Fundamentals of Statistical Learning Theory

INSIDE-HEART Spring School – AI

Matteo Papini

March 26, 2026

UNIVERSITÀ DEGLI STUDI DI MILANO

Statistical Learning Theory

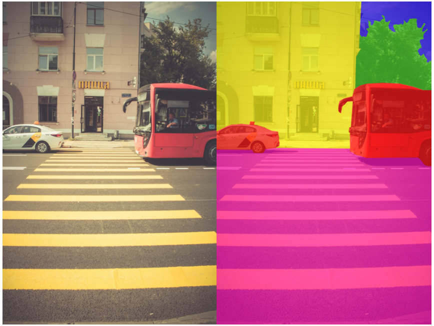
In the words of Vapnik:[1]

- *"...theoretical analysis of the problem of function estimation from a given collection of data..."*
- *"...a tool for creating practical algorithms for estimating multidimensional functions."*

[1] Vladimir Vapnik (1999). "An overview of statistical learning theory". In: *IEEE Trans. Neural Networks* 10.5, pp. 988–999.

The Problems of Machine Learning...

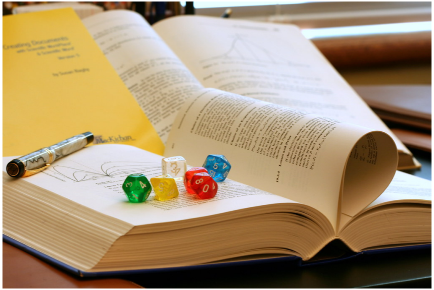
- **Classification**
- Regression
- Clustering
- Prediction
- Decision-making
- ...



...Under the Lens of Probability Theory

Assumptions on the *distribution* of data

- Independence
- Stationarity
- ...



Some Important Questions

- What are the theoretical limits of a learning algorithm?
- How much data do I need?
- What is a good model?
- *Why do I need a test set?*

- The Learning Problem
- Bias-Variance Tradeoff
- PAC Learning
- VC Theory

(Supervised) Learning Problem

- $\mathcal{X}$: input space
- $\mathcal{Y}$: target space
- $\mathcal{D}$: data distribution over $\mathcal{X} \times \mathcal{Y}$
- $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$: loss function

Example: Binary Classification

- $\mathcal{X} \subseteq \mathbb{R}^d$: samples identified by their *features*
- $\mathcal{Y} = \{0, 1\}$: negative or positive class
- $\mathcal{D}$: distribution of samples with their *true* labels (unknown)
- Zero-one loss

$$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{otherwise} \end{cases}$$

Example: Single-Output Regression

- $\mathcal{X} \subseteq \mathbb{R}^d$: samples identified by their *features*
- $\mathcal{Y} = \mathbb{R}$: numerical measurement
- $\mathcal{D}$: distribution of samples with their measurements
- Quadratic loss: $\ell(y, \widehat{y}) = (y - \widehat{y})^2$

Data and Datasets

The fundamental assumption on **data**:

$$(X, Y) \sim \mathcal{D} \qquad \textbf{i.i.d.}$$

A **dataset** is formed by sampling $(X_i, Y_i) \sim \mathcal{D}$ *independently* for $i = 1, \dots, n$

$$S = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$$

I will write $S \sim \mathcal{D}^n$

Predictors

A **predictor** is a function

$$h : \mathcal{X} \to \mathcal{Y}$$

Examples:

- Classifier: $h : \mathbb{R}^d \to \{0, 1\}$
- Regressor: $h : \mathbb{R}^d \to \mathbb{R}$

Statistical Risk

A *good* predictor has small **Statistical Risk** (a.k.a. Population Risk a.k.a. Expected Loss)

$$\mathcal{L}(h) = \mathbb{E}\left[\ell(h(X), Y)\right]$$

where $(X, Y) \sim \mathcal{D}$

It's the expected loss of a "test" sample

Bayes Optimal Predictor

$$h^*(x) = \arg\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)\middle|X = x\right]$$

Bayes risk: $\mathcal{L}(h^*)$ (typically *larger* than zero)

Theorem

$\mathcal{L}(h^*) \le \mathcal{L}(h)$ for every predictor h

Bayes Optimal Predictor

$$h^*(x) = \arg\min_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\left[\ell(\widehat{y}, Y) \middle| X = x\right]$$

Bayes risk: $\mathcal{L}(h^*)$ (typically *larger* than zero)

Theorem

$\mathcal{L}(h^*) \leq \mathcal{L}(h)$ for every predictor h

To compute the Bayes optimal predictor you need to know $\mathcal{D}$

Bayes Optimal Predictor: Examples (1/2)

