Machine Learning
EDS
VC-dimension and learning guarantees for infinite models in binary classification

Janneke van Brummelen

Vrije Universiteit Amsterdam

Period 3.4
2023/2024

So far

- We have introduced the approximation/estimation error decomposition
  - Approximation error: how well a model can, theoretically, approximate a certain feature/output relationship
  - Estimation error: how close a predictor learnt from a sample is to the best performance possible within the chosen model
- More complex models can approximate more complex distributions, but make learning a predictor from the sample harder
- We obtained first learning guarantees for finite models: we quantified with probabilities how close an ERM predictor can be to the best performance possible within a model

Towards infinite models

- These guarantees are uninformative for infinite models
- For instance, in the non-deterministic case we obtained that with a probability of at least $1-\delta$

$$\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + C\sqrt{\frac{2\ln(\frac{2}{\delta}) + 2\ln\left(\mathrm{Card}S\right)}{n}}$$

  which is non-informative when $\mathrm{Card}S$ is infinite.
- In practice, infinite models are omnipresent: e.g. any model depending on even a single continuous parameter.

  Even linear models are uncountably infinite:

$$S = \{f_{\boldsymbol{w}} : \ f_{\boldsymbol{w}}(x) = \boldsymbol{w}^\top x \text{ for } \boldsymbol{w} \in \mathbb{R}^p\}$$

Towards infinite models

- Despite being infinite, some models show some very appreciable performance in practice
  Example: neural networks and SVMs for image classification
- This indicates that the latter learning guarantee might be too pessimistic, at least in some cases
- We need to go into a finer analysis

We will focus on binary classification: $\mathcal{Y} = \{0, 1\}$.

Table of Contents

Linear classification

- The plain number of different predictors a model contains might not be the right way to quantify its actual complexity
- To see this, let's consider the linear classification problem

Definition: Set of linear classifiers

Consider $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{Y} = \{0, 1\}$ and let us define linear predictors, also called in this case *linear classifiers*, as

$$f^{\text{class}}_{\boldsymbol{w},b} : x \in \mathcal{X} \mapsto \mathbb{1}_{\boldsymbol{w}^\top x + b \geq 0}, \qquad \text{for any} \quad \boldsymbol{w} \in \mathbb{R}^p, b \in \mathbb{R},$$

Define then the model $S^{\text{class}}_{\text{lin}}$ as the set of linear classifiers:

$$S^{\text{class}}_{\text{lin}} := \{ f^{\text{class}}_{\boldsymbol{w},b} : \quad \boldsymbol{w} \in \mathbb{R}^p, b \in \mathbb{R} \}.$$

