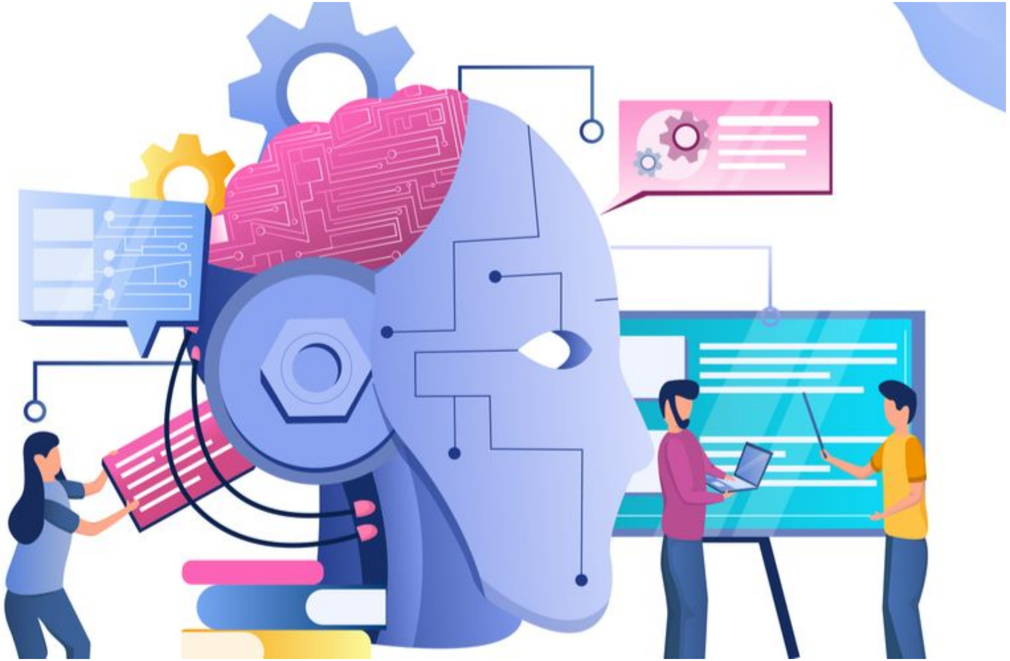


Sauer-Shelah Lemma

Zahra Bayramli
20200812

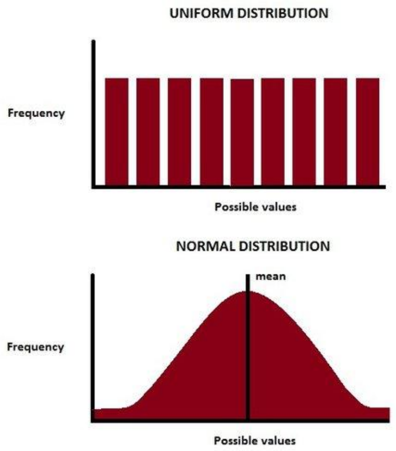
Table of Contents

Definitions

- **X** is the population or universe. This is the set of objects that we wish to classify
- **μ** is a distribution over **X. μ(x)** is the probability to sample **x**
- Training data: n instances of x sampled from the population according to **μ**

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

- **c*** is an optimal classifier: correctly classifies every point of the population **X**
- **C** is the family of possible classifiers (or a set of functions from X to {0,1})



