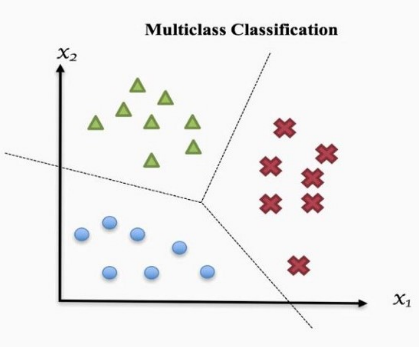
3.5 Extension of the VC dimension

- The definition of the VC dimension is exclusive to binary classification, as well the statement of the fundamental theorem (Theorem 3.33)

- However, the concept and statements can be extended to multiclass classification and regression



Source: medium.com

- There we want to learn hypotheses $h\colon \mathcal{X} \to \{1, \ldots, L\}$ or $h\colon \mathcal{X} \to \mathbb{R}$, respectively

- We briefly discuss those extensions and start with the multiclass case, i.e., let now $\mathcal{Y} = \{1, \ldots, L\}$ mit $L \geq 2$ and ℓ denote the 0-1 loss.

Proof of Th 3.36:

Since $\mathcal{H} = \mathcal{H}_{V,E,sgn}$ satisfies $\mathcal{H} = \mathcal{H}_L \circ \cdots \circ \mathcal{H}_1$,

$\mathcal{H}_k = \mathcal{L}_{d_{k,1}} \times \cdots \times \mathcal{L}_{d_{k,n_k}}$? $\}$ $h(x) = \operatorname{Sgn}(W_k x + b)$

$W_k \in \mathbb{R}^{n_k \times n_{k-1}}, b \in \mathbb{R}^{n_k}$

neuron $v_{k,1} \in V_k$ (which specify a hypothesis class)

Thus we get by proposition 3.37:

$$\Rightarrow \tau_{\mathcal{H}}(m) \leq \prod_{k=1}^{L} \tau_{\mathcal{H}_k}(m)$$

$$\leq \prod_{k=1}^{L} \prod_{j=1}^{n_k} \tau(\mathcal{L}_{d_{k,j}})(m)$$

By Shaler-Sauer lemma / Th 3.31:

we have for $\mathcal{L}_d$, $d \in \mathbb{N}$, $\tau_{\mathcal{L}_d}(m) \le \left(\frac{e}{d+1}\right)^{d+1} m^{d+1}$

Since $VCD(\mathcal{L}_d) = d+1$

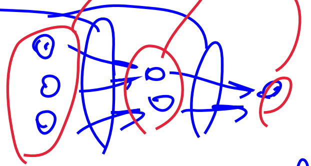
$\le (em)^{d+1}$ (See Th 3.35) in exam

Thus we get that $\tau_{H}(m) \le \prod_{k=1}^{L} \prod_{j=1}^{n_k} (em)^{d_{k,j}+1}$.

$= (em)^{\sum_{k=1}^{L} \sum_{j=1}^{n_k} (d_{k,j}+1)}$

Since, $\underbrace{\sum_{k=1}^{L}\sum_{j=1}^{n_k} d_{k,j}}_{=|E|}$, $\underbrace{\sum_{k=1}^{L}\sum_{j=1}^{n_k} 1}_{|V_1|+\dots|V_L|=n}$ => it requires Vertices & Edges

b cuz of this:

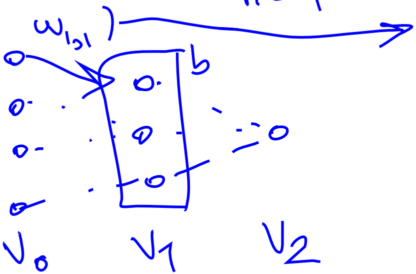


we would get $\overline{\Pi}_H(m) \le (em)^{P_{V,E}}$, based on Th 3.37

$$VCD(\underbrace{\mathcal{H}_{V,E,sgn}}_{FFN}) \le 16 \cdot P_{V,E} \ln(P_{V,E}) \quad \#$$

for number of parameters, based on the FNN's layers of the FNN, we would not consider the input layer but only hidden layer. only in this case we would have to sum up all of V, E together.