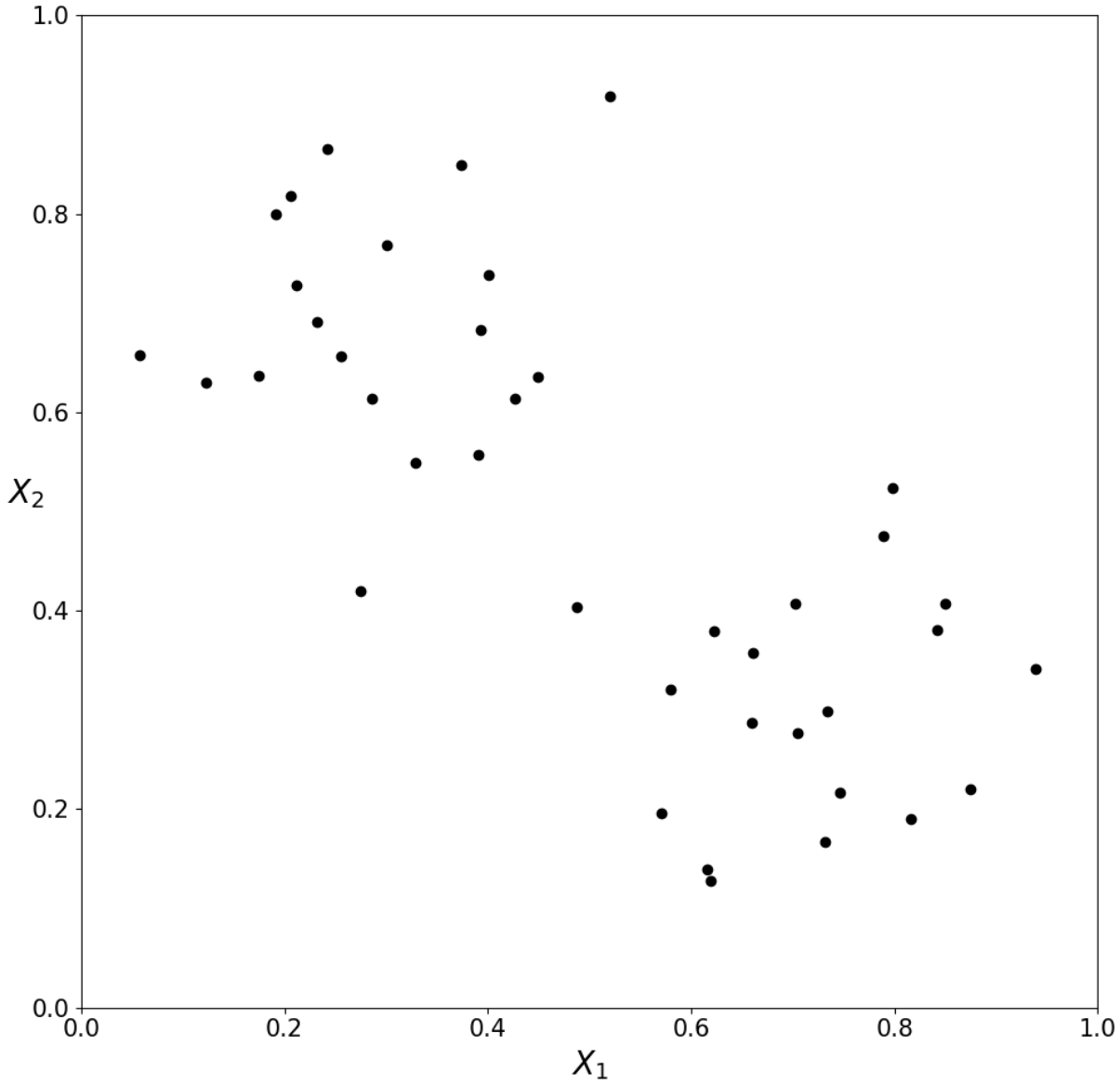
Interpretation of linear classification

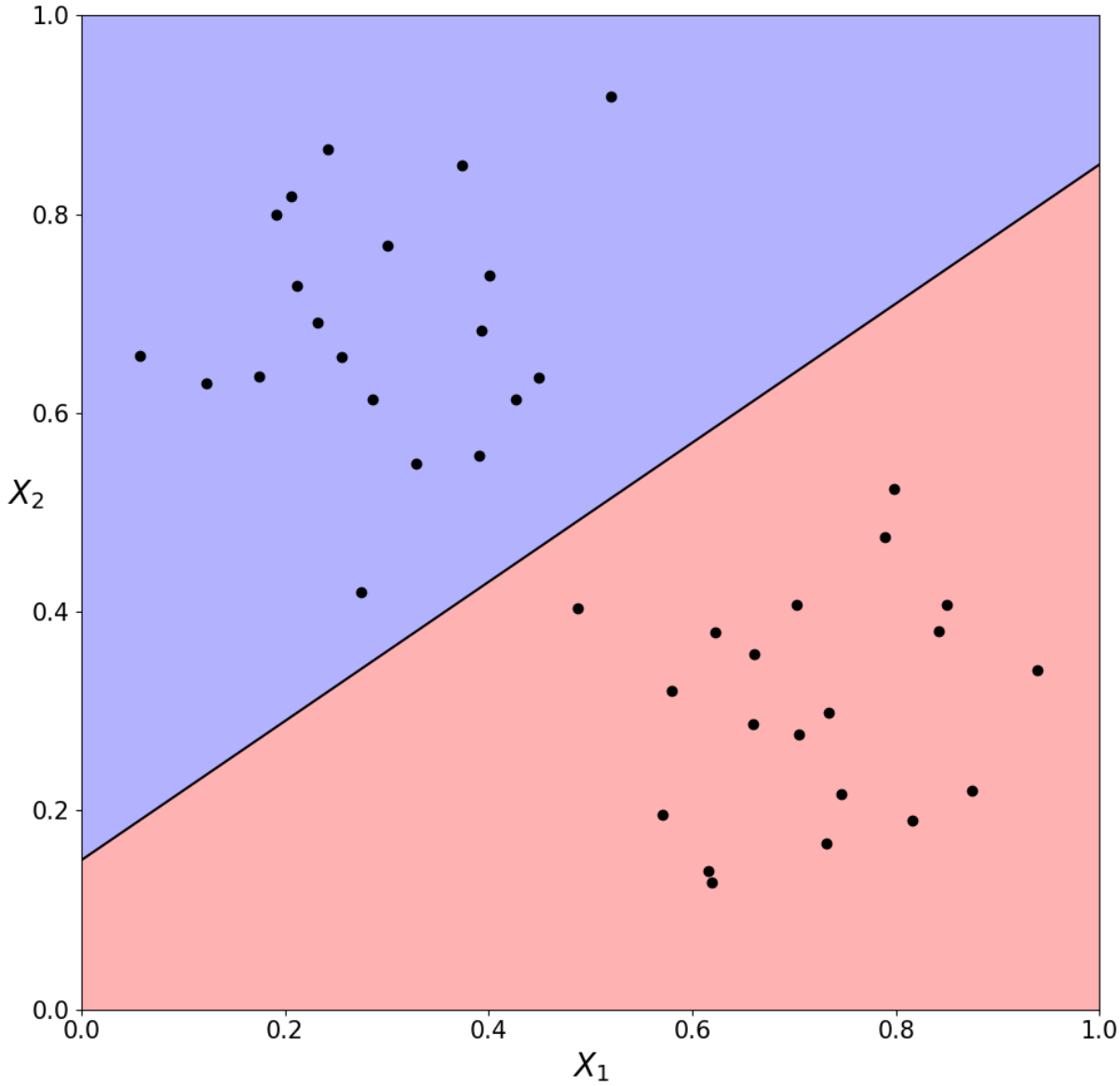
Interpretation

- For a given vector $\boldsymbol{w}$, the classifier $f^{\text{class}}_{\boldsymbol{w},b}$ basically separates $\mathbb{R}^p$ into two half-spaces.
- The separating border is the hyperplane orthogonal to $\boldsymbol{w}$.
- The constant b, often referred to as the bias, corresponds to a shift of the hyperplane away from the origin.

Remark: Towards SVMs

We will see next week that linear classifiers are the starting point of Support Vector Machine (SVM) algorithms.

