Fundamentals of Learning and Information Processing

Session 13: Statistical Machine Learning (III)

Gang Li

Deakin University, Australia

2021-08-09

Table of Content

No Free Lunch Theorem

PAC Learning

Leslie Valiant, Turing Award 2010

- For transformative contributions to the theory of computation, including
  - the theory of probably approximately correct (PAC) learning,
  - the complexity of enumeration and of algebraic computation,
  - and the theory of parallel and distributed computing.

What is Learnable and How to Learn?

For any finite hypothesis class $\mathscr{H}$, we have shown that:

- $\mathscr{H}$ is PAC learnable with sample complexity

$$m_{\mathscr{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathscr{H}|/\delta)}{\epsilon} \rceil$$

- $\mathscr{H}$ is Agnostic PAC learnable with sample complexity

$$m_{\mathscr{H}}(\epsilon, \delta) \leq \lceil \frac{2\log(2|\mathscr{H}|/\delta)}{\epsilon^2} \rceil$$

- This sample complexity is obtained using the $ERM_{\mathscr{H}}$ learning rule

What is more?

- What about infinite hypothesis classes?
- What is the sample complexity of a given class?
- Is there a generic learning algorithm that achieves the optimal sample complexity?

□

The No-Free-Lunch Theorem

Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain $\mathscr{X}$, and m be any number representing a training set size: $m \le |\mathscr{X}|/2$. Then, there exists a distribution $\mathscr{D}$ over $\mathscr{X} \times \{0,1\}$ such that

- There exists a function $f : \mathscr{X} \to \{0,1\}$ with $L_{\mathscr{D}}(f) = 0$
- With probability at least 1/7 over the choice of $S \sim \mathscr{D}^m$, we have $L_{\mathscr{D}}(A(S)) \ge 1/8$

Intuition.

- It means that: *for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner*.
- We need to show
  1. For any algorithm A that receives a training set S of m examples from $\mathscr{X} \times \{0,1\}$, there exists a function $f : \mathscr{X} \to \{0,1\}$ such that $L_{\mathscr{D}}(f) = 0$ and

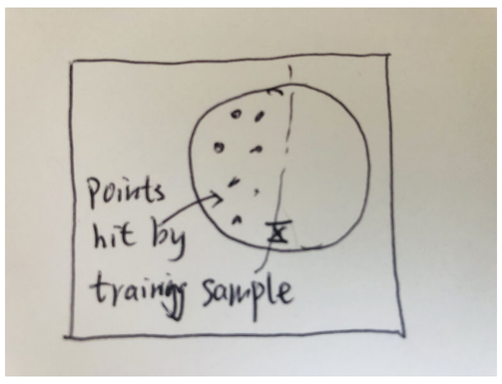
  $$E_{S \sim \mathscr{D}^m}[L_{\mathscr{D}}(A(S))] \ge 1/4$$

  2. From above, we derive that $P[L_{\mathscr{D}}(A(S))] \ge 1/8] \ge 1/7$

□

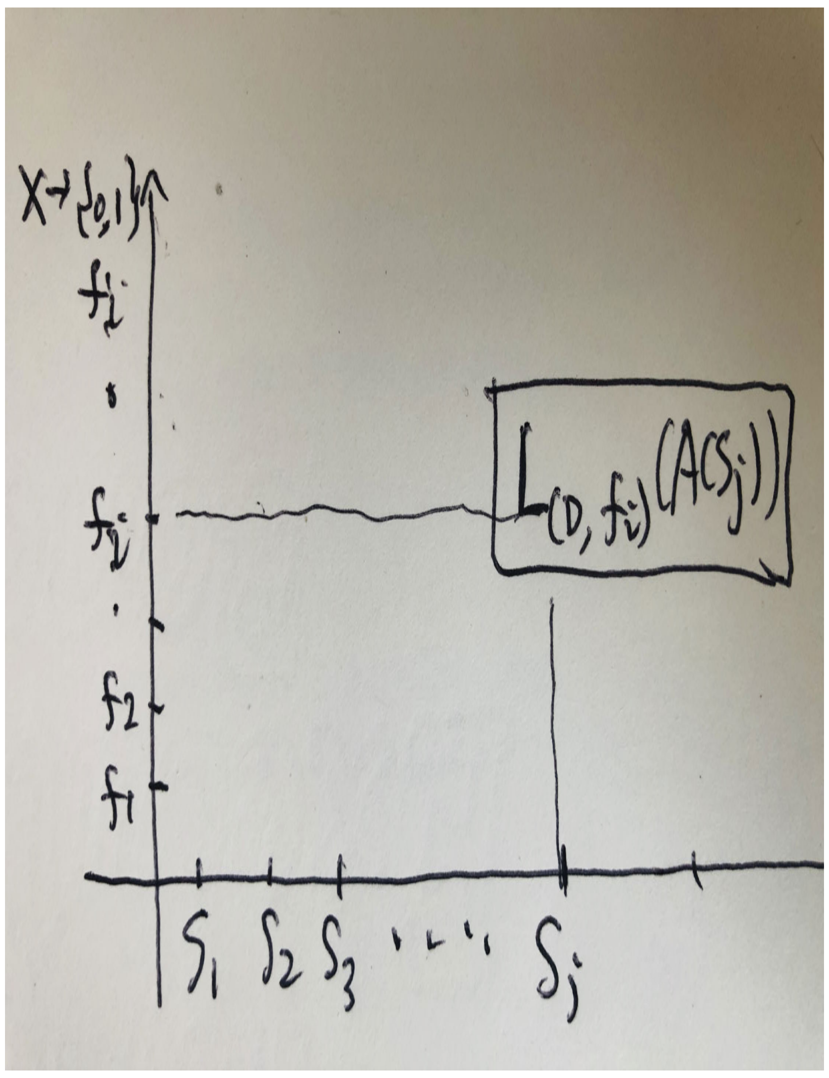
Proof-1.

