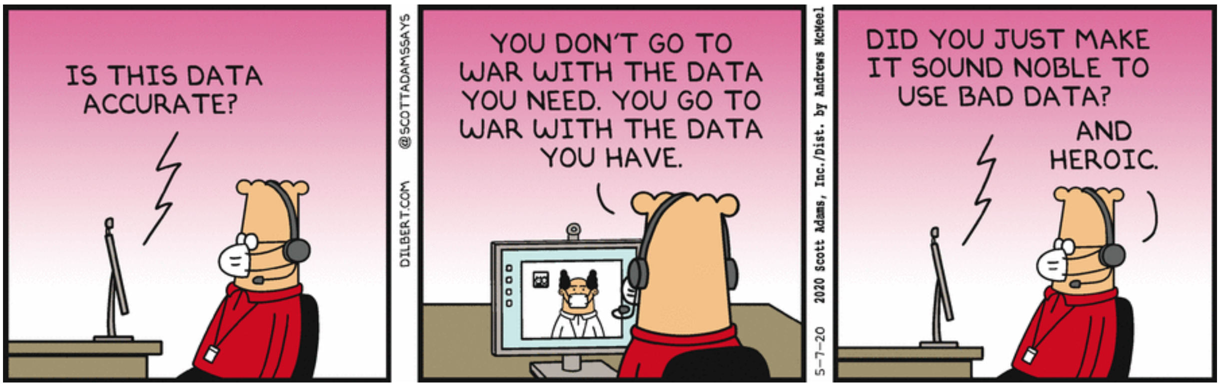
Notes: Statistical Learning Theory II



1 Pre-lecture Work

Problem 1. (optional) Chapter 4 of Shalev-Shwartz and Ben-David covers a concept called *uniform convergence*. This is a mathematical tool that was historically used before the discovery of the VC-dimension, and is currently used in situations where the VC-dimension is not applicable. We will not cover the concept in this class, but if you are particularly interested in machine learning theory, then I recommend reading section 4.1 from this chapter. (It's only 2 pages.) Then in Section 4.2, the authors generalize Corollary 2.3 (finite hypothesis classes are PAC learnable) to the agnostic setting.

Problem 2. Read Chapter 5 of Shalev-Shwartz and Ben-David. (You may skip section 5.1, which formally defines the *No Free Lunch Theorem*.) Complete the following notes as you read.

1. Equation (5.7)

Problem 3. Read Chapter 6 of Shalev-Shwartz and Ben-David. (You may skip Section 6.5, which is only concerned with proofs.) Complete the following notes as you read.

1. Restriction of $\mathcal{H}$ to C (Definition 6.2)

2. Shattering (Definition 6.3)

3. VC-Dimension (Definition 6.5)

4. Theorem 6.6

5. The Fundamental Theorem of Statistical Learning (Theorem 6.7). You may ignore result 1 about uniform convergence (we're not covering uniform convergence in this class). You only need to know results 2-6.

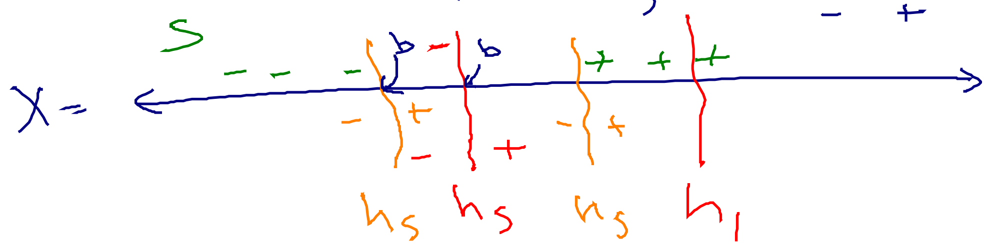
6. The Fundamental Theorem of Statistical Learning - Quantitative Version (Theorem 6.8). You may ignore result 1 about uniform convergence (we're not covering uniform convergence in this class). You only need to know results 2 and 3 about PAC and agnostic PAC learnability.

VC dim captures "how complicated" of a dataset can $\mathcal{H}$ do well on?

$\underbrace{\text{do well}}_{100\ L_S}$

Problem 4. For each hypothesis class below, formally define the hypothesis class and state its VC-dimension. You can find all of these answers in Section 6.3. You do not need to provide the proof of the VC-dimension below.