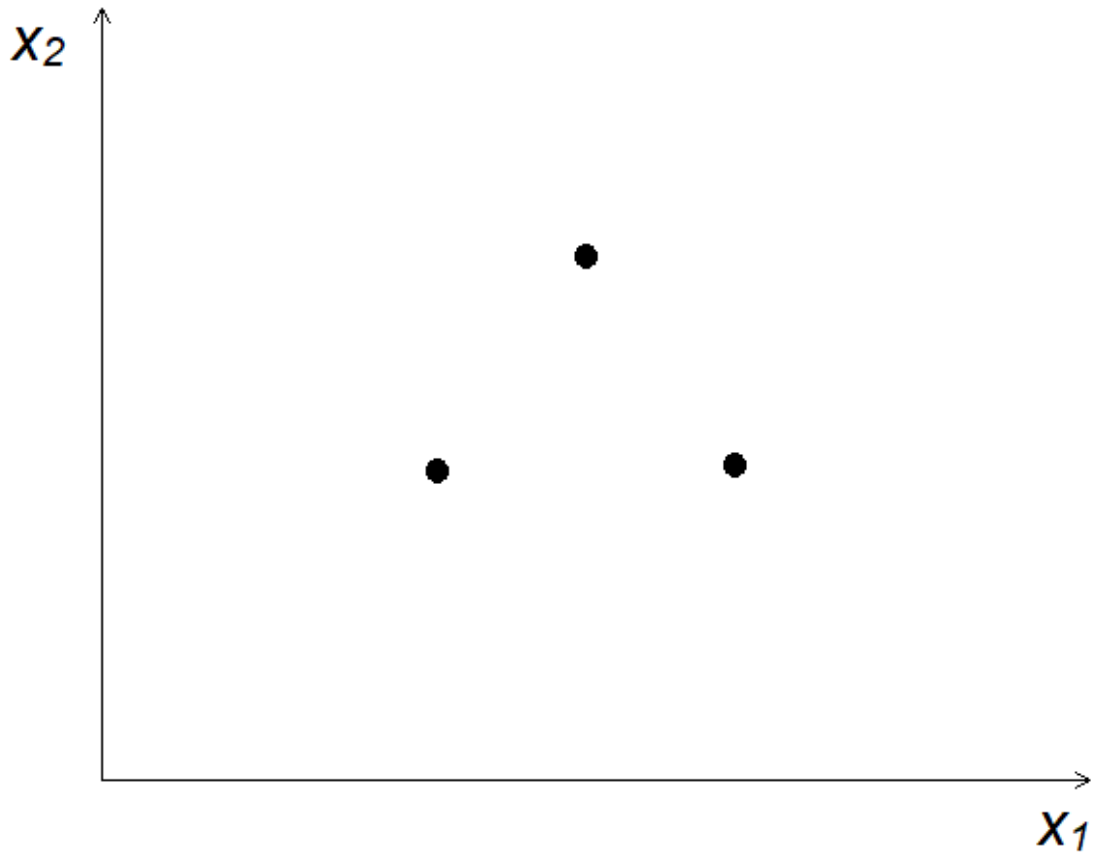

1.0
0.8
0.6
0.4
0.2
0.0
X_2
0.0
0.2
0.4
0.6
0.8
1.0
X_1


1.0
0.8
0.6
0.4
0.2
0.0
X_2
0.0
0.2
0.4
0.6
0.8
1.0
X_1

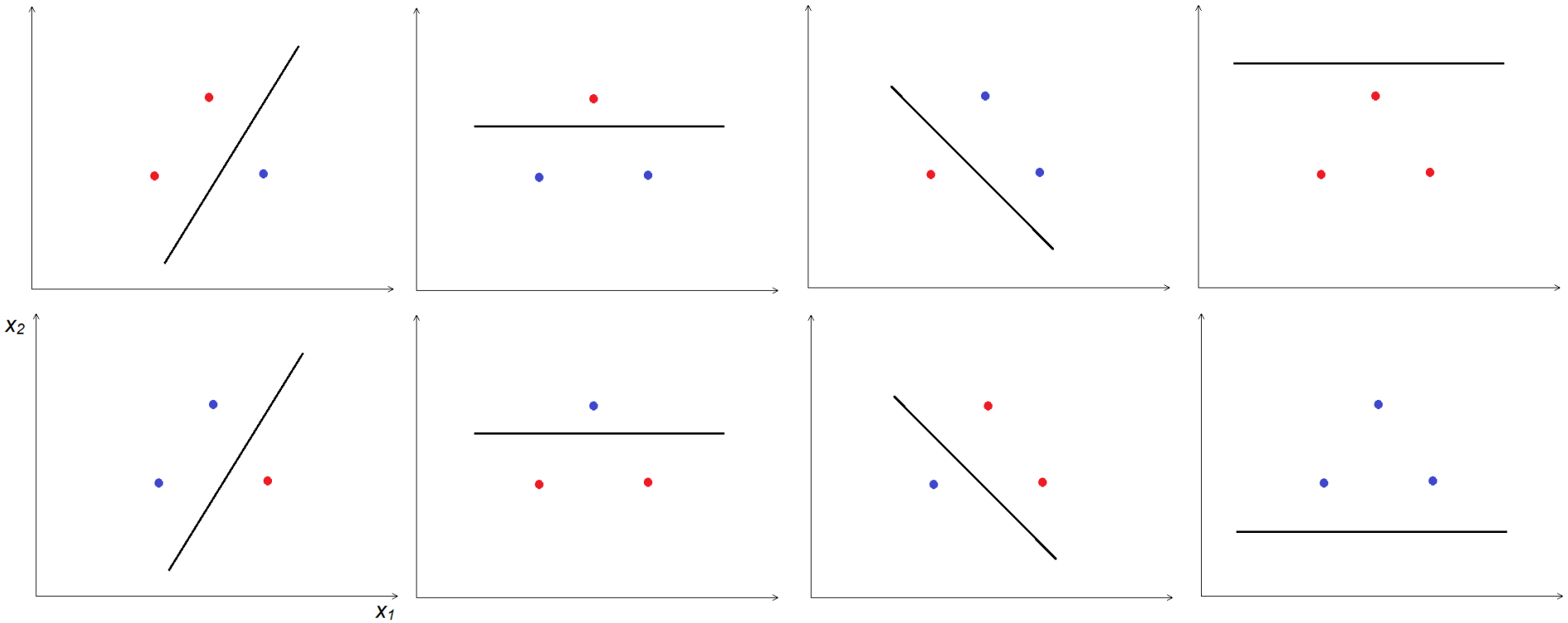
Complexity and discriminative capacity

- The model $S_{\text{lin}}^{\text{class}}$ contains an uncountably infinite number of predictors.