$$\sum_{n=1}^{L} sgn(V, E):$$ Z.B.:



weight of V_1 to layer 1, the biases would be our hidden layers & only our weights would be counted

Shattering for multiple classes $|\mathcal{Y}| \geq 2$

Definition 3.38:

A hypothesis class $\mathcal{H} \subseteq \{1,\ldots,L\}^{\mathcal{X}}$, $L \geq 2$, shatters a set $M \subseteq \mathcal{X}$, if

- there exists two functions $f, g\colon M \to \mathcal{Y}$ with $f(x) \neq g(x)$ for all $x \in M$
- and for each subset $B \subseteq M$ there exists an $h \in \mathcal{H}_M$ such that

$$h(x) = f(x) \quad \forall x \in B, \qquad h(x) = g(x) \quad \forall x \in M \setminus B.$$

Remarks:

- Definition 3.38 again examines whether $\mathcal{H}$ can reproduce any arbitrary labelling of elements $x \in M$ based on the labels $f(x) \neq g(x) \in \mathcal{Y}$.
- For $L = 2$ we get the binary version of shattering: For $f(x) \equiv 0$ and $g(x) \equiv 1$ and any $B \subseteq M$ there need to exist an $h \in \mathcal{H}_M$ such that

$$h(x) = 0 \quad \forall x \in B, \qquad h(x) = 1 \quad \forall x \in M \setminus B.$$

  Hence: $\mathcal{H}$ shatters M if and only if $\mathcal{H}_M = \{0,1\}^M$.

Example 3.39:

Let $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\mathcal{Y} = \{1, \dots, 4\}$. We consider the class

$$\mathcal{H} = \{h \colon \mathcal{X} \to \mathcal{Y} \mid h(x_1) \in \{1,2,4\},\ h(x_2) \in \{1,2,3\},\ h(x_3) \in \{3,4\}\}.$$

There holds $|\mathcal{H}| = 18 = 3 \cdot 3 \cdot 2$. Further, $\mathcal{X}$ is shattered by $\mathcal{H}$ – e.g., with $f,\ g \colon \mathcal{X} \to \mathcal{Y}$ as follows:

$$[f(x_1)\ f(x_2)\ f(x_3)] = [1\ 2\ 3], \qquad [g(x_1)\ g(x_2)\ g(x_3)] = [2\ 3\ 4].$$

Definition 3.40:

The Natarajan dimension $\mathrm{ND}(\mathcal{H})$ of a class $\mathcal{H} \subseteq \{1, \dots, L\}^{\mathcal{X}}$, $L \geq 2$, is given by

$$\mathrm{ND}(\mathcal{H}) := \underbrace{\sup}_{\text{the lagest set of shatterings}}\{|M| \colon M \subseteq \mathcal{X} \text{ is shattered by } \mathcal{H}\}.$$

labels

$\Rightarrow$ For $L = 2$ we have $\mathrm{ND}(\mathcal{H}) = \mathrm{VCD}(\mathcal{H})$.

The multiclass fundamental theorem

The infinte ND would not satisfies UC cuz it ain't PAC-learnable

Theorem 3.41:

For a class $\mathcal{H} \subseteq \{1,\dots,L\}^{\mathcal{X}}$, $L \geq 2$, the following statements are equivalent w.r.t. the 0-1 loss:

1. $\mathcal{H}$ satisfies uniform convergence (UC). → there are no data distribution in True hypothesis in hypothesis class. (but unknown)
2. $\mathcal{H}$ is (agnostic) PAC-learnable by $A = \mathrm{ERM}_{\mathcal{H}}$.
3. $\mathcal{H}$ is (agnostic) PAC-learnable.
4. $\mathcal{H}$ has finite Natarajan dimension. also satisfies UC.

In particular, there are universal constants $c, C < \infty$, such that:

$$c\frac{\mathrm{ND}(\mathcal{H}) + \ln(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon,\delta) \leq C\frac{\mathrm{ND}(\mathcal{H})\ln(L) + \ln(1/\delta)}{\epsilon^2}$$

$\tau_{\mathcal{H}}(m) = \sup |H_M|$

$m = |M|$

$M \subset \mathcal{X}$

sample complexity according to PAC-learnability

Scalar regression

- We now consider the label space $\mathcal{Y} = \mathbb{R}$ as well as the quadratic loss $\ell(h,(x,y)) = |h(x) - y|^2$.

