqrt{\frac{VC(H)}{|S|}}$$

for all functions $h \in H$ simultaneously.

80's: PAC Learning

Computational complexity is an emerging topic...

PAC Learning

A class H of binary predictors $h : \mathcal{X} \to \{0, 1\}$ is PAC (Probably Approximately Correct) learnable, if there exists an algorithm $\mathcal{A}$ such that for all $\epsilon, \delta > 0$, there exists a sample size n such that, for all data generating distributions P, algorithm $\mathcal{A}$ computes a predictor $\mathcal{A}(S)$ in time polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ and we have

$$\Pr_{S \sim P^n}[\mathcal{L}_P(\mathcal{A}(S)) \leq \inf_{h \in H} \mathcal{L}_P(h) + \epsilon] \geq 1 - \delta$$

Valiant '84: A Theory of the Learnable

80's: PAC Learning

Statistical aspects of PAC learnability satisfied if:

Finite VC-dimension

Bounded compression sizes

$$\mathcal{L}_P(h) \;\leq\; \mathcal{L}_S(h) + \sqrt{\frac{d}{|S|}}$$

$\Rightarrow$ For a good trade-off, the class needs to have bounded capacity.

However, for many classes empirical risk minimization (ERM) is NP-hard...

90's: Boosting and Support Vector Machines

Emergence of efficient, practical learning algorithms..

Freund, Shapire

Weak learning (returning a classifier that is slightly better than random guessing) is as difficult as strong learning.

This was turned into a practical tool ("boosting")

Vapnik, Cortes, Schölkopf, Smola

The concept of large margin classifiers lead to development of Support Vector Machines.

Theory of kernels turned this into a highly successful tool.

2000's: More bounds, better understanding

Various concepts formalize that, as long as a learner will not be allowed enough flexibility to fit to random noise, it will generalize.

- Rademacher complexities
- Data dependent generalization bounds
- PAC Bayes bounds
- Compression bounds
- Algorithmic Stability

2010's: Deep Learning Taking over...

Very large neural networks solve complex learning tasks.

Issues:

- Classes of networks have huge capacity... they "shouldn't" generalize.
- Loss minimization is computationally hard...

Belief:
There still must be some inherent property in either learning method or the predictors that leads to generalization.

2017: Understanding deep learning requires rethinking generalization (Zhang, Bengio, Hardt, Recht, Vinyals)

$$\mathcal{L}_P(h) \;\leq\; \mathcal{L}_S(h) + |\mathcal{L}_P(h) - \mathcal{L}_S(h)|$$

1.) Training deep nets on image data with correct labels

- Training error 0
- Small test error
- Generalizes well

2.) Training deep nets on image data with randomly permuted labels

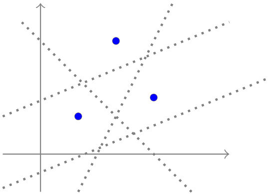
- Training error 0
- Large test error
- Does not generalize!

Both the class of predictors and the training method were identical in the two experiments!

...

The VC-dimension–Shattering

Instance space: $\mathcal{X} \subseteq \mathbb{R}^d$
Label set: $\mathcal{Y} = \{0, 1\}$
Hypothesis class: $H \subseteq \{0, 1\}^{\mathcal{X}}$



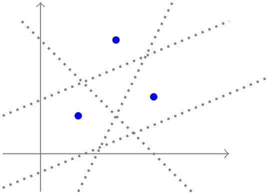
Definition: Shattering
We say that class H shatters a set of points $U \subseteq \mathcal{X}$ if

$$H|_U = \{0, 1\}^U$$

The restriction of the function class H to the set U contains all binary functions over U.

The VC-dimension–Definition

Instance space: $\mathcal{X} \subseteq \mathbb{R}^d$
Label set: $\mathcal{Y} = \{0, 1\}$
Hypothesis class: $H \subseteq \{0, 1\}^{\mathcal{X}}$

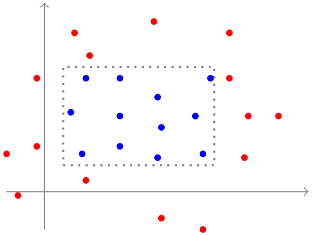


Definition: VC-Dimension
The VC-dimension of H is the maximal size of a set $U \subseteq \mathcal{X}$ that is shattered by H (or ∞ if H can shatter sets of arbitrarily large size).

Quiz 1: What does the above illustration tell us about the VC-dimension of halfspace classifiers in $\mathbb{R}^2$? $\Rightarrow$ It is at least 3!

Quiz 2

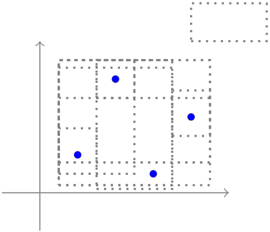
What is the VC-dimension of (axis-aligned) rectangle classifiers in $\mathbb{R}^2$?



H = classifiers that are defined by a fixed range in each feature.

Quiz 2–Solution

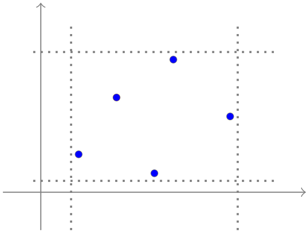
The VC-dimension of rectangle classifiers in $\mathbb{R}^2$ is 4!