- Despite this apparent richness, linear classifiers are rigid predictors: they can only perform elementary dichotomies of the observed sample points
- If we consider n points $x_1, \ldots, x_n \in \mathbb{R}^2$, in how many ways can we separate the points using a single hyperplane?

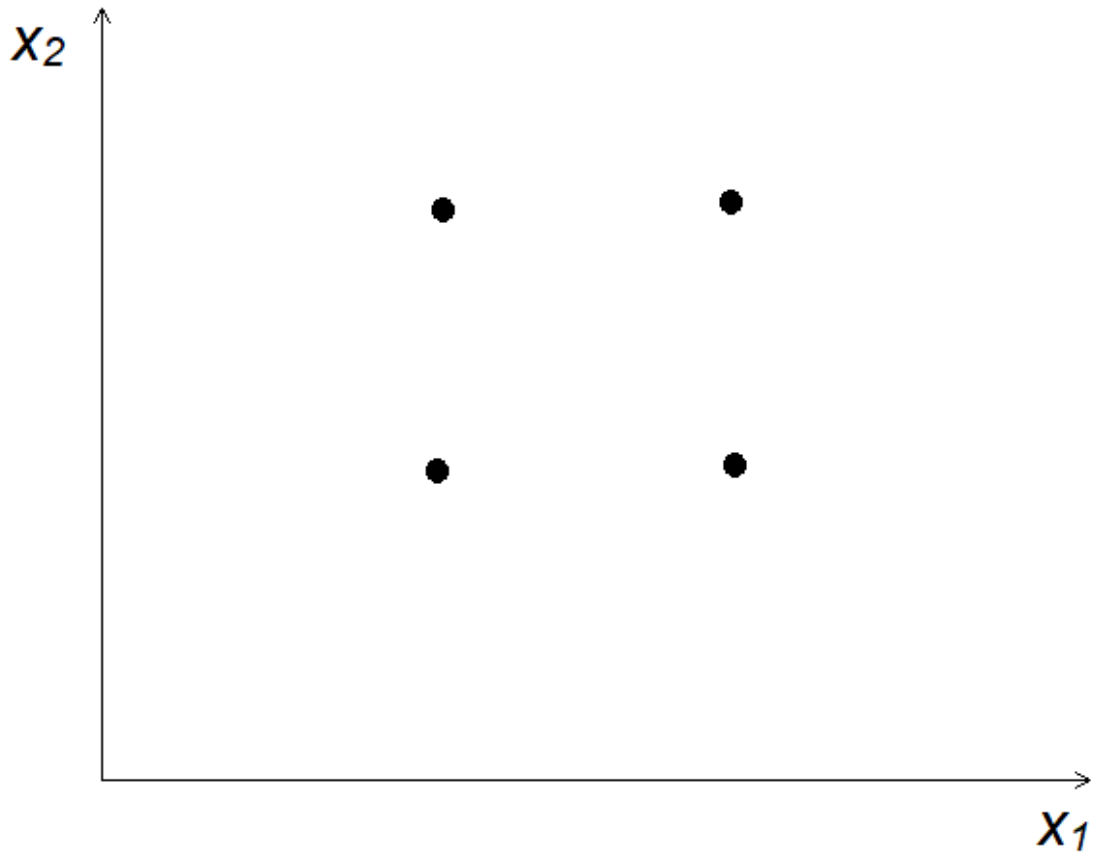
All possible labellings of 3 points by a linear classifier