Binary Classification: let $p(x) = \mathbb{P}(Y = 1|X = x)$

$$h^*(x) = \begin{cases} 0 \text{ if } p(x) < 1/2 \\ 1 \text{ otherwise} \end{cases}$$

Bayes risk: $\mathcal{L}(h^*) = \mathbb{E}[\min\{p(X), 1 - p(X)\}]$

Bayes risk is **zero** if labels are assigned deterministically

Bayes Optimal Predictor: Examples (2/2)

Regression:

$$h^*(x) = \mathbb{E}\left[Y | X = x\right]$$

Bayes risk: $\mathcal{L}(h^*) = \mathbb{E}\left[\mathrm{Var}[Y|X]\right]$

Bayes Optimal Predictor: Examples (2/2)

Regression:

$$h^*(x) = \mathbb{E}[Y|X = x]$$

Bayes risk: $\mathcal{L}(h^*) = \mathbb{E}[\mathrm{Var}[Y|X]]$

Gaussian model: $Y = f(X) + \eta, \quad \eta \sim \mathcal{N}(0; \sigma^2)$ i.i.d.

$$h^*(x) = f(x)$$

Bayes risk: $\mathcal{L}(h^*(x)) = \sigma^2$

Empirical Risk

a.k.a. Average Loss

Given i.i.d. dataset $S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$

$$\mathcal{L}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

- Binary classification: $\mathcal{L}_S(h)$ = ratio of incorrect classifications
- Regression: $\mathcal{L}_S(h)$ = Mean Squared Error (MSE)

Theorem

For a **fixed** h (independent of S), $\mathbb{E}\left[\mathcal{L}_S(h)\right] = \mathcal{L}(h)$

- The Learning Problem
- **Bias-Variance Tradeoff**
- PAC Learning
- VC Theory

Hypothesis Classes

Let $\mathcal{S}$ be the set of all possible datasets, $\mathcal{F}$ be the set of all functions $\mathcal{X} \to \mathcal{Y}$

A **hypothesis class** is a set of predictors

$$\mathcal{H} \subseteq \mathcal{F}$$

A **learning algorithm** maps *datasets* to *predictors*

$$\mathtt{A} : \mathcal{S} \to \mathcal{F}$$

A learning algorithm (implicitly) defines a hypothesis class

$$\mathcal{H} = \{\mathtt{A}(S) \mid S \in \mathcal{S}\}$$

No-Free-Lunch Theorem

Consider *binary classification* with the zero-one loss

Theorem

Let $n < |\mathcal{X}|/2$. For every learning algorithm $\mathtt{A}$, there exists a data distribution $\mathcal{D}$ such that

1. There exists a predictor h^* such that $\mathcal{L}(h^*) = 0$
2. $\mathbb{P}(\mathcal{L}(\mathtt{A}(S)) \geq 1/8) \geq 1/7$

where $S \sim \mathcal{D}^n$

No-Free-Lunch Theorem's Lesson

There is no universal learner

- For every learner, there exists a task on which it fails
- We need prior knowledge about the **specific** task at hand
- Equivalent to restricting the hypothesis class $\mathcal{H}$

Figure 1: Not what we do

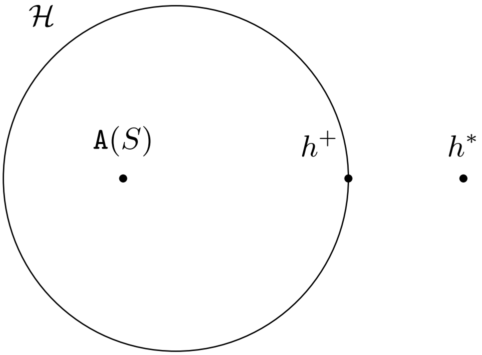
Bias-Variance Decomposition

Fix a learning problem $(\mathcal{X}, \mathcal{Y}, \mathcal{D}, \ell)$ and an algorithm $\mathtt{A}$ with hypothesis class $\mathcal{H}$

- Let h^+ be the best predictor in $\mathcal{H}$

$$h^+ \in \arg\min_{h \in \mathcal{H}} \mathcal{L}(h)$$

- By definition $\mathtt{A}(S) \in \mathcal{H}$
- Bayes optimal predictor h^* may *not* belong to $\mathcal{H}$