Proof step 1: There exist 4 points that are shattered.

Quiz 2–Solution

The VC-dimension of rectangle classifiers in $\mathbb{R}^2$ is 4!



Proof step 2: No set of 5 points is shattered!

The VC-dimension–more examples

Consider domain $\mathcal{X} = \mathbb{R}^d$:

The VC-dimension of linear halfspaces in $\mathbb{R}^d$:

$$d + 1$$

The VC-dimension of all predictors h with $|h^{-1}(1)| < \infty$:

$$\infty$$

The VC-dimension of neural networks with fixed architecture and sign activation:

$$O(|E| \log(|E|))$$

Lemma (Sauer-Shelah-Perles)
Let H be a class of VC-dimension d. Then, for every finite domain subset $U \subseteq \mathcal{X}$, we have

$$|H|_U| \leq \sum_{i=1}^{d} \binom{|U|}{i} \leq \left(\frac{\mathrm{e}|U|}{d}\right)^d \simeq |U|^d$$

$\Rightarrow$ The number of "patterns" that the class H can induce on any domain subset U grows polynomially in $|U|$ once $|U| > d$.

Learnability of bounded VC-classes

Theorem

- Let $\mathcal{X}$ be some domain.
- Let $H \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with $\mathrm{VC}(H) = d < \infty$.
- Let P be a distribution over $\mathcal{X} \times \{0,1\}$.
- Let $\delta > 0$.

With probability at least $1 - \delta$ (over the draw of a data-sample $S \sim P$) we have

$$\sup_{h \in H} |\mathcal{L}_S(h) - \mathcal{L}_P(h)| \;\leq\; c \cdot \sqrt{\frac{\mathrm{VC}(H) + \log(1/\delta)}{n}}$$

(where c is a constant that does not depend on P).

Learnability of VC-classes – proof idea

Recall our proof for finite classes:

$$\Pr\left[\max_{h\in H} |\mathcal{L}_P(h) - \mathcal{L}_S(h)| > \epsilon/2\right]$$
$$\leq \Pr\left[\bigvee_{h\in H} |\mathcal{L}_P(h) - \mathcal{L}_S(h)| > \epsilon/2\right]$$
$$\leq |H| \cdot 2\mathrm{e}^{-2n\epsilon^2} = N \cdot 2\mathrm{e}^{-2n\epsilon^2} := \delta$$

Solving for ϵ gave, w.h.P. $> 1-\delta$:

$$|\mathcal{L}_P(h) - \mathcal{L}_S(h)| \leq \sqrt{\frac{2(\log(2N) + \log(1/\delta))}{n}}$$

for all $h \in H$.

For classes of bounded VC (sketch!):

$$\Pr\left[\sup_{h\in H} |\mathcal{L}_P(h) - \mathcal{L}_S(h)| > \epsilon\right]$$
$$\lesssim \Pr\left[\sup_{h\in H} |\mathcal{L}_{S'}(h) - \mathcal{L}_S(h)| > \epsilon/2\right]$$
$$\lesssim \Pr\left[\bigvee_{h\in H|_{S'\cup S}} |\mathcal{L}_{S'}(h) - \mathcal{L}_S(h)| > \epsilon/2\right]$$
$$\lesssim |S' \cup S|^d \cdot 2\mathrm{e}^{-2n\epsilon^2} := \delta$$

We get w.h.P. $> 1-\delta$:

$$|\mathcal{L}_P(h) - \mathcal{L}_S(h)| \leq c \cdot \sqrt{\frac{d + \log(1/\delta))}{n}}$$

for all $h \in H$.

No Free Lunch Theorem

Theorem
Let $\mathcal{X}$ be an infinite domain. For every learning algorithm $\mathcal{A}$, and every sample size $n \in \mathbb{N}$, there exists a distribution P such that

$$\mathop{\mathbb{E}}_{S \sim P^n} [\mathcal{L}_P(\mathcal{A}(S))] \geq \frac{1}{4}$$

The distribution P is rather benign:

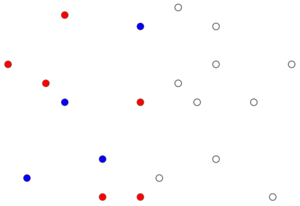
- There is a predictor h with $\mathcal{L}_P(h) = 0$
- The support of P has size $2n$.

No Free Lunch – Proof idea

$$\forall \mathcal{A} \forall n \exists P : \underset{S \sim P^n}{\mathbb{E}} [\mathcal{L}_P(\mathcal{A}(S))] \geq \frac{1}{4}$$

As **possible distributions** P, consider:

- support of $2n$ points
- all possible labelings over these points
- $\mathcal{A}$ gets to "see" labels on half the points
- for each point not in S, $\mathcal{A}$ mis-predicts with probability $1/2$



Implications of the No Free Lunch Theorem

The same proof technique can be used to prove lower bounds for learning VC-classes.

The sample complexity of learning a class H with $\mathrm{VC}(H) = d$, even in the realizable case, is lower bounded by $\Omega(\frac{d}{\epsilon})$.

There is no universally successful learning algorithm.
(Every learner has an "inductive bias", and needs this bias to be successful on some tasks).

Thank You!