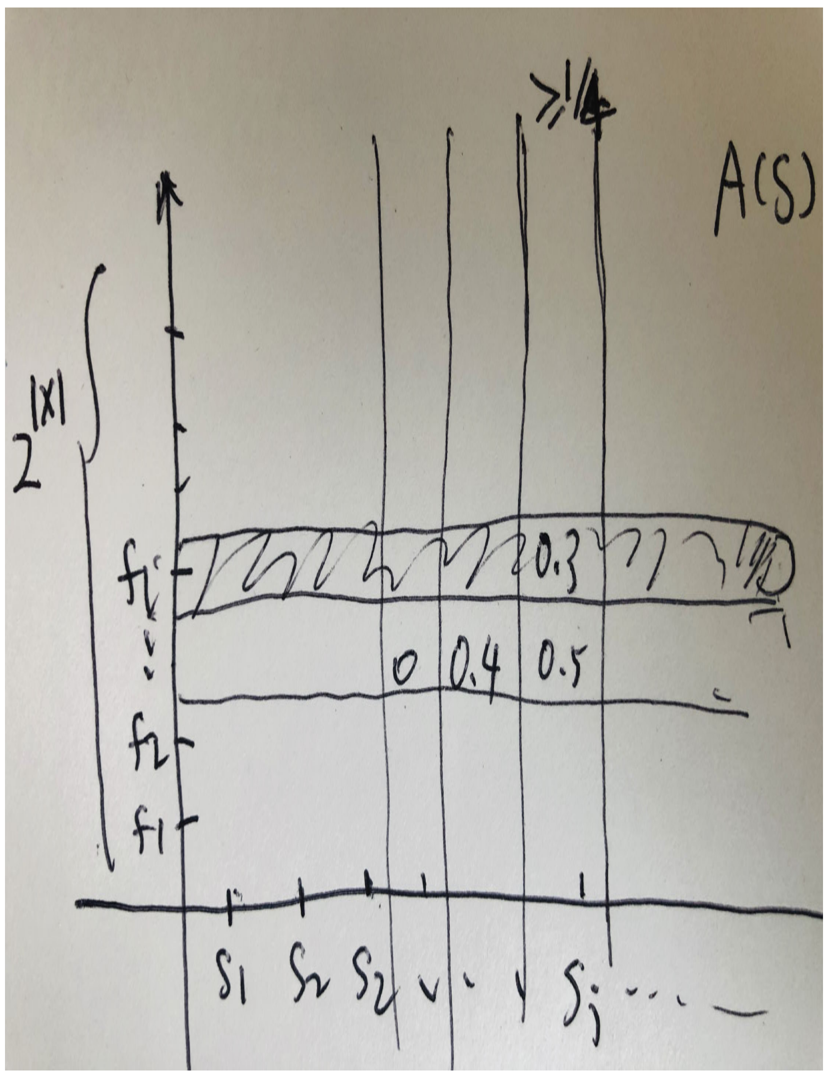
1. Estimate the probabilities that any given learner A will err on a random point: the probability that a random test point x, is not covered by S is at least 1/2.
2. Any hypothesis from A will has the probability 50 % of making error on x, so the expected error will be at least 1/4.