- For the fundamental theorem we restrict ourselves to $\mathcal{Y} = [0,1]$ but the results can be generalized to $\mathcal{Y} = [-a,a]$, $0 < a < \infty$, and $\mathcal{Y} = \mathbb{R}$ with suitable modifications.

- In the case of scalar regression, the expected and empirical risk are, respectively,

$$\mathcal{R}_\mu(h) = \mathbb{E}_\mu\left[|Y - h(X)|^2\right] \qquad \mathcal{R}_s(h) = \frac{1}{m}\sum_{i=1}^{m} |y_i - h(x_i)|^2\,,$$

  with sample $s = ((x_i, y_i) \colon i = 1, \ldots, m) \in (\mathcal{X} \times [0,1])^m$.

- $\mathcal{R}_s$ reminds to linear regression and corresponds to least squares curve fitting.

The pseudo dimension

Definition 3.42:

A finite set $M = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ is pseudo shattered by the class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ if there are real numbers $r_1, \ldots, r_m$ such that for every binary m-bit pattern $\mathbf{b} \in \{0,1\}^m$ there exists a $h_{\mathbf{b}} \in \mathcal{H}$ with

$$b_i = \begin{cases} 0, & h_{\mathbf{b}}(x_i) < r_i, \\ 1, & h_{\mathbf{b}}(x_i) \geq r_i, \end{cases} \qquad i = 1, \ldots, m.$$

Moreover, the pseudo dimension of $\mathcal{H}$ is given by

$$\mathrm{PD}(\mathcal{H}) := \sup\{|M| \colon M \subseteq \mathcal{X} \text{ is pseudo shattered by } \mathcal{H}\}.$$

- For binary hypothesis $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}} \subset \mathbb{R}^{\mathcal{X}}$ shattering coiincides with pseudo shattering (where $\mathbf{r} = \mathbf{1}$) and $\mathrm{VCD}(\mathcal{H}) = \mathrm{PD}(\mathcal{H})$. which satisfie ReLU function.
- The pseudo dimension generalizes the VC dimension to real valued hypotheses.

Illustrating pseudo shattering

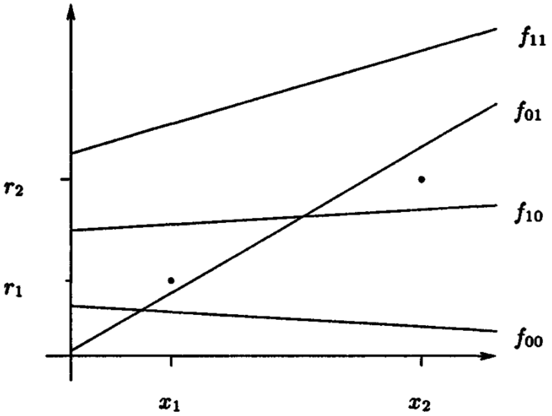


Fig. 11.1. The set $\{x_1, x_2\} \subset \mathbb{R}$ is shattered by the class F of affine functions on $\mathbb{R}$, $F = \{x \mapsto ax + b : a, b \in \mathbb{R}\}$. The points r_1, r_2 witness the shattering.

Source: "Neural Network Learning: Theoretical Foundations" (2009)

Example 3.43: Logistic regression

Let $\mathcal{X} = \mathbb{R}$ and consider

$$\mathcal{H} = \mathcal{L}_{1,\mathrm{sig}} = \{h_{w,b}(x) = \mathrm{sig}\,(wx + b) : \; w, b \in \mathbb{R}\} \; \subseteq \; [0,1]^{\mathcal{X}}.$$

- For $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 < x_2$, and $r_1 = r_2 = \mathrm{sig}^{-1}(0.5) = 0$ we can find corresponding $w, b \in \mathbb{R}$ for each of the following cases:

  green $\quad wx_1 + b < 0, wx_2 + b < 0$

  yellow $\quad wx_1 + b < 0, wx_2 + b > 0$

  red $\quad wx_1 + b > 0, wx_2 + b < 0$

  purple $\quad wx_1 + b > 0, wx_2 + b > 0$

  Thus, M is pseudo-shattered by $\mathcal{H}$
- However, any set $M = \{x_1, x_2, x_3\} \subset \mathbb{R}$, $x_1 < x_2 < x_3$, can not be shattered by $\mathcal{H}$. This is, in fact, a consequence of Theorem 3.35 and the monotonicity of $\mathrm{sig}\colon \mathbb{R} \to [0,1]$.