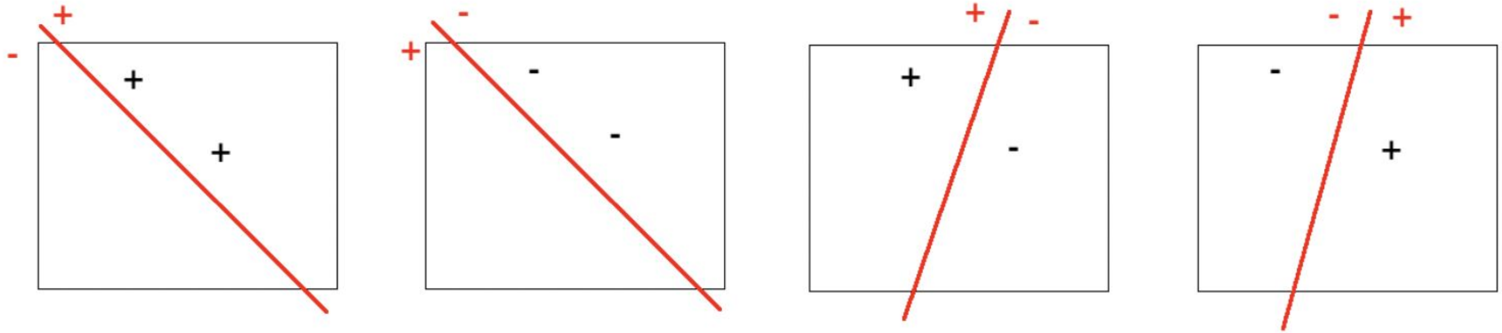
Shattering

A set **U**⊆**X** is ***shattered*** by family **C** if the classifiers in **C** are able to give all the $2^{|U|}$ possible classifications of these points

|**U**| = cardinality of the set **U**

Shattering (Example)

Suppose our population **X = R**^{2} and **C** is a family of linear classifiers



U is the set of two points
U is shattered by **C** $\Rightarrow$ **C** shatters any set of two points in **R**^{2}

Shattering (Example)

If **U** is a set of 4 points, **C** is not able to shatter it ⇒ 4 points: impossible to shatter using linear classifiers

Shattering

Going a little further, we can check the following results:

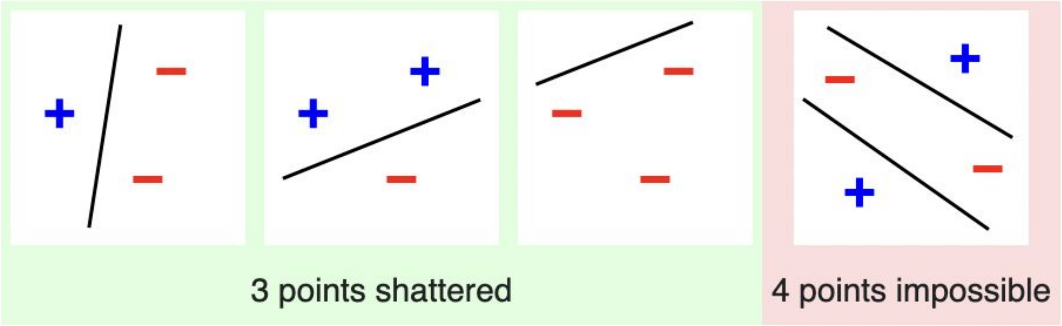
- In $\mathbb{R}^2$, $\exists$ a set U of 3 points that is shattered by $\mathcal{C}$, a family of linear classifiers.
- In $\mathbb{R}^3$, $\exists$ a set U of 4 points that is shattered by $\mathcal{C}$, a family of linear classifiers.
- In $\mathbb{R}^2$, $\nexists$ a set U of 4 points that is shattered by $\mathcal{C}$, a family of linear classifiers.
- In $\mathbb{R}^3$, $\nexists$ a set U of 5 points that is shattered by $\mathcal{C}$, a family of linear classifiers.

VC dimensions

Let **X** be the population and **C** be the family of classifiers. The **Vapnik-Chervonenkis (VC)** dimension of **C** is **d** if:

- There exist a set **U**, with |**U**| = **d**, shattered by **C**
- There does not exist a set **U**, with |**U**| = **d+1**, shattered by **C**

In our previous example, with **X = R**^{2} and **C** the set of linear classifiers, the VC dimension is 3.



VC dimensions

Finite ***C***

When the family of classifiers is finite, can we give an upper bound for the VC dimension?

Yes!

Suppose a set of m points:

$$|\mathcal{C}| \geq 2^m$$

$$m \leq \log_2(|\mathcal{C}|)$$

VC dimension is the *maximum number* of points that can be arranged so that a classifier from ***C*** can shatter them ⇒

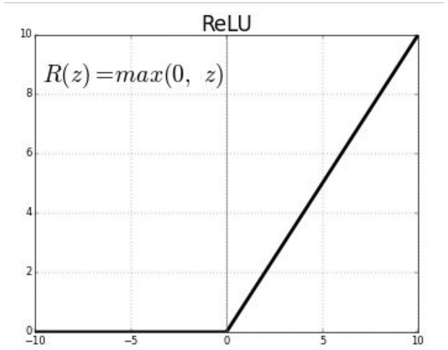
$$\text{VC dimension} \leq \log_2(|\mathcal{C}|)$$