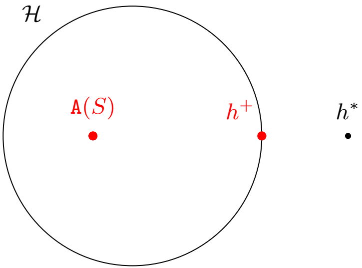


Bias-Variance Decomposition

$$
\begin{aligned}
\mathcal{L}(\mathtt{A}(S)) = \quad & \mathcal{L}(\mathtt{A}(S)) - \mathcal{L}(h^+) && \textbf{variance} \\
& + \mathcal{L}(h^+) - \mathcal{L}(h^*) && \textbf{bias} \\
& + \mathcal{L}(h^*) && \text{Bayes risk (irreducible)}
\end{aligned}
$$

Variance or Estimation Error

$$\mathcal{L}(\mathtt{A}(S)) - \mathcal{L}(h^+)$$



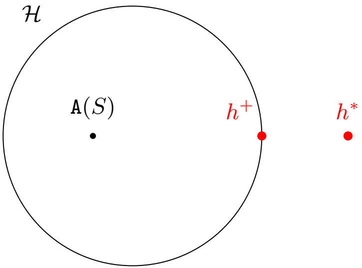
Large estimation error

$\implies$ S is not informative enough to identify h^+ within $\mathcal{H}$

$\implies$ **overfitting**

Bias or Approximation Error

$$\mathcal{L}(h^+) - \mathcal{L}(h^*)$$



Large approximation error

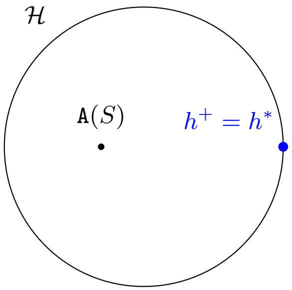
$\Longrightarrow$ $\mathcal{H}$ does not contain a good approximation of h^*

$\Longrightarrow$ **underfitting**

Realizability

If $\mathcal{H}$ contains the Bayes optimal predictor

$$h^+ = h^* \implies \mathcal{L}(h^+) - \mathcal{L}(h^*) = 0$$



No approximation error

- The Learning Problem
- Bias-Variance Tradeoff
- **PAC Learning**
- VC Theory

The Importance of Statistical Risk

- If **empirical risk** $\mathcal{L}_S(\mathtt{A}(S))$ is large, we are *underfitting* ($\mathcal{H}$ is too small)
- Even if $\mathcal{L}_S(\mathtt{A}(S))$ is small, we may be *overfitting* the training set S
- How would $\mathtt{A}(S)$ perform on a test sample $(X, Y) \sim \mathcal{D}$?
- This is measured by **statistical risk** $\mathcal{L}(\mathtt{A}(S))$

Measuring the Statistical Risk with a Test Set

Let $Q = \{(X'_1, Y'_1), \dots, (X'_m, Y'_m)\}$ be a **test set** of size m independent of S

Test error:

$$\mathcal{L}_Q(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(X'_i), Y'_i)$$

Theorem

The test error is an unbiased estimate of the statistical risk of $\mathtt{A}(S)$:

$$\mathbb{E}\left[\mathcal{L}_Q(\mathtt{A}(S)) \mid S\right] = \mathcal{L}(\mathtt{A}(S))$$

Upper-Bounding the Statistical Risk with the Test Error

Pick a *failure probability* $\delta \in (0, 1)$

Theorem

With probability $1 - \delta$

$$\mathcal{L}(\mathtt{A}(S)) \leq \mathcal{L}_Q(\mathtt{A}(S)) + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}}$$

Measuring the Statistical Risk without a Test Set

- All we have is **training error** $\mathcal{L}_S(\mathtt{A}(S))$
- $\mathcal{L}_S(\mathtt{A}(S))$ is **not** an unbiased estimate of the statistical risk!
- $\mathtt{A}(S)$ computed and evaluated on the same data $S \rightarrow$ loss **underestimation**
- It's a problem of *statistical dependence*

Upper Bounding the Statistical Risk with the Training Error

Consider a *finite* hypothesis class $\mathcal{H}$

Theorem (Generalization Bound)

With probability $1 - \delta$

$$\mathcal{L}(\mathtt{A}(S)) \leq \mathcal{L}_S(\mathtt{A}(S)) + \sqrt{\frac{1}{2n} \log \frac{|\mathcal{H}|}{\delta}}$$

Larger hypothesis class $\implies$ I can trust empirical risk less (more prone to overfitting)