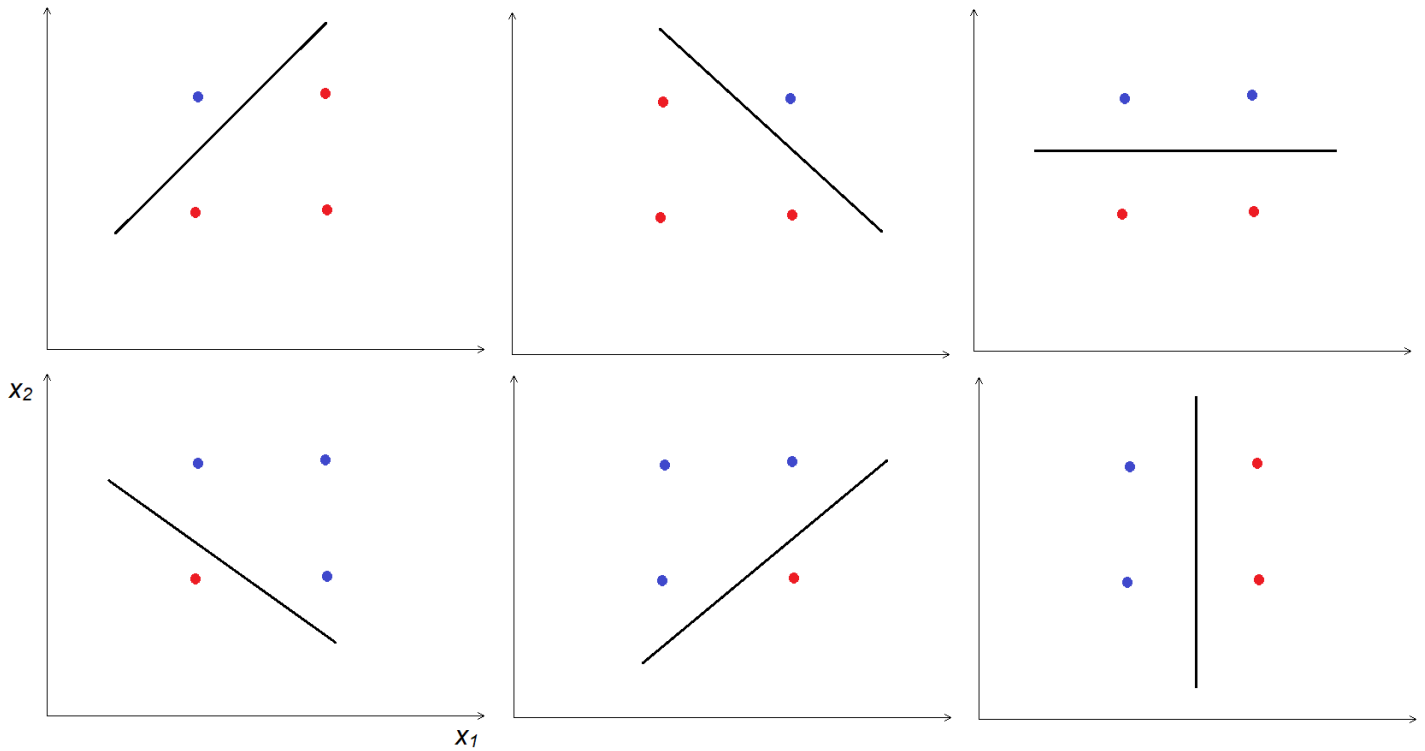
All possible labellings of 3 points by a linear classifier