VC dimensions

C *= every possible classifier,* ***X*** *= [0, 1]*

Since every classifier is in C, this family shatters sets of any size.
In this case,

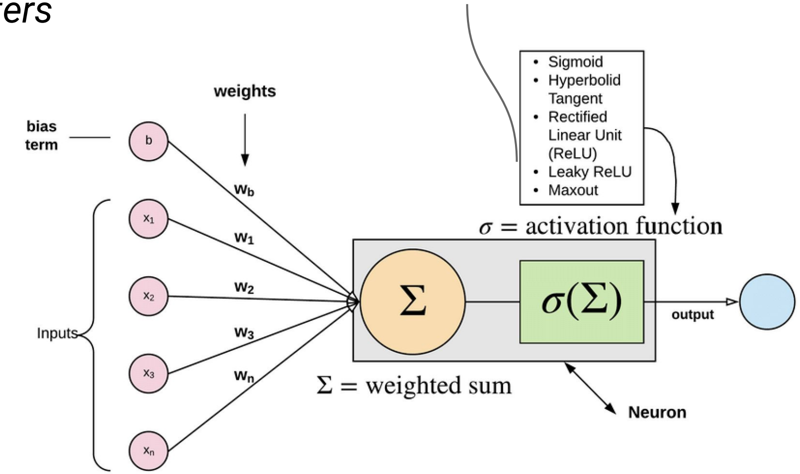
$$VC dimension = \infty.$$



Neural Networks with activation function ***ReLU*** *and* ***k*** *parameters*

For this family of classifiers,

$$VC dimension \approx k.$$



What does Sauer-Shelah lemma state?

Question: How does the VC dimension capture the flexibility of our classifiers?

A few more notations…

- For any, $S \subseteq X$ we can denote it as $S = \{x_1, \cdots, x_m\}$
- $C(S)$ Set of all labelings on *S* that are induced by *C*,

  $C(S) = \{(c(x_1), \cdots, c(x_m)); c \in \mathcal{C}\}$ and $C(S) \subseteq \{0,1\}^m$

- For any natural number m, the maximum number of ways to split *m* points using classifiers in *C*:

  $C[m] = \max\{|C(S)|; |S| = m, S \subseteq X\}$

Sauer-Shelah Lemma and Proof

Lemma 1 *If $d = VCdim(C)$, then for all m, $C[m] \leq \Phi_d(m)$, where $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.*

Proof

The proof proceeds by induction on both d and m.

two base cases: when $m = 0$ and d is arbitrary, and when $d = 0$ and m is arbitrary.

- $m = 0$, there can only be one subset $\Rightarrow$ $C[0] \leq 1 = \Phi_d(0)$
- $d = VCdim(C) = 0$, no set of points can be shattered $\Rightarrow$ $C[m] = 1 \leq \Phi_0(m)$

We are done with the base case!

Proof

We assume for induction that for all m', d' such that $m' \leq m$ and $d' \leq d$ and at least one of these inequalities is strict, we have $C[m'] \leq \Phi_{d'}(m')$.

Now suppose we have a set $S = \{x_1, x_2, \ldots, x_m\}$ of cardinality m. Let H be a class of functions defined only over $\{x_1, x_2, \ldots, x_m\}$ such that $C(S) = H(S) = H$. Since any $\tilde{S} \subseteq S$ that is shattered by H is also shattered by C, we have $VCdim(H) \leq VCdim(C)$.

we add a representative function from H to H_1; we let $H_2 = H \setminus H_1$.

for each $h \in H_2$, $\exists \tilde{h} \in H_1$ such that $h(x_i) = \tilde{h}(x_i)$ for $i \in \{1, \ldots, m-1\}$ and $h(x_m) \neq \tilde{h}(x_m)$.