Empirical Risk Minimization (ERM) Algorithm

Given hypothesis class $\mathcal{H}$

$$\texttt{ERM(S)} \in \arg\min_{h \in \mathcal{H}} \mathcal{L}_S(h)$$

- ERM minimizes the training loss
- It reflects the practice of fitting function approximators on training data
- Implementation can vary greatly depending on $\mathcal{H}$
- Minimization can be inexact in practice

PAC Bound

Probably Approximately Correct

PAC bound for ERM

With probability $1 - \delta$

$$\mathcal{L}(\mathtt{ERM}(S)) \leq \min_{h \in \mathcal{H}} \mathcal{L}(h) + \epsilon$$

where

$$\epsilon = \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}}$$

ERM is *probably* $(1 - \delta)$ *approximately* (ϵ) correct

PAC Learnability

Hypothesis class $\mathcal{H}$ is **PAC-learnable** if it admits a learning algorithm $\mathtt{A}$ and a **sample complexity** $n_{\epsilon,\delta}$ with the following property:

> *For every $\mathcal{D}$ and every $\epsilon > 0, \delta \in (0,1)$, with probability $1-\delta$ over datasets S of size $n \geq n_{\epsilon,\delta}$*

$$\mathcal{L}(\mathtt{A}(S)) \leq \min_{h \in \mathcal{H}} \mathcal{L}(h) + \epsilon$$

PAC Learnability of Finite Classes

A simple consequence of the ERM PAC bound

Theorem

Finite hypothesis classes are PAC learnable with sample complexity

$$n_{\epsilon,\delta} = \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{H}|}{\delta}$$

- The Learning Problem
- Bias-Variance Tradeoff
- PAC Learning
- VC Theory

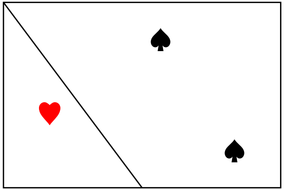
Linear Classifier

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{♠, ♥\}$ and

$$\mathcal{H} = \{h_{a,b} | \, a, b \in \mathbb{R}\}$$

where

$$h_{a,b}(x) = \begin{cases} ♠ & \text{if } a^\top x + b < 0 \\ ♥ & \text{otherwise} \end{cases}$$



$$|\mathcal{H}| = |\mathbb{R}^2| = \infty$$

Infinite Hypothesis Classes

If $|\mathcal{H}| = \infty$, PAC bounds become *vacuous*

$$\epsilon = \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}} = \infty$$

We need a different **measure of complexity** for $\mathcal{H}$

VC Dimension

Vapnik-Chervonenkis Dimension[2]

Consider binary classification with the zero-one loss

Hypothesis class $\mathcal{H}$ **shatters** finite set $E \subseteq \mathcal{X}$ if, *for any binary labeling* of the elements of E, there is $h \in \mathcal{H}$ that *perfectly* classifies them

[2] V. N. Vapnik and A. Ya. Chervonenkis (1971). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability & Its Applications* 16.2, pp. 264–280.

VC Dimension

Vapnik-Chervonenkis Dimension[2]

Consider binary classification with the zero-one loss

Hypothesis class $\mathcal{H}$ **shatters** finite set $E \subseteq \mathcal{X}$ if, *for any binary labeling* of the elements of E, there is $h \in \mathcal{H}$ that *perfectly* classifies them

$\mathrm{VC}(\mathcal{H})$ **is the cardinality of the largest set that can be shattered by** $\mathcal{H}$

If this is unbounded, $\mathrm{VC}(\mathcal{H}) = \infty$

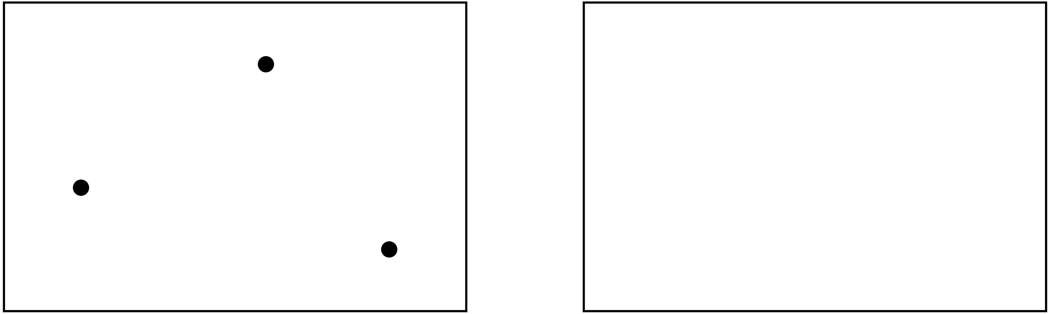
[2] V. N. Vapnik and A. Ya. Chervonenkis (1971). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability & Its Applications* 16.2, pp. 264–280.

VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$

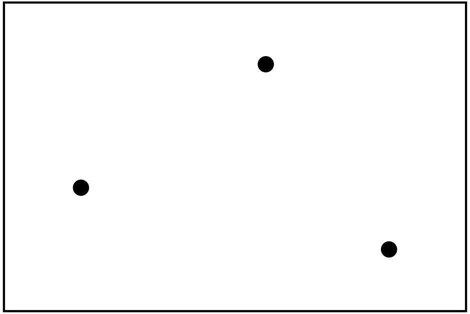
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

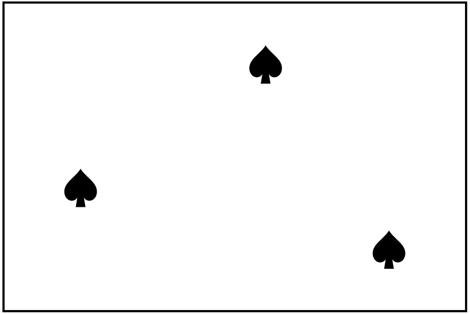
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

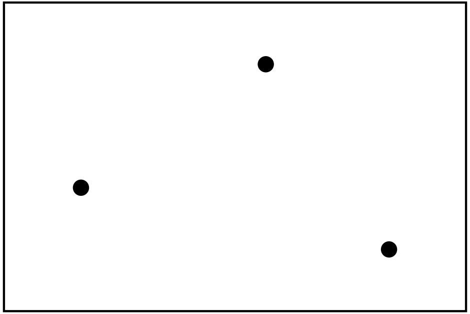
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$

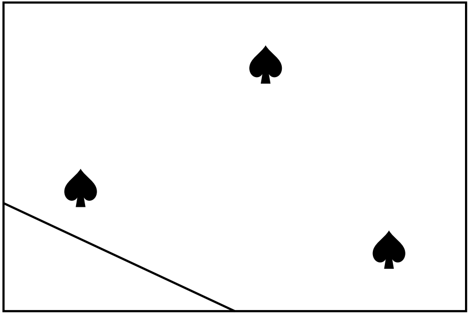




VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

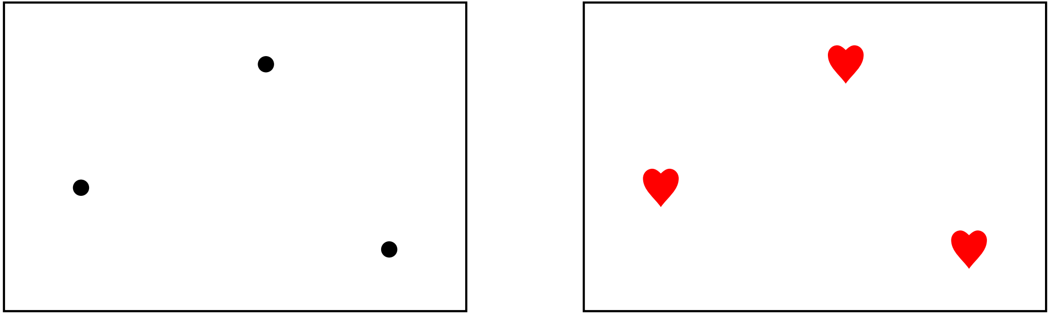
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$





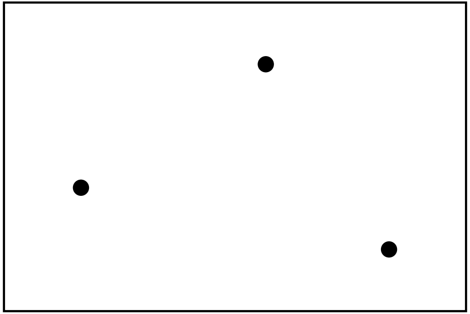
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

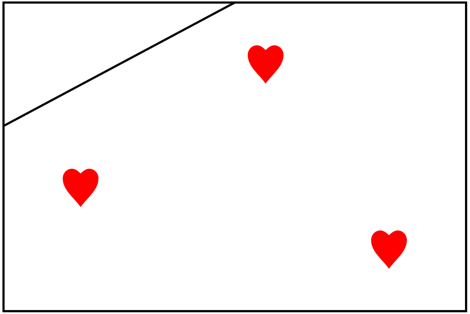
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