□

Proof-1.





□

Proof-2.

1. Let θ be a random variable that receives values in $[0,1]$ and $E[\theta] \ge 1/4$.
2. From Markov's Inequality,

$$P[\theta \ge 1/8] \ge \frac{1/4 - 1/8}{1 - 1/8} = 1/7$$

□

NFL Theorem and Prior Knowledge

Let $\mathscr{X}$ be an infinite domain set and $\mathscr{H}$ be the set of all functions $\mathscr{X} \to \{0,1\}$, then $\mathscr{H}$ is not PAC learnable.

Proof-V1.

- By way of contradiction: if $\mathscr{H}$ is learnable, choose $\epsilon < 1/8$ and $\delta < 1/7$.
- By the definition of PAC learnability: there exists a learning algorithm A and an integer $m = m(\epsilon, \delta)$, such that, for any distribution $\mathscr{D}$ over $\mathscr{X} \times \{0,1\}$, if for some function $f : \mathscr{X} \to \{0,1\}$, $L_{\mathscr{D}}(f) = 0$, then with probability greater than $1-\delta$, $L_{\mathscr{D}}(A(S)) \leq \epsilon$.
- From NFL, since $|\mathscr{X}| > 2m$, for every learning algorithm, there exists a distribution $\mathscr{D}$ such that with probability greater than $1/7 > \delta$, $L_{\mathscr{D}}(A(S)) > 1/8 > \epsilon$. contradiction!

□

Proof-V2.

- By way of contradiction: if $\mathscr{H}$ is learnable, then for some ϵ and δ we have $m_{\mathscr{H}} : (0,1)^2 \to \mathscr{N}$, for every distribution $\mathscr{D}$ over $\mathscr{X}$ and function f:

  $$P_{S \sim \mathscr{D}^m}[L_{\mathscr{D},f}(A(S)) > \epsilon] < \delta$$

  whenever $m > m_{\mathscr{H}}(\epsilon, \delta)$
- Let $m = 2m_{\mathscr{H}}(1/8, 1/7)$, $\mathscr{H}$ includes every possible function from S to $\{0,1\}$.
- From NFL, we have

  $$m_{\mathscr{H}}(1/8, 1/7) \geq m/2 > m_{\mathscr{H}}(1/8, 1/7)$$

  Contradiction!!!

□

We can escape the hazards foreseen by the NFL theorem by using our prior knowledge about a specific learning task, to avoid the distributions that will cause us to fail when learning that task.

- Such prior knowledge can be expressed by restricting our hypothesis class.

Error Decomposition

Let h_S be an $ERM_{\mathscr{H}}$ hypothesis: $h_S \in \text{argmin}_{h \in \mathscr{H}} L_S(h)$. Assume $f \in \mathscr{Y}^{\mathscr{X}}$ be the true hypothesis, and h^* is the best hypothesis in $\mathscr{H}$: $h^* = \min_{h^* \in \mathscr{H}} L_{\mathscr{D}}(h^*)$. Then we can have:

$$L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(f) = L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(h^*) + L_{\mathscr{D}}(h^*) - L_{\mathscr{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

The Approximation Error $\epsilon_{app} = L_{\mathscr{D}}(h^*) - L_{\mathscr{D}}(f)$

- How much risk do we have due to restricting to $\mathscr{H}$
- It does not depend on S
- It decreases with the complexity (size, or VC dimension) of $\mathscr{H}$

The Estimation Error $\epsilon_{est} = L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(h^*)$