Some possible labellings of 4 points by a linear classifier

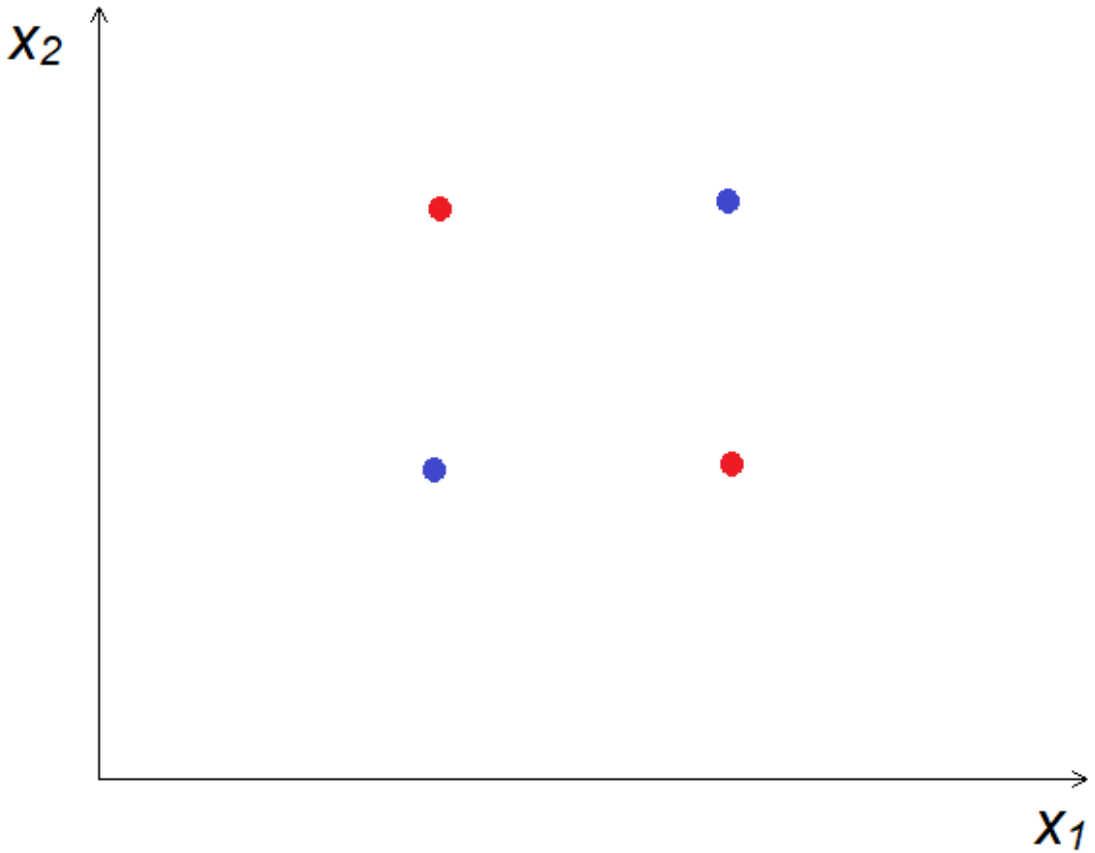


Some possible labellings of 4 points by a linear classifier

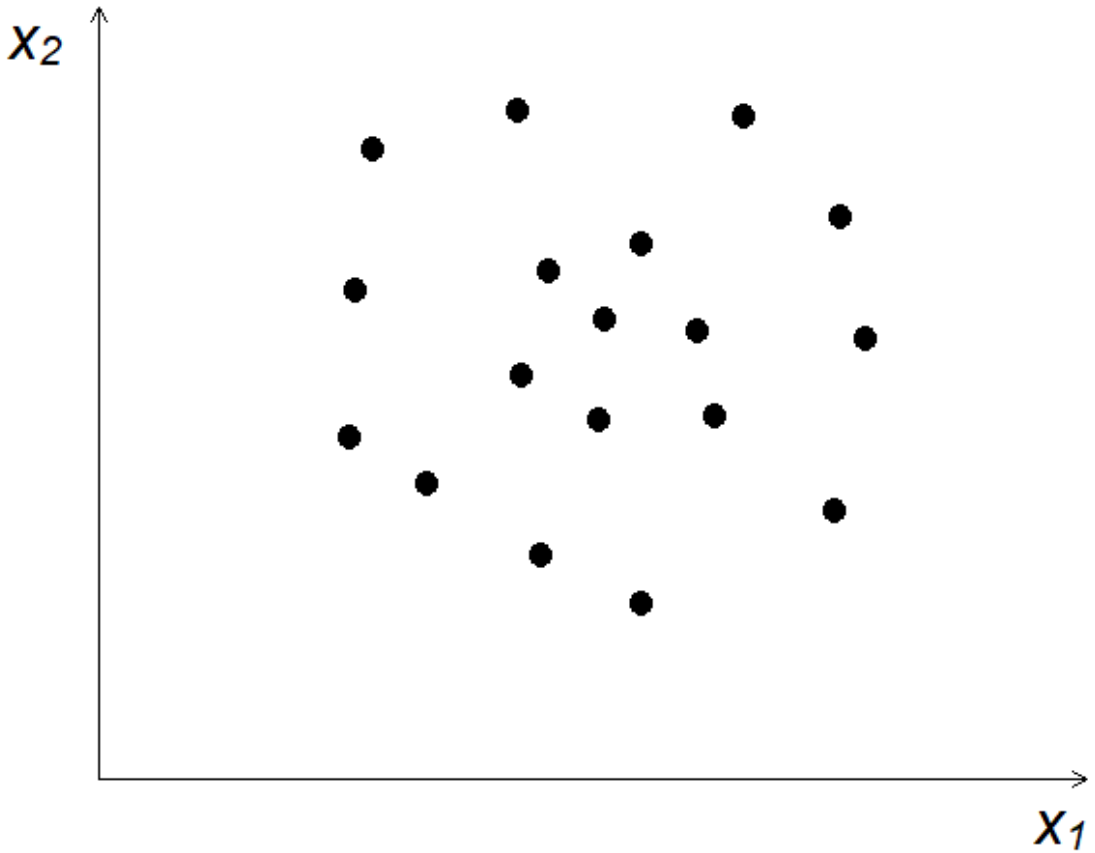


Etc...

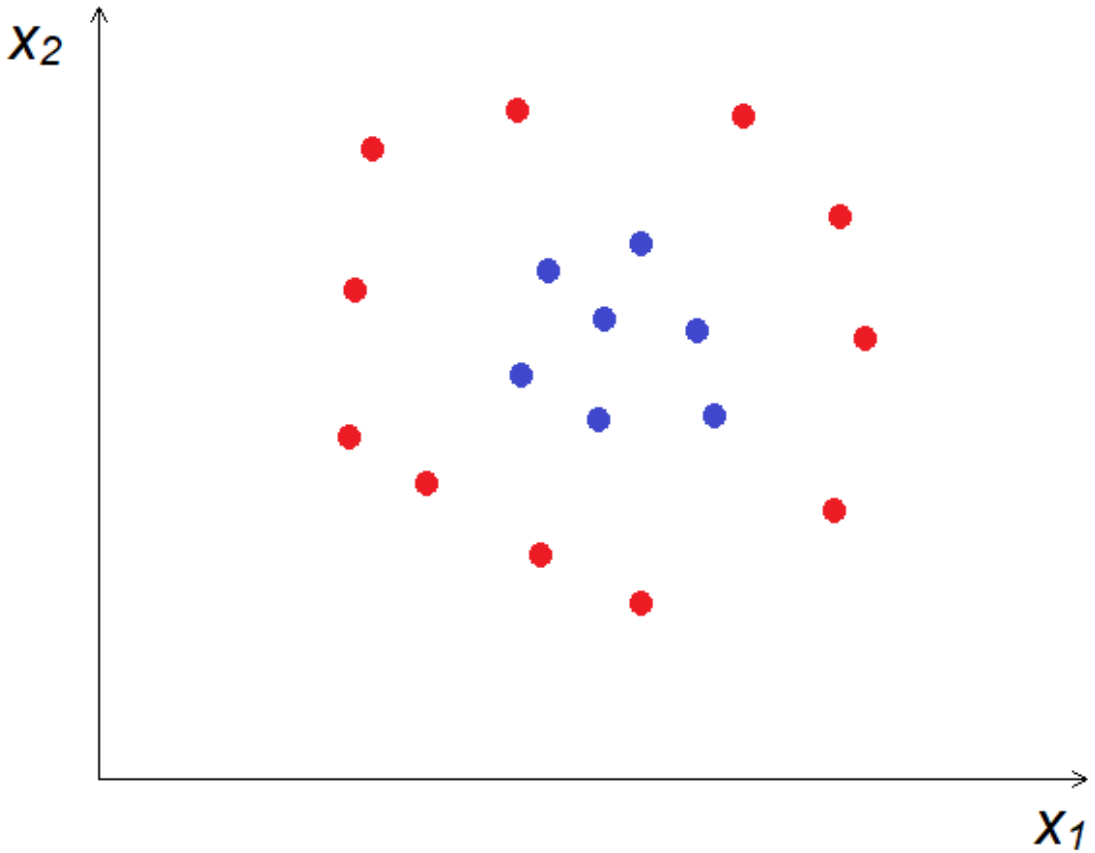
Infeasible labelling of 4 points for linear classifiers