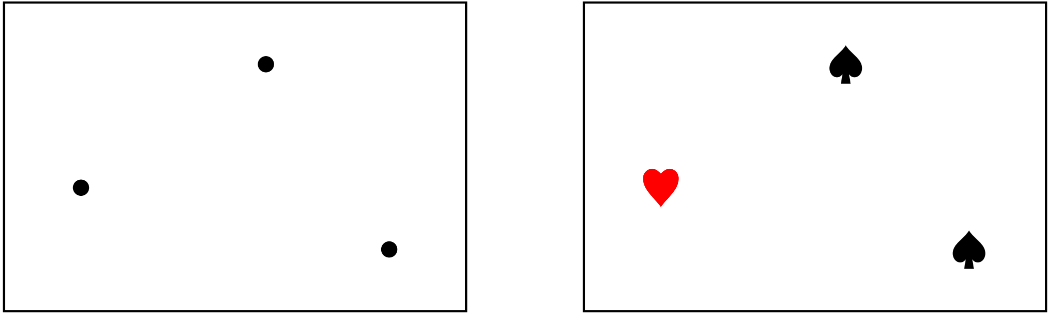
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$





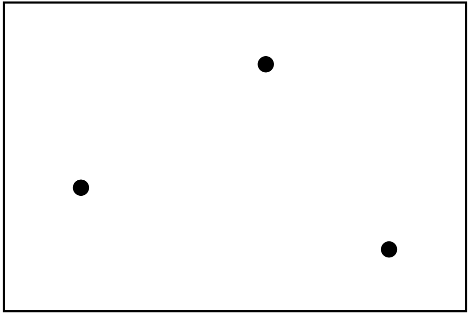
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

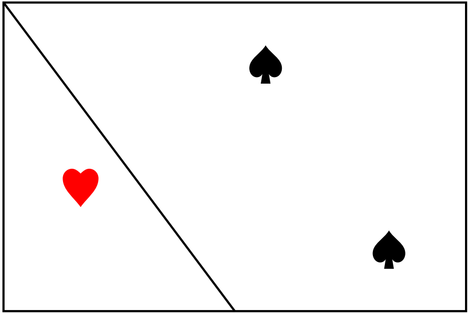
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

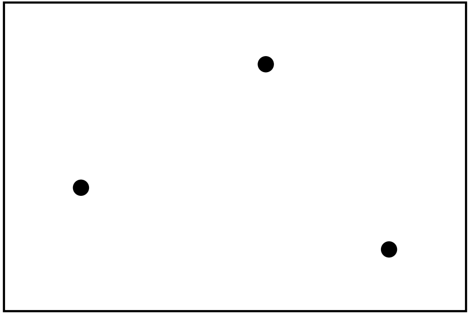
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$

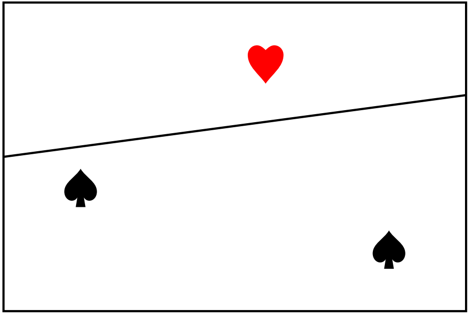




VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

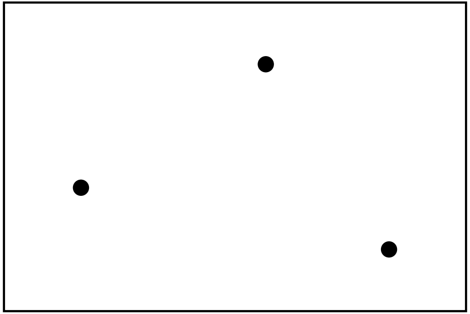
There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$

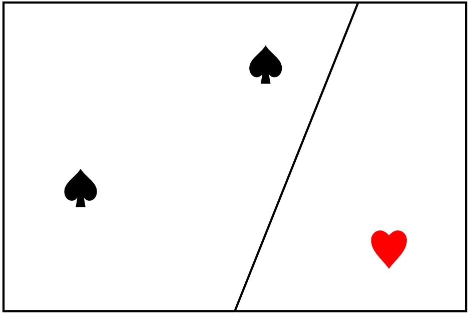




VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Lower Bound)

There is *at least one* set of 3 points that is shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) \geq 3$