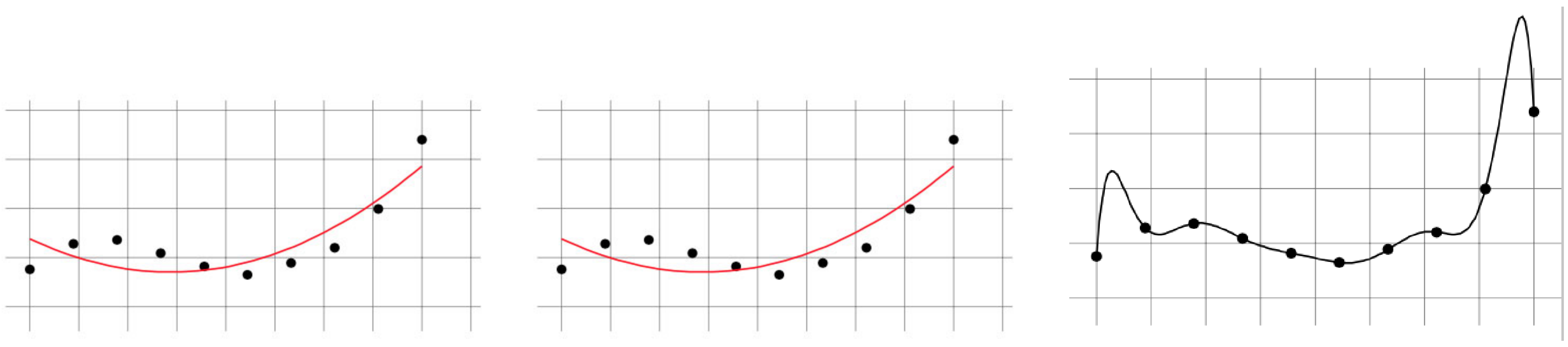
- Result of L_S being only an estimate of $L_{\mathscr{D}}$
- It decreases with the size of S
- It increases with the complexity of $\mathscr{H}$
- We have $\epsilon_{est} = L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(h^*) < 2\sup_{h \in \mathscr{H}} L_S(h) - L_{\mathscr{D}}(h)$. From Hoffding's Inequality, this provides an upper bound for ϵ_{est}

Bias-Complexity Tradeoff.

Overfitting Choosing $\mathscr{H}$ to be a very rich class will decrease the ϵ_{app}, but at the same time might increase the ϵ_{est}, as a rich $\mathscr{H}$ might lead to overfitting.
Underfitting Choosing $\mathscr{H}$ to be a very small class will reduce the estimation error ϵ_{est}, but at the same time might increase the approximation error ϵ_{app}, as a small $\mathscr{H}$ might lead to underfitting.
Bayes Optimal However, Bayes optimal classifier depends on the underlying distribution $\mathscr{D}$, which we do not know.

□

Bias-Complexity Tradeoff: Example.



(a) degree 2 (b) degree 3 (c) degree 10

□

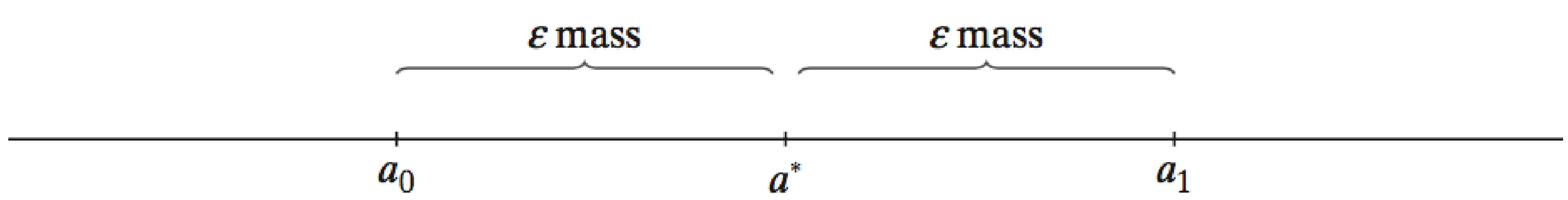
Is *Infinite Class* PAC learnable?

Threshold functions $\mathscr{X} = \mathscr{R}$, $\mathscr{Y} = \pm 1$, let $\mathscr{H}_\theta = \{x \mapsto sign(x - \theta) : \theta \in \mathscr{R}\}$

$\mathscr{H}_\theta$ is PAC learnable using ERM rule, with sample complexity of $m_{\mathscr{H}}(\epsilon, \delta) \le \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

1. Let a^* be the perfect threshold, namely $L_{\mathscr{D}}(h_{a^*}) = 0$. let $a_0 < a^* < a_1$ be such at $P[x \in (a_0, a^*)] = P[x \in (a^*, a^1)] = \epsilon$



2. Let b_S be the threshold for ERM result h_S, let $b_0 = \max\{x : (x, -1) \in S\}$ and $b_1 = \min\{x : (x, +1) \in S\}$.
3. A sufficient condition for $L_{\mathscr{D}}(h_s) > \epsilon$ can be rewritten as:

   $$P[L_{\mathscr{D}}(h_s) > \epsilon] \le P[b_0 < a_0 \vee b_1 > a_1] \le P[b_0 < a_0] + P[b_1 > a_1]$$

□

Proof.