Infeasible labelling for linear classifiers



Infeasible labelling using linear classifier



Discriminative capacity of linear classifiers

- A priori, with n sample points, there are 2^n possible ways of assigning some to class 0 and others to class 1
- Because of their rigidity, linear classifiers can only realise certain of these possible assignments

Number of possible splits of n sample points with a hyperplane

For n points in dimension p, it can be shown that there are at most

$$2\sum_{i=0}^{p}\binom{n-1}{i} \leq 2(n-1)^p + 2$$

ways of splitting the sample in two using a single hyperplane.

Here, $\binom{n-1}{i} = \frac{(n-1)!}{i!(n-i-1)!}$ denotes the binomial coefficient.

- This is **typically much smaller than 2^n**

Model complexity and shattering

- In classification, the capacity of a model to arbitrarily assign labels 0 or 1 to sample points by fine-tuning its free parameters is a more sensible measure of its complexity.
- Intuitively, the higher this capacity, the more prone to overfitting the model is.
- When a model has the capacity, by simply fine-tuning its parameters, to assign any possible labelling $\{0,1\}^n$ to n sample points $(x_1,\ldots,x_n)$, it is said that

$$\text{the model shatters the points } (x_1,\ldots,x_n).$$

Such a model is very likely to overfit samples of n observations and to generalise extremely poorly.

Table of Contents

Shattering

Definition: Shattering

Let S be model, i.e. a set of classifiers from $\mathcal{X}$ into $\{0,1\}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ be n fixed points of the feature space.

It is said that the model S shatters the points $x_1, \ldots, x_n$ if for any $(y_1, \ldots, y_n) \in \{0,1\}^n$, there is a classifier $f \in S$ such that

$$(f(x_1), \ldots, f(x_n)) = (y_1, \ldots, y_n),$$

or equivalently if: $\text{Card}\Big\{ \big(f(x_1), \ldots, f(x_n)\big) : \quad f \in S \Big\} = 2^n$.

Interpretation: Shattering

A model that shatters n given points is a model that is able to propose any arbitrary labelling to these specific points.

Shatter coefficients

Definition: Shatter coefficients

Let S be model, i.e. a set of classifiers from $\mathcal{X}$ into $\{0,1\}$. For any number of points n, we define the n^{th}-shatter coefficient of S (or growth function of S if seen as function of n) as the quantity:

$$\mathcal{C}(S,n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \text{Card}\Big\{ \underbrace{\big(f(x_1),\ldots,f(x_n)\big)}_{\in\{0,1\}^n} : \quad f \in S \Big\}$$

Interpretation: Shatter coefficients

- For a given n, $\mathcal{C}(S,n)$ is the maximum number of distinct labellings that the model S is able to produce on n points.
- It is a purely combinatorial quantity based on the flexibility of the model, distributions do not play any role in its definition.

Remarks on shattering

Remark

For a given n, there are at most 2^n different ways of assigning labels 0-1 to n points. Thus, it always holds that

$$\mathcal{C}(S, n) \leq 2^n.$$

Remark

Whenever there exist points $x_1, \ldots, x_n \in \mathcal{X}$ that are shattered by the model S, then $\mathcal{C}(S, n) = 2^n$.

Remark

Whenever $\mathcal{C}(S, n) = 2^n$, then for any $1 \leq m < n$ we must have $\mathcal{C}(S, m) = 2^m$

Remarks on shattering

Remark

All models are susceptible of shattering a small number n of points:

- In $\mathbb{R}^2$, linear classifiers can obviously shatter any(?) 3 points
- The model of step-functions over a partition $\mathcal{A} = (A_1, \ldots, A_m)$ can potentially shatter up to m points, provided they each fall in a different cell of $\mathcal{A}$

However, when n is too high, a given model may not have the ability to shatter n points anymore:

- In $\mathbb{R}^2$, linear classifiers cannot shatter 4 points or more
- The model of step-functions over a partition $\mathcal{A} = (A_1, \ldots, A_m)$ cannot shatter more than m points.

VC-dimension

Definition: VC-dimension of a model

A model S is said to be a class of Vapnik-Chervonenkis (VC-class) if there exists a maximum number of points n that S can shatter.

Formally, a model S is a VC-class if

$$V(S) := \sup\{n \geq 1 : \mathcal{C}(S, n) = 2^n\} < +\infty.$$

$V(S)$ is called the Vapnik-Chervonenkis dimension of S (or VC-dimension).