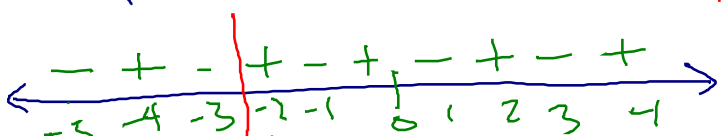
1. Threshold functions $X = \mathbb{R}$, $Y = \{\underset{-}{0}, \underset{+}{1}\}$



$\mathcal{H} =$ + if true, − if false

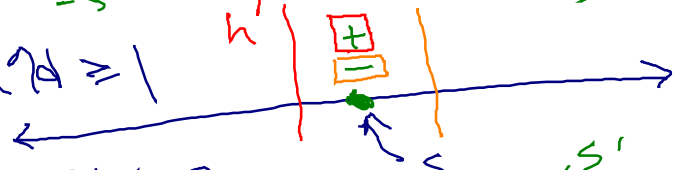
$$\mathcal{H} = \{ x \longmapsto \mathbb{1}[x > \underset{\uparrow \text{ parameter}}{b}] : b \in \mathbb{R} \}$$

VCdim = 1

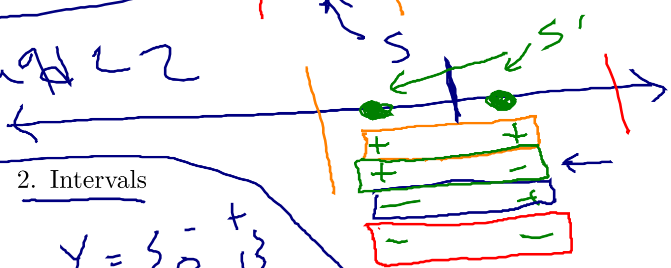


$\mathcal{H}$ shatters S b/c it can label every datapoint as either + or −

VCdim $\mathcal{H} \geq 1$



VCdim $\mathcal{H} < 2$



$\mathcal{H}$ does not shatter S'

VCdim is the size of the largest set shattered by $\mathcal{H}$

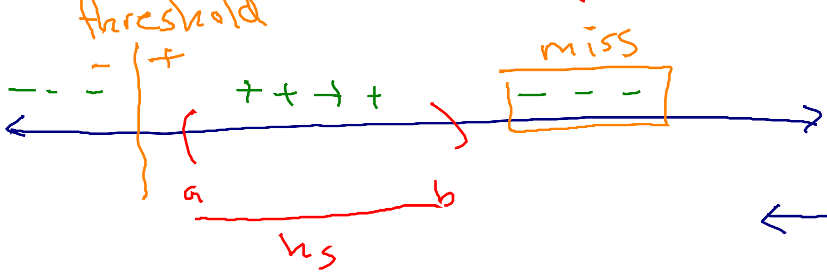
2. Intervals

$X = \mathbb{R}$, $Y = \{\underset{-}{0}, \underset{+}{1}\}$

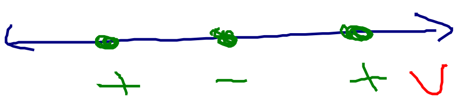
$$\mathcal{H} = \{ x \mapsto \mathbb{1}[x \in (a, b)] : a, b \in \mathbb{R} \}$$

parameter

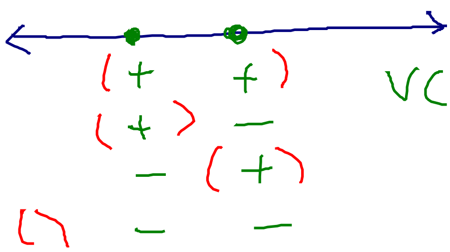
VCdim = 2



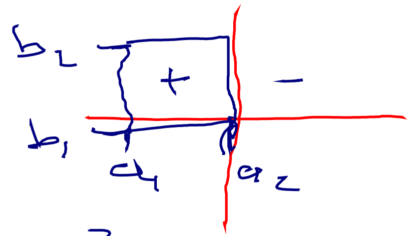
$L_S(h_S) = 0$



VCdim < 3



VCdim ≥ 2

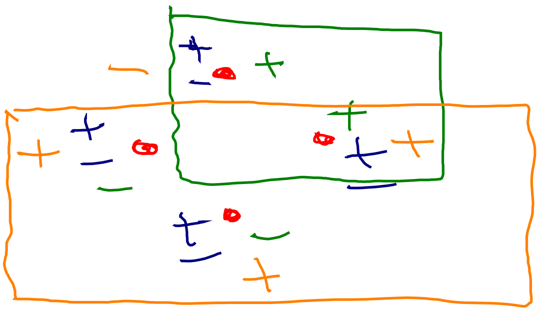


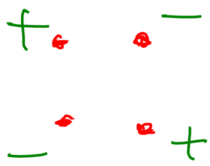
3. Axis Aligned Rectangles

$\mathcal{X}$ $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{-1, +1\}$

$h \in \mathcal{H}$ $\mathcal{H} = \{ \underline{h(a_1, a_2, b_1, b_2)} : \overbrace{a_1, a_2, b_1, b_2 \in \mathbb{R}}^{\text{parameters}} \}$