4. The event $b_0 < a_0$ happens if and only if all examples in S are not in the internal (a_0, a^*), whose probability mass is defined to be ϵ:

   $$P[b_0 < a_0] = P[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \le e^{\epsilon m}$$

5. When we have $m > \log \frac{2/\delta}{\epsilon}$, $P[b_0 < a_0]$ is at most $\frac{\delta}{2}$.
6. Similarly, we have $P[b_1 < a_1] \le \frac{\delta}{2}$.
7. Combining with above, we conclude the proof.

□

Is *Infinite Hypothesis Class* PAC learnable?

Finite-Set functions Let $\mathscr{H}_F = \{x \mapsto h_F(x) : \text{ finite set } F \subseteq \mathscr{R}\} \bigcup \{x \mapsto 1\}$

$\mathscr{H}_F$ is not PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$

Proof.

1. Let $\mathscr{D}$ be the uniform distribution over $\mathscr{X}$, and the labelling function is $f = \{x \mapsto 1\}$.
2. Let $S \sim \mathscr{D}^m$ be a sample with size m from distribution $\mathscr{D}$: $S = \{(x_1, 1), \cdots, (x_m, 1)\}$
3. An ERM learner may pick a $h_F \in \mathscr{H}_F$ for $F = \{x_1, \cdots, x_m\}$, and with $L_S(h_F) = 0$.
4. What is $L_{\mathscr{D}}(h_F)$?
   - h_F fails on every test point which is outside F.
   - $L_{\mathscr{D}}(h_F) = 1$

□

The VC Dimension

Let $C = \{x_1, \cdots, x_m\} \subset \mathscr{X}$, and $\mathscr{H}$ be a class of hypothesis from $\mathscr{X}$ to $\{\pm 1\}$. The restriction $\mathscr{H}_C$ of $\mathscr{H}$ to C is the set of hypothesis from C to $\{\pm 1\}$ that can be derived from $\mathscr{H}$: $\forall x_i \in C$, $h_C(x_i) = h(x_i)$

- we can represent $\mathscr{H}_C$ as the vector $(h(x_1), \cdots, h(x_m)) \in \{\pm 1\}^m$

Defn

A hypothesis class $\mathscr{H}$ shatters a finite set $C \subset \mathscr{X}$ if the restriction of $\mathscr{H}$ to C is the set of all possible functions from C to $\{\pm 1\}$.

- in this case, we have $|\mathscr{H}_C| = 2^{|C|}$

The VC-dimension of a hypothesis class $\mathscr{H}$, denoted $VCdim(\mathscr{H})$, the maximal size of a set $C \subset \mathscr{X}$ that can be shattered by $\mathscr{H}$.

- If $\mathscr{H}$ can shatter sets of arbitrary large size, we say that $VCdim(\mathscr{H}) = \infty$

- To show that $VCdim(\mathscr{H}) = d$, we need to show
  - There exists a set C of size d which is shattered by $\mathscr{H}$
  - Any set C of size $d+1$ is not shattered by $\mathscr{H}$

The VC Dimension — Examples (1)

Threshold functions $\mathscr{X} = \mathscr{R}$, $\mathscr{Y} = \pm 1$, let $\mathscr{H}_\theta = \{x \mapsto sign(x - \theta) : \theta \in \mathscr{R}\}$

$VCdim(\mathscr{H}_\theta) = 1$

Demo.

- Show that $\{0\}$ is shattered
- Show that any two points can not be shattered

□

The VC Dimension — Examples (2)

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{ finite set } F \subseteq \mathcal{R}\} \bigcup \{x \mapsto 1\}$

$$VCdim(\mathcal{H}_F) = \infty$$

Demo.

- Show that any set can be shattered

□

The VC Dimension — Examples (3)

Finite Hypothesis Classes Let $\mathcal{H}$ be a finite hypothesis class,

$$VCdim(\mathcal{H}) < \log_2(|\mathcal{H}|)$$

Demo.

- Show that any set with at most $\log_2(|\mathcal{H}|)$ size can be shattered

□

The VC Dimension — Examples (4)

Intervals functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_{a,b} = \{h_{a,b}(x) : a < b \in \mathcal{R}\}$, where $h_{a,b}(x) = 1$ iff $x \in (a, b)$.

$$VCdim(\mathcal{H}_{a,b}) = 2$$

Demo.