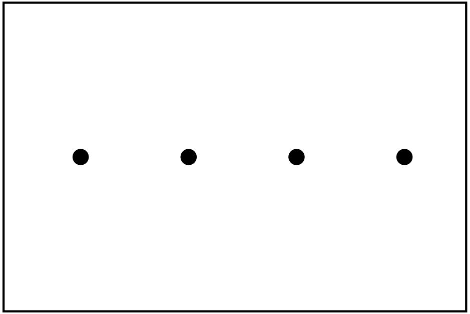


VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

No set of 4 (distinct) points can be shattered by $\mathcal{H} \implies \mathtt{VC}(\mathcal{H}) < 4 \implies \mathtt{VC}(\mathcal{H}) = 3$

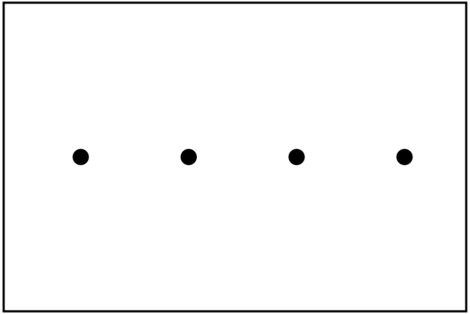
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

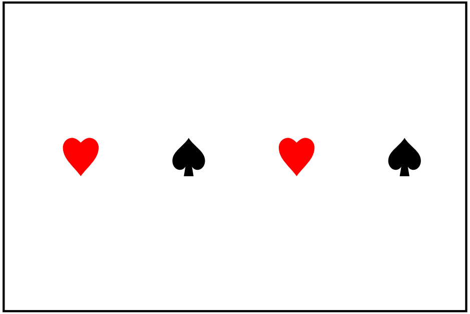
No set of 4 (distinct) points can be shattered by $\mathcal{H} \implies \mathrm{VC}(\mathcal{H}) < 4 \implies \mathrm{VC}(\mathcal{H}) = 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

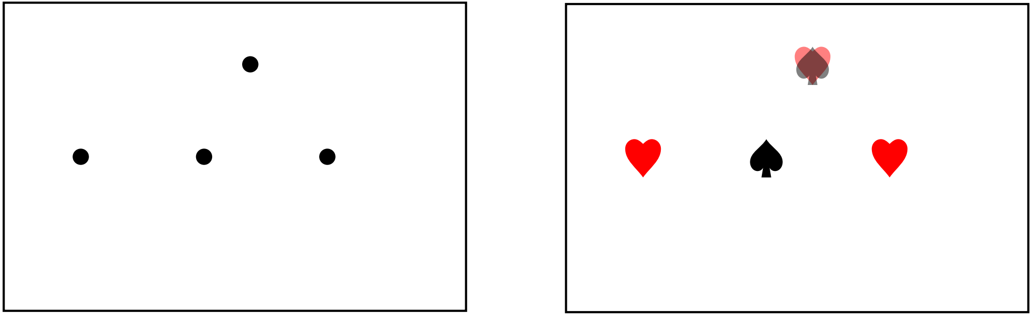
No set of 4 (distinct) points can be shattered by $\mathcal{H}$ $\implies$ $\mathtt{VC}(\mathcal{H}) < 4$ $\implies$ $\mathtt{VC}(\mathcal{H}) = 3$





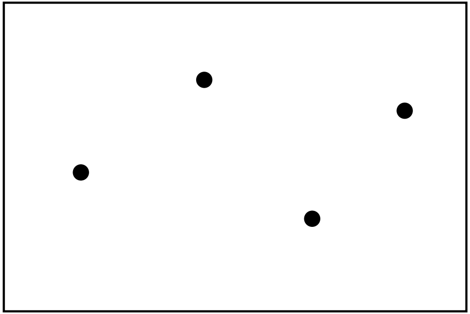
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

No set of 4 (distinct) points can be shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) < 4$ $\implies$ $\mathrm{VC}(\mathcal{H}) = 3$



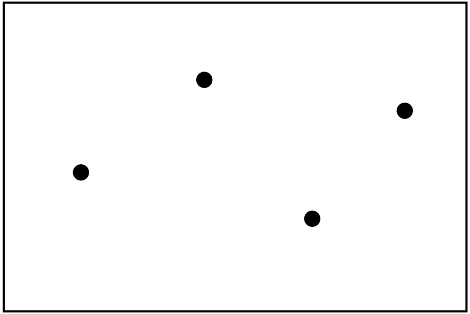
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