Thus, we have $\mathrm{PD}(\mathcal{L}_{1,\mathrm{sig}}) = 2$ and, in general, $\mathrm{PD}(\mathcal{L}_{d,\mathrm{sig}}) = \underbrace{d}_{\text{inputs}} + 1$.

Using covering numbers for hypotheses classes $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ and their relation to the pseudo dimension $\mathrm{PD}(\mathcal{H})$ one can show an analogue of Theorem 3.32:

Theorem 3.44:

For hypotheses classes $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and quadratic loss ℓ we have for arbitrary $\epsilon \in (0,1)$ and $m \in \mathbb{N}$ and any distribution μ on $\mathcal{X} \times [0,1]$ that

$$\mathbb{P}_{\mu^m}\left(\exists h \in \mathcal{H}\colon |\mathcal{R}_\mu(h) - \mathcal{R}_S(h)| > \epsilon\right) \;\leq\; 4\left(\frac{32m}{\epsilon}\right)^{\mathrm{PD}(\mathcal{H})} \exp\left(-\frac{\epsilon^2 m}{32}\right).$$

Thus, if $\mathrm{PD}(\mathcal{H}) < \infty$, then $\mathcal{H}$ is (agnostic) PAC-learnable by $A = \mathrm{ERM}_{\mathcal{H}}$ and

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{128}{\epsilon^2}\left(2\mathrm{PD}(\mathcal{H}) \ln\left(\frac{34}{\epsilon}\right) + \ln\left(\frac{16}{\delta}\right)\right)\right\rceil.$$

Proof: See Chapter 17 and 19 in "Neural Network Learning: Theoretical Foundations".

More on covering numbers

- The growth function

$$\tau_{\mathcal{H}}(m) := \sup_{M \subset \mathcal{X}, |M| = m} |\mathcal{H}_M|, \qquad m \in \mathbb{N},$$

  is not useful anymore for real-valued $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, because we usually have $|\mathcal{H}_M| = \infty$

- It is replaced by the covering number of $\mathcal{H}_M \subseteq \mathbb{R}^m$, $|M| = m$,

$$\mathcal{N}_{\mathcal{H}}(m, \epsilon) := \sup_{M \subset \mathcal{X}, |M| = m} \mathcal{N}_{\infty}(\mathcal{H}_M, \epsilon)$$

  where $\mathcal{N}_{\infty}(C, \epsilon)$ is the smallest number of balls of radius ϵ w.r.t. the maximum distance which cover $C \subseteq \mathbb{R}^m$ completely, i.e.,

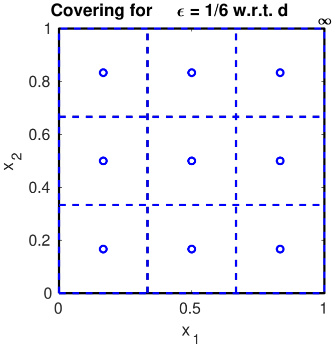
$$C = \bigcup_{j=1}^{n} B_{\epsilon}(\mathbf{b}_j), \qquad B_{\epsilon}(\mathbf{b}_j) = \{\mathbf{b} \in \mathbb{R}^m : \|\mathbf{b} - \mathbf{b}_j\|_{\infty} = \max_k |b_k - b_{k,i}| \leq \epsilon\}$$

- **Example:** Let $\mathcal{X} = \mathbb{R}$, then obviously

$$\mathcal{N}_\infty([0,1],\epsilon) = \left\lceil \frac{1}{2\epsilon} \right\rceil,$$

and for $\mathcal{X} = \mathbb{R}^m$ we have

$$\mathcal{N}_\infty([0,1]^m,\epsilon) = \left\lceil \frac{1}{2\epsilon} \right\rceil^m.$$