- Show that $\{0, 1\}$ is shattered
- Show that any three points can not be shattered

□

The VC Dimension — Examples (5)

Axis aligned rectangles $\mathcal{X} = \mathcal{R}^2$, let $\mathcal{H}_{a_1,a_2,b_1,b_2} = \{h_{a_1,a_2,b_1,b_2}(x) : a_1 < a_2 \wedge b_1 < b_2\}$, where $h_{a_1,a_2,b_1,b_2}(x_1, x_2) = 1$ iff $x_1 \in (a_1, a_2)$ and $x_2 \in (b_1, b_2)$.

$$VCdim(\mathcal{H}_{a_1,a_2,b_1,b_2}) = 4$$

Demo.

□

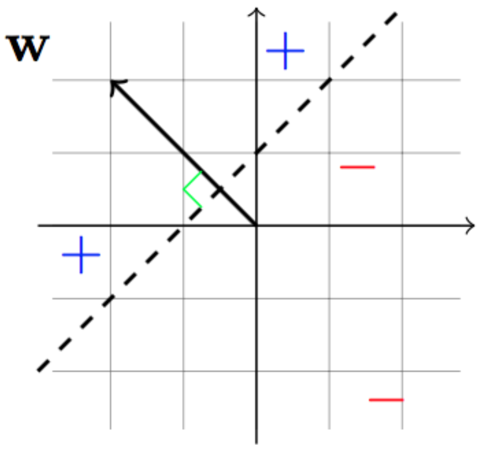
The VC Dimension — Examples (6)

Halfspaces $\mathcal{X} = \mathcal{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto sign(<\vec{w}, \vec{x}>) : \vec{w} \in \mathcal{R}^d\}$, where the inner product $<\vec{w}, \vec{x}> = \vec{w} \cdot \vec{x} = \sum_{i=1}^{d} w_i x_i$.

$VCdim(\mathcal{H}_S) = d$

Demo.

- Show that $\{e_1, \cdots, e_d\}$ is shattered
- Show that any $d+1$ points can not be shattered



□

Proof-1. Show that the across zero half space $VCdim(\mathcal{H}_S^0) \geq d$

- Consider the points $\{e_1, \cdots, e_d\}$, where $e_i = (0, \cdots, 1, \cdots, 0)$.
- Pick $B \subset \{e_1, \cdots, e_d\}$, let $h_B = (w_1, \cdots, w_d)$, with $w_i = \{\begin{matrix} 1 & \text{if } e_i \in B \\ -1 & \text{if } e_i \notin B \end{matrix}$
- $\forall e_i$, we have $h_B(e_i) = 1$ if $e_i \in B$, and -1 otherwise:

  $h_B(x) = sign(<\vec{w}, \vec{e}>) = w_i$

□

Proof-2. Show that the across zero half space $VCdim(\mathcal{H}_S^0) < d+1$

- Aim: given any set $A = \{x_1, \cdots, x_{n+1}\}$, which can not be shattered by $\mathcal{H}_S$.
- From linear algebra, we have $\sum_{i=1}^{d+1} a_i x_i = 0$
- We decompose them into positive set P and negative set N, and then have $\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$
- Let $B = \{x_i : i \in P\}$, let us apply $\mathcal{H}_B$ on both sides:

  $$\mathcal{H}_B(\sum_{i \in P} a_i x_i) = \sum_{i \in P} a_i \mathcal{H}_B(x_i) = \mathcal{H}_B(\sum_{j \in N} |a_j| x_j) = \sum_{j \in N} |a_j| \mathcal{H}_B(x_j)$$