$$h(a_1, a_2, b_1, b_2)(x_1, x_2) \mapsto \begin{cases} +1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ -1 & \text{else} \end{cases}$$





$VCdim(\mathcal{H}) < 5$

$\mathcal{H}$ shatters ⋰ $\Rightarrow VCdim(\mathcal{H}) \geq 4$

4. Finite Classes

3.2

VCdim $\Leftrightarrow$ worst case analysis

$VCdim(\mathcal{H}) = 4$

$*$ $VCdim(\mathcal{H}) \leq |\mathcal{H}|$

combined w/ Thm 6-8

$\Rightarrow$ alternative proof of 3.2

Problem 5. Prove or disprove the following statements. Note that all of the proofs/disproofs follow immediately from the definitions above, and that is why they are included in this section. You do not have to complete all of these problems before the start of lecture. We will discuss some of these problems during lecture, but I recommend you solve as many as you can on your own before lecture.

1. The following equation always holds:

$$L_{\mathcal{D}}(h_S) - \epsilon_{\text{est}} = \epsilon_{\text{app}} \tag{1}$$

True

2. The following equation always holds:

$$L_{\mathcal{D}}(h_S) - \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{= \epsilon_{\text{app}}} = \epsilon_{\text{est}} \tag{2}$$

True

3. The following equation always holds:

$$\epsilon_{\text{app}} \geq 0 \tag{3}$$

True

$$\min_{h \in \mathcal{H}} L_D(h) = \epsilon_{\text{app}}$$

≥ 0

$$L_D(h) = \mathop{\mathbb{E}}_{z \sim D} \ell(z, h)$$

4. The following equation always holds:

True

$$\epsilon_{\text{est}} \geq 0 \tag{4}$$

$$\mathcal{E}_{est} = L_D(h_S) - \mathcal{E}_{app}$$

$$\geq 0$$

5. The following equation always holds:

$$\epsilon_{\text{est}} < \epsilon_{\text{app}} \tag{5}$$

False

6. As the number of data points in the training set increases, the approximation error decreases.

False

depends only on $\mathcal{H}$

$$\min_{h \in \mathcal{H}} L_D(h)$$

7. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be hypothesis classes where $\mathcal{H}_1 \subset \mathcal{H}_2$. Then, the approximation error of $\mathcal{H}_1$ is greater than or equal to the approximation error of $\mathcal{H}_2$.

$\mathcal{H}_{HS\varphi_1} \subset \mathcal{H}_{HS\varphi_2}$

True

$\text{VCdim}\,\mathcal{H}_1 \le \text{VCdim}\,\mathcal{H}_2$

$\text{VCdim}\uparrow \Rightarrow \varepsilon_{app}\downarrow$

8. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be hypothesis classes where $\mathcal{H}_1 \subset \mathcal{H}_2$. Then, the estimation error of $\mathcal{H}_1$ is less than or equal to the estimation error of $\mathcal{H}_2$.

$\text{VCdim}\uparrow \Rightarrow \varepsilon_{est}\uparrow$

True

9. A model with a high approximation error and a low estimation error is underfitting.

True

10. A model with a low approximation error and a high estimation error is overfitting.

True

11. If VCdim($\mathcal{H}$) is finite, then $\epsilon_{\text{app}} = 0$.

False

12. If VCdim($\mathcal{H}$) is infinite, then $\epsilon_{\text{app}} = 0$.

False

For any $\mathcal{H}$ we actually use

True.

nearest neighbor

no $\mathcal{D}$

13. If VCdim($\mathcal{H}$) is finite, then the Bayes optimal predictor $f_{\mathcal{D}} \in \mathcal{H}$.

False

depends on $\mathcal{D}$

14. If VCdim($\mathcal{H}$) is infinite, then the Bayes optimal predictor $f_{\mathcal{D}} \notin \mathcal{H}$.

False

15. If VCdim($\mathcal{H}$) is infinite, then $L_{\mathcal{D}}(h) > 0$ for all distributions $\mathcal{D}$, datasets $S \sim \mathcal{D}^m$, and $h \in \mathcal{H}$.