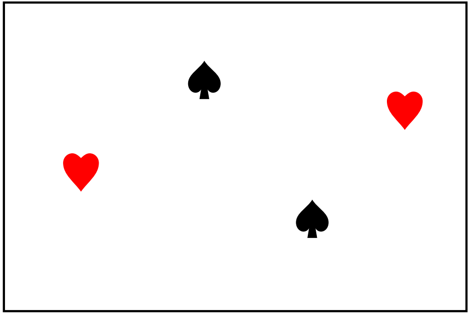
No set of 4 (distinct) points can be shattered by $\mathcal{H} \implies \mathrm{VC}(\mathcal{H}) < 4 \implies \mathrm{VC}(\mathcal{H}) = 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

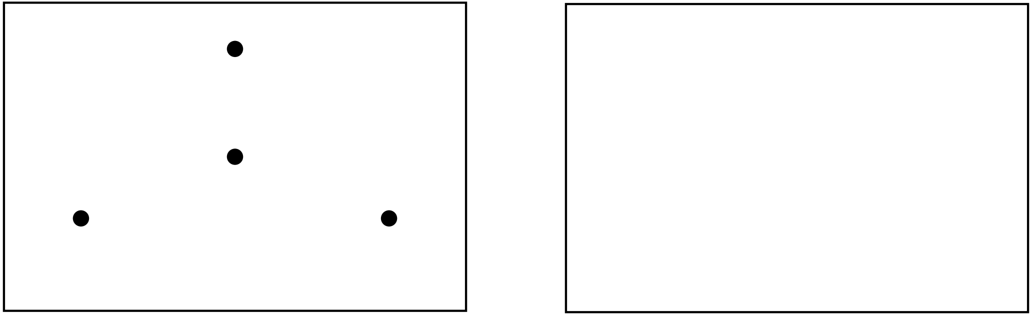
No set of 4 (distinct) points can be shattered by $\mathcal{H}$ $\implies$ $\mathrm{VC}(\mathcal{H}) < 4$ $\implies$ $\mathrm{VC}(\mathcal{H}) = 3$





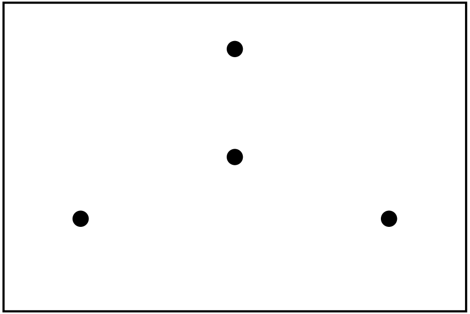
VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

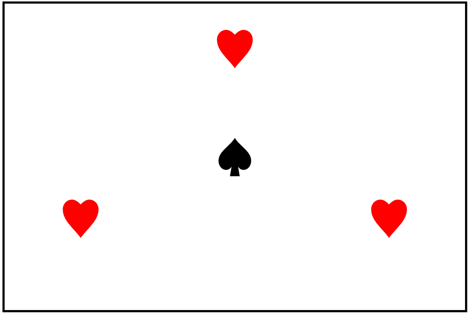
No set of 4 (distinct) points can be shattered by $\mathcal{H} \implies \mathrm{VC}(\mathcal{H}) < 4 \implies \mathrm{VC}(\mathcal{H}) = 3$



VC Dimension of Linear Classifiers in $\mathbb{R}^2$ (Upper Bound)

No set of 4 (distinct) points can be shattered by $\mathcal{H} \implies \mathrm{VC}(\mathcal{H}) < 4 \implies \mathrm{VC}(\mathcal{H}) = 3$





VC Dimension of Linear Classifiers

Theorem

The VC dimension of the class of linear classifiers in $\mathbb{R}^d$ is $d + 1$

Other examples

- If $|\mathcal{H}| < \infty$, $\text{VC}(\mathcal{H}) \leq \log_2 |\mathcal{H}|$
- Threshold functions in $\mathbb{R}$ have VC dimension 1
- Intervals in $\mathbb{R}$ have VC dimension 2
- Axis-aligned rectangles in $\mathbb{R}^2$ have VC dimension 4
- The VC dimension of neural networks with W weights is $O(W)$

VC and PAC Learnability

Consider binary classification with the zero-one loss

Theorem

$\mathcal{H}$ is PAC-learnable if and only if $\mathrm{VC}(\mathcal{H}) < \infty$

The sample complexity is

$$n_{\epsilon,\delta} = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})\log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$$

Beyond Fundamentals

The results presented here, their proofs, and more advanced results can be found in the following books:

Shai Shalev-Shwartz and Shai Ben-David (2014). *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press

Statistical Learning Theory

Thank you for listening!

Any questions?