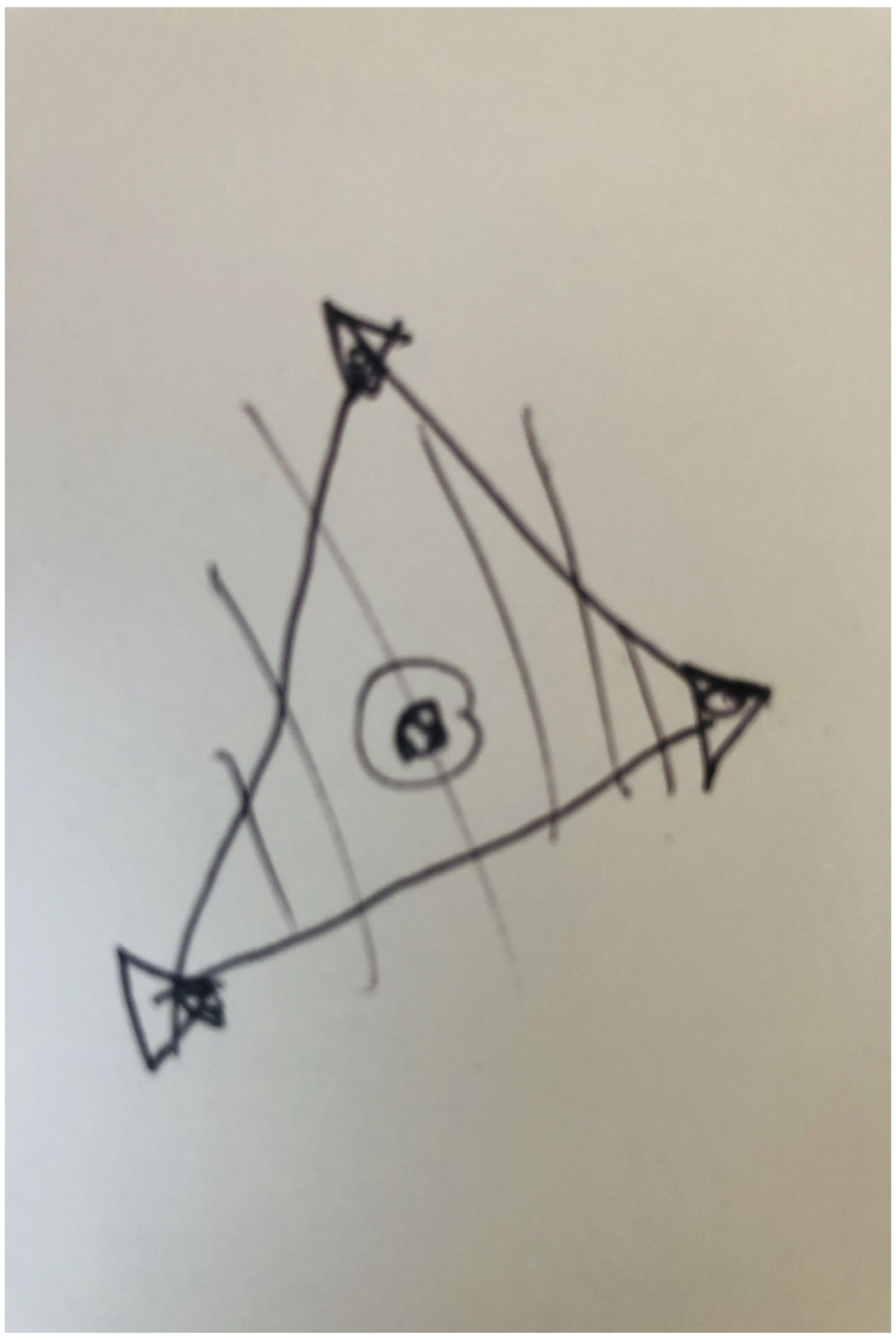
  the left hand side is positive, the right hand side is negative. Contradiction!!!