16. If VCdim($\mathcal{H}$) is infinite, then $L_S(h_S) = 0$ for all distributions $\mathcal{D}$, datasets $S \sim \mathcal{D}^m$, and any ERM $h_S \in \mathcal{H}$.

17. If VCdim($\mathcal{H}$) is finite, then $\mathcal{H}$ is PAC learnable.

True FTSL

18. If $\mathcal{H}$ is agnostic PAC learnable, then it is also PAC learnable.

True FTSL

19. For every two hypotheses class $\mathcal{H}_1$ and $\mathcal{H}_2$, if $\mathcal{H}_1 \subset \mathcal{H}_2$, than $\text{VCdim}(\mathcal{H}_1) \leq \text{VCdim}(\mathcal{H}_2)$.

True

20. For every two hypotheses class $\mathcal{H}_1$ and $\mathcal{H}_2$, if $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2)$, then $\mathcal{H}_1 = \mathcal{H}_2$.

False

21. The ordinary least squares (OLS) hypothesis class discussed in the previous lecture notes has a finite VC dimension.

OLS : $Y = \mathbb{R}$ fat shattering

VCdim $\{-1, +1\}$

$Y = \{1, 2, 3 \dots, c\}$ Nataranj

rademacher complexity, covering numbers

doubling dim

realizability

PAC $\Rightarrow L_D(h^*) = 0$

2 Lecture

Problem 6. Combine Theorem 6.8 and the definition of (agnostic) PAC learnability to bound the generalization error of a hypothesis class based on its VC-dimension.

agnostic PAC: $L_D(h_S) \leq L_D(h^*) + \varepsilon$ ε_{est}

ε_{app}

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_D(h)$$

theorem $$m \leq C_2 \frac{VCdim(\mathcal{H}) + \log 1/\delta}{\varepsilon^2}$$

$$\varepsilon \leq \sqrt{C_2 \frac{VCdim(\mathcal{H}) + \log 1/\delta}{m}}$$

$$L_D(h_S) \leq L_D(h^*) + \sqrt{C_2 \frac{VCdim(\mathcal{H}) + \log 1/\delta}{m}}$$

$$* \quad L_D(h_S) - L_D(h^*) = O\left(\sqrt{\frac{VCdim(\mathcal{H})}{m}}\right)$$

$$* \quad |L_D(h_S) - L_S(h_S)| \leq \sqrt{C_2 \frac{VCdim\,\mathcal{H} + \log 1/\delta}{m}}$$

$$= O\left(\sqrt{\frac{VCdim\,\mathcal{H}}{m}}\right)$$

PAC $$m \leq \frac{VCdim\,\mathcal{H}\,\boxed{\log 1/\varepsilon} + \log 1/\delta}{\varepsilon}$$

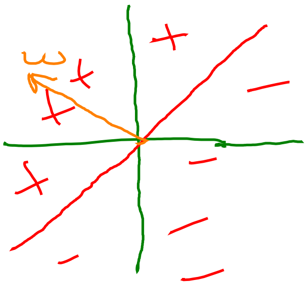
$$\varepsilon \leq \frac{VCdim\,\mathcal{H}\,\boxed{\log 1/\varepsilon} + \log 1/\delta}{m}$$

Half spaces

Problem 7. Linear models are one of the main tools in data mining. Chapter 9 in Shalev-Swartz and Ben-David discuss linear predictors in significantly more detail than we need for this class. In this problem, we will review all the relevant concepts.

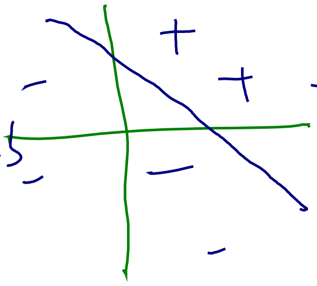
1. Define the hypothesis class of halfspaces.