- Moreover, for binary $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}} \subseteq [0,1]^{\mathcal{X}}$ we have for any $\epsilon \in (0,1)$

$$\mathcal{N}_{\mathcal{H}}(m,\epsilon) = \tau_{\mathcal{H}}(m), \qquad \forall m \in \mathbb{N}.$$

Thus, the covering number $\mathcal{N}_{\mathcal{H}}$ **generalizes** the growth function $\tau_{\mathcal{H}}$.

Uniform convergence and covering numbers

Analogously, to the uniform convergene theorem one can show

Theorem 3.45:

For hypotheses classes $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and quadratic loss ℓ we have for arbitrary $\epsilon \in (0,1)$ and $m \in \mathbb{N}$ and any distribution μ on $\mathcal{X} \times [0,1]$ that

$$\mathbb{P}_{\mu^m}\left(\exists h \in \mathcal{H} \colon |\mathcal{R}_\mu(h) - \mathcal{R}_S(h)| > \epsilon\right) \;\leq\; 4\,\mathcal{N}_{\mathcal{H}}(2m, \epsilon/16)\exp\left(-\frac{\epsilon^2 m}{32}\right).$$

And analogously to the Shelah–Sauer Lemma, we have

$$\mathcal{N}_{\mathcal{H}}(m,\epsilon) \;\leq\; \sum_{i=0}^{\mathrm{PD}(\mathcal{H})} \binom{m}{i}\left(\frac{1}{\epsilon}\right)^i \in \mathcal{O}\left(\left(\frac{m}{\epsilon}\right)^{\mathrm{PD}(\mathcal{H})}\right) \qquad \forall m \in \mathbb{N}, \epsilon \in (0,1),$$

which yields Theorem 3.45.

Remarks

- The statement of Theorem 3.45 holds even for approximate ERM rules $A\colon \bigcup_{m\in\mathbb{N}} \mathcal{D}^m \to \mathcal{H}$ satisfying

$$\mathcal{R}_s(A(s)) \leq \inf_{h\in\mathcal{H}} \mathcal{R}_s(h) + \frac{16}{\sqrt{m}} \qquad \forall s \in \mathcal{D}^m \ \forall m \in \mathbb{N}$$

- However, there are PAC-learnable classes $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with $\mathrm{PD}(\mathcal{H}) = \infty$.

- Thus, the pseudo dimension is not a characterizing property for learnability in scalar regression.

- A “refined” dimension which indeed characterizes PAC learnability is the so called *fat shattering dimension* of $\mathcal{H}$ which relies on a particular definition of *fat shattered sets*.

- The *fat shattering dimension* is always smaller or equal to the pseudo dimension and yields also lower bounds on $m_{\mathcal{H}}(\epsilon, \delta)$. For further details, see "Neural Network Learning: Theoretical Foundations" (2009).

Recall comparing hypotheses classes

Now we have the knowledge to compare the three methods we have learned in Chapter 2 by completing the table below using

$$1 \text{ (best)}, \quad 2 \text{ (medium)}, \quad 3 \text{ (worst)}$$

for the performance regarding the corresponding important errors:

VCD is important in terms of learning in here

VCD of perceptron and logistic is d+1 & NN is $P_{V,E} \ln p(V,E)$

Method	ε_{app}	ε_{est}	ε_{opt}
Perceptron	3	1	2
Logistic regression	2	1	1
Neural networks	1	3	3

for heavy computational we use, besides in cases the ε_{app} are most important so NN is best

Take home messages

- What does PAC mean (in words and formulas)?
- What is the reazibility assumption?
- What are the implications of the no-free-lunch theorem?
- Which one is the stronger statement: uniform convergence or PAC learnability? And why?
- What is the growth function and why is it growing behaviour important?
- What is the meaning of shattering (mathematically and illustrative)?
- What is the VC dimension and how can one calculate it?
- What is the statement of the fundamental theorem?
- What about PAC learnability for multiclasses or regression?