□

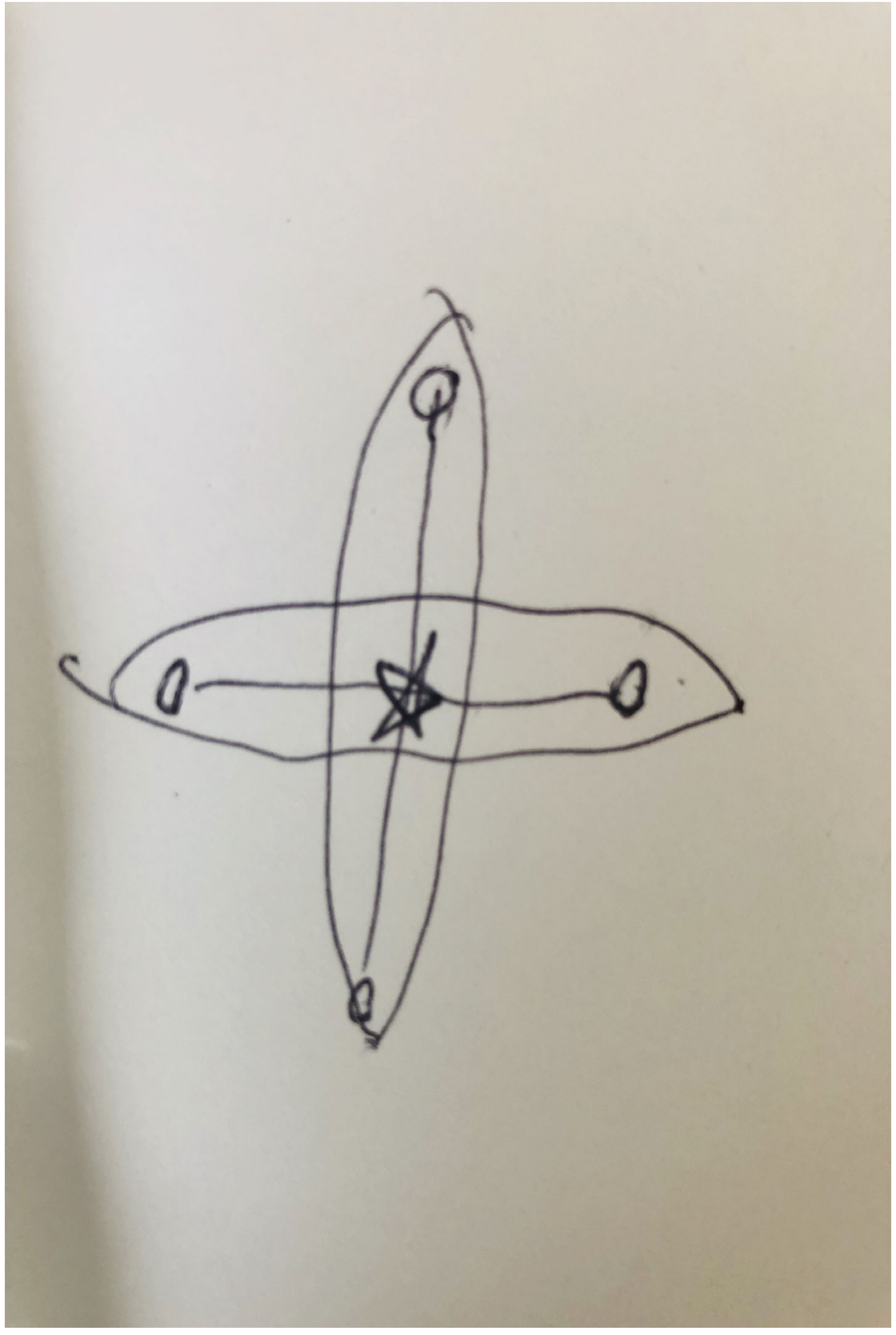
Radon's Lemma

For every n, for every $\{x_1, x_2, \cdots, x_{n+2}\} \subset \mathscr{R}^n$, there exists $B \subset \{x_1, x_2, \cdots, x_{n+2}\}$,

$$CovHull(B) \cap CovHull(\{x_1, x_2, \cdots, x_{n+2}\} - B) \neq \phi$$



(d)



(e)

No Free Lunch Theorem

1. To design a learning algorithm to predict whether patients are going to suffer a heart attack based on features includin *blood pressure* (BP), *body-mass index* (BMI), *age* (A), *level of physical activity* (P), and *income* (I).
   You have to choose between two algorithms: the first picks an axis aligned rectangle in the 2-dimensional space spanned by the features BP and BMI; and the other picks an axis aligned rectangle in the 5-dimensional space spanned by all the preceding features.
   - Explain the pros and cons of each choice.
   - Explain how the size of training sample affects your choice.
2. Prove that if $|\mathcal{X}| \geq km$ for a positive integer $k \geq 2$, then we can replace the lower bound of 1/4 in the No-Free-Lunch theorem with $\frac{k-1}{2k} = \frac{1}{2} - \frac{1}{2k}$. Namely, let A be a learning algorithm for the task of binary classification. Let m be any number smaller than $|\mathcal{X}|/k$, representing a training sample size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:
   - There exists a function f, which maps $\mathcal{X}$ to $\{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.
   - $E_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq \frac{1}{2} - \frac{1}{2k}]$.

The VC Dimension

1. Show the following monotonicity property of VC-dimension: for every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$, then $VCDim(\mathcal{H}') \leq VCDim(\mathcal{H})$.
2. For a finite hypothesis clas $\mathcal{H}$, $VCDim(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor$. However, this is just an upper bound. The VC-dimension of a class can be much lower than that:
   - Find an example of a class $\mathcal{H}$ of functions over the real interval $\mathbb{X} = [0,1]$, such that $\mathbb{H}$ is finite while $VCDim(\mathcal{H}) = 1$.
   - Give an example of a finite hypothesis class $\mathcal{H}$ over the domain $\mathbb{X} = [0,1]$, where $VCDim(\mathcal{H}) = \lfloor \log(|\mathcal{H}|) \rfloor$.

Questions?

Contact Information

Associate Professor **Gang Li**
School of Information Technology
Deakin University
Geelong, Victoria 3216, Australia

gangli@tulip.org.au
Open Resources of TULIP-Lab
Team for Universal Learning and Intelligent Processing