$X = \mathbb{R}^d, \; Y = \{+1, -1\}$



homogenous

non-homogenous



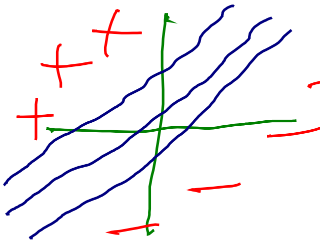
$$\mathcal{H}_{HS} = \{ x \mapsto \operatorname{sign}(x^T w + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

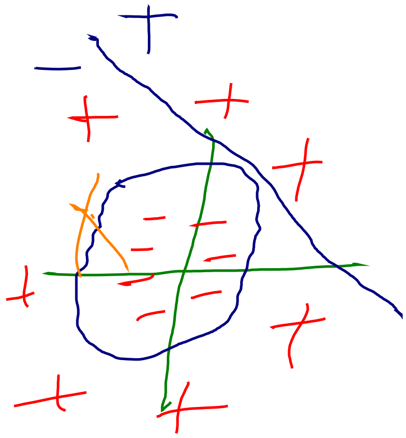
$$\operatorname{sign}(a) = \begin{cases} +1 & \text{if } a \geq 0 \\ -1 & \text{if } a < 0 \end{cases}$$

$$\operatorname{sign} : \mathbb{R} \to \{+1, -1\}$$

2. What is separability?

realizability for $\mathcal{H}_{HS}$





3. What is the VC-Dimension?

$$VCdim(\mathcal{H}_{HS}) = d + 1 = \Theta(d)$$

4. What is the computational complexity of computing the ERM in the separable and agnostic cases?

support vector machine (SVM)

logistic regression

perceptron

linear descriminant analysis (LDA)

linear naive baye

polynomial d^3

NP-hard $\approx 2^d$

Problem 8. Kernel functions are tools that let us manipulate the VC-dimension of hypothesis classes. They also have nice computational properties, but in this problem we are only concerned with their statistical properties.

↑ VCdim

1. Define the polynomial kernel.

$\phi \quad \varphi_p : \mathbb{R}^d \longrightarrow \mathbb{R}^{d'} \qquad d' > d \qquad p = \text{degree}$

$\varphi_p(x) = (x_1, x_2, \dots, x_d) \leftarrow p = 1 \qquad d$

$p = 3 \quad d^3$

$x_1 x_1 x_1$

$x_1 x_1 x_2$

$\vdots$

$x_1 x_1, x_1 x_2, x_1 x_3, \dots, \boxed{x_1 x_d},$

$x_2 x_1, x_2 x_2, \dots, \boxed{x_2 x_d} \leftarrow p = 2$

$\vdots$

$\boxed{x_d x_1}, \boxed{x_d x_2}, \dots, x_d x_d$

d^2

$\frac{d(d-1)}{2} = \Theta(d^2)$

φ_p contains all degree p terms that can be created from x

2. What is the VC-dimension of halspaces with the polynomial kernel?

$\mathcal{H}_{HS\varphi_p} = \mathcal{H}_{HS} \circ \varphi_p = \{ x \mapsto \text{sign}(\varphi(x)^T w), \; w \in \mathbb{R}^{d'} \}$

$\text{VCdim}(\mathcal{H}_{HS\varphi_p}) = d' = \binom{p+d}{p}$

$= \Theta(\min\{d^p, p^d\})$

$f \circ g$

3. When would we use the polynomial kernel?

$\mathcal{E}_{app}^{\mathcal{H}_{HS\varphi_2}} \le \mathcal{E}_{app}^{\mathcal{H}_{HS\varphi_1}}$

if $\mathcal{E}_{app}$ is large,

$\downarrow \mathcal{E}_{app}$ by $\uparrow p$

$\mathcal{E}_{est}^{p=1} \leq \mathcal{E}_{est}^{p=2}$ if $\mathcal{E}_{est}$ is large, $\downarrow \mathcal{E}_{est}$

φ $\quad \downarrow$ VCdim $\quad \downarrow p$

4. Define the random projection kernel.

ψ

$$\psi : \mathbb{R}^d \to \mathbb{R}^p \qquad p \ll d$$

$$\psi(x) = Ax \qquad A : \mathbb{R}^{p \times d}$$

$$\mathcal{H}_{HS\psi} = \mathcal{H}_{HS} \circ \psi = \{x \mapsto \underbrace{w^T\psi(x)}_{w^TAx} : w \in \mathbb{R}^p\}$$

also parameters $O(pd)$

5. What is the VC-dimension of halspaces with the random projection kernel?

$$\text{VCDim}\ \mathcal{H}_{HS\psi} = p$$

6. When would we use the random projection kernel?

$\mathcal{E}_{app}$ is small, but $\mathcal{E}_{est}$ is large

$\uparrow \qquad \downarrow\downarrow\downarrow$