For convenience, we let $h(x_m) = 1$ and $\tilde{h}(x_m) = 0$

Proof

By construction we have

$$|C(S)| = |H(S)| = |H_1(S)| + |H_2(S)|.$$

Since $H_1 \subseteq H$ we have $VCdim(H_1) \leq VCdim(H) \leq d$. Moreover, we can show

$$|H_1(S)| = |H_1(S \setminus \{x_m\})|.$$

By induction we have:

$$|H_1(S)| \leq \Phi_d(m-1).$$

We omit the details, but we can show that,

$$VCdim(H_2) \leq VCdim(H) - 1 \leq d - 1.$$

$$|H_2(S)| = |H_2(\{x_1, x_2, \ldots, x_{m-1}\})|,$$

By induction we obtain: $|H_2(S)| \leq \Phi_{d-1}(m-1).$

Proof

Now we have two important results:

$$|H_1(S)| \leq \Phi_d(m-1) \quad \text{and} \quad |H_2(S)| \leq \Phi_{d-1}(m-1)$$

Let us combine them:

$$|C(S)| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1)$$

Since,

RHS = $$\sum_{i=0}^{d}\binom{m-1}{i} + \sum_{i=0}^{d-1}\binom{m-1}{i} = \binom{m}{0} + \sum_{i=1}^{d}\binom{m-1}{i} + \sum_{i=1}^{d}\binom{m-1}{i-1} = \sum_{i=0}^{d}\binom{m}{i}$$

So we get $|C(S)| \leq \Phi_d(m)$ **We are done!**

Polynomial Bound

So far, we proved:

$$C[m] \leq \Phi_d(m), \; where \; \Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$$

It's cool! But can we introduce a polynomial bound to make it even cooler?

Yes, we can!

Polynomial Bound

Lemma 2 *For $m \geq d$ we have:*

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d.$$

Proof

Since $m \geq d$, we have $0 \leq \frac{d}{m} \leq 1$. Therefore:

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = \left(1 + \frac{d}{m}\right)^m \leq e^d.$$

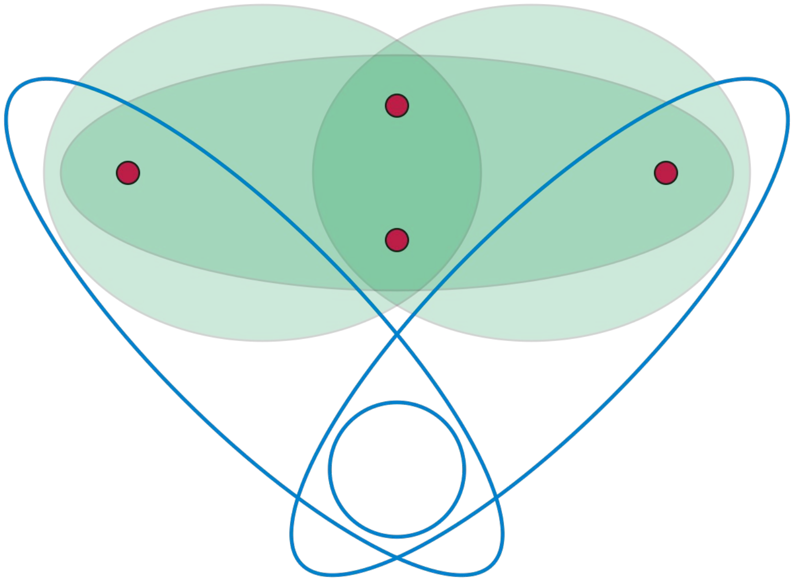
First inequality due to $d/m \leqslant 1$
Second inequality due to $d \leqslant m$
Next equality is binomial theorem
Last inequality is $1 + x \leqslant e^x$ for all real x

!

Sauer-Shelah lemma shows that when m becomes larger than d, the function C[m] increases polynomially rather than exponentially with sample size m.

Pajor’s formulation of Sauer-Shelah Lemma

For every finite family of sets (green) there is another family of equally many sets (blue outlines) such that each set in the second family is shattered by the first family



Applications

Machine Learning

- Probably approximately correct learning (PAC)

Graph Theory

- Strong orientations of a graph

Computational Geometry

- Range searching
- Derandomization
- Approximation algorithms

References

Understanding Machine Learning: From Theory to Algorithms, 2014, by y Shai Shalev-Shwartz and Shai Ben-David

https://www.cse.cuhk.edu.hk/~siuon/csci4230/notes/notes13.pdf

http://www-di.inf.puc-rio.br/~mmolinaro/algInc18-2/lecture11.pdf

https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_dimension

https://en.wikipedia.org/wiki/Sauer%E2%80%93Shelah_lemma

Thank you!