Remark

Similar to shatter coefficients, VC-dimension describes the flexibility of the model S only based on combinatoric considerations.

VC-dimension: perspectives for learning

VC-dimension of a model

The VC-dimension of a model will turn out to be a powerful measure of complexity of a model in classification.

We will provide several generalisation guarantees of sample-based predictors based on it.

Again, intuitively, a model with higher complexity is much more prone to overfitting and makes it also harder to find ("estimate") a good predictor within such model.

A lower VC-dimension means a lower model complexity and can be a very good thing for learning!

VC-dimension example (I): Intervals

Set of interval classifiers

Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and define the predictors:

$$f_{a,b} : x \in \mathbb{R} \mapsto \mathbb{1}_{x \in [a,b]}, \qquad \text{for any} \quad a, b \in \mathbb{R},\ a < b\,.$$

Then define the set of predictors

$$S_{\text{int}} := \left\{ f_{a,b} : \quad a, b \in \mathbb{R}\,, a < b \right\},$$

Proposition: VC-dim of interval classifiers

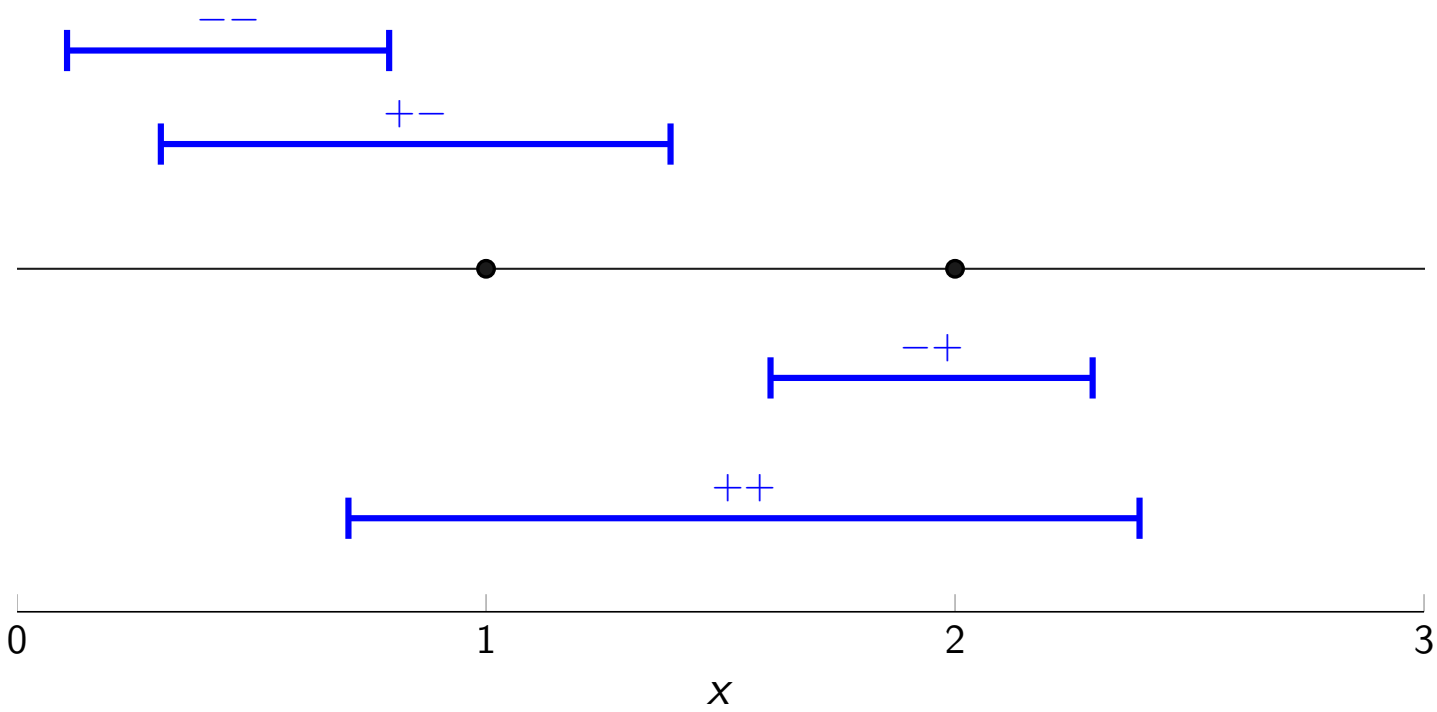
The model of interval classifiers defined above is a VC-class with

$$V(S_{\text{int}}) = 2$$

Proof (1/2).

First we prove $V(S_{\text{int}}) \geq 2$, by showing S_{int} shatters some pair $x_1, x_2 \in \mathbb{R}$.

For example take $x_1 = 1$ and $x_2 = 2$, then [verify yourself]



Proof (2/2).

Next, we will argue that $V(S_{\text{int}}) < 3$.

Take some $x_1, x_2, x_3 \in \mathbb{R}$ and without loss of generality assume $x_1 \leq x_2 \leq x_3$. Then the labeling $(1, 0, 1)$ cannot be obtained by an interval, so S_{int} does not shatter (x_1, x_2, x_3).

So we can conclude that $2 \leq V(S_{\text{int}}) < 3 \implies V(S_{\text{int}}) = 2$.

VC-dimension Example (II): Sine function

Models with a high VC-dimension (i.e. with high complexity), do not necessarily have a high number of parameters:

Model defined by a family of sine functions

Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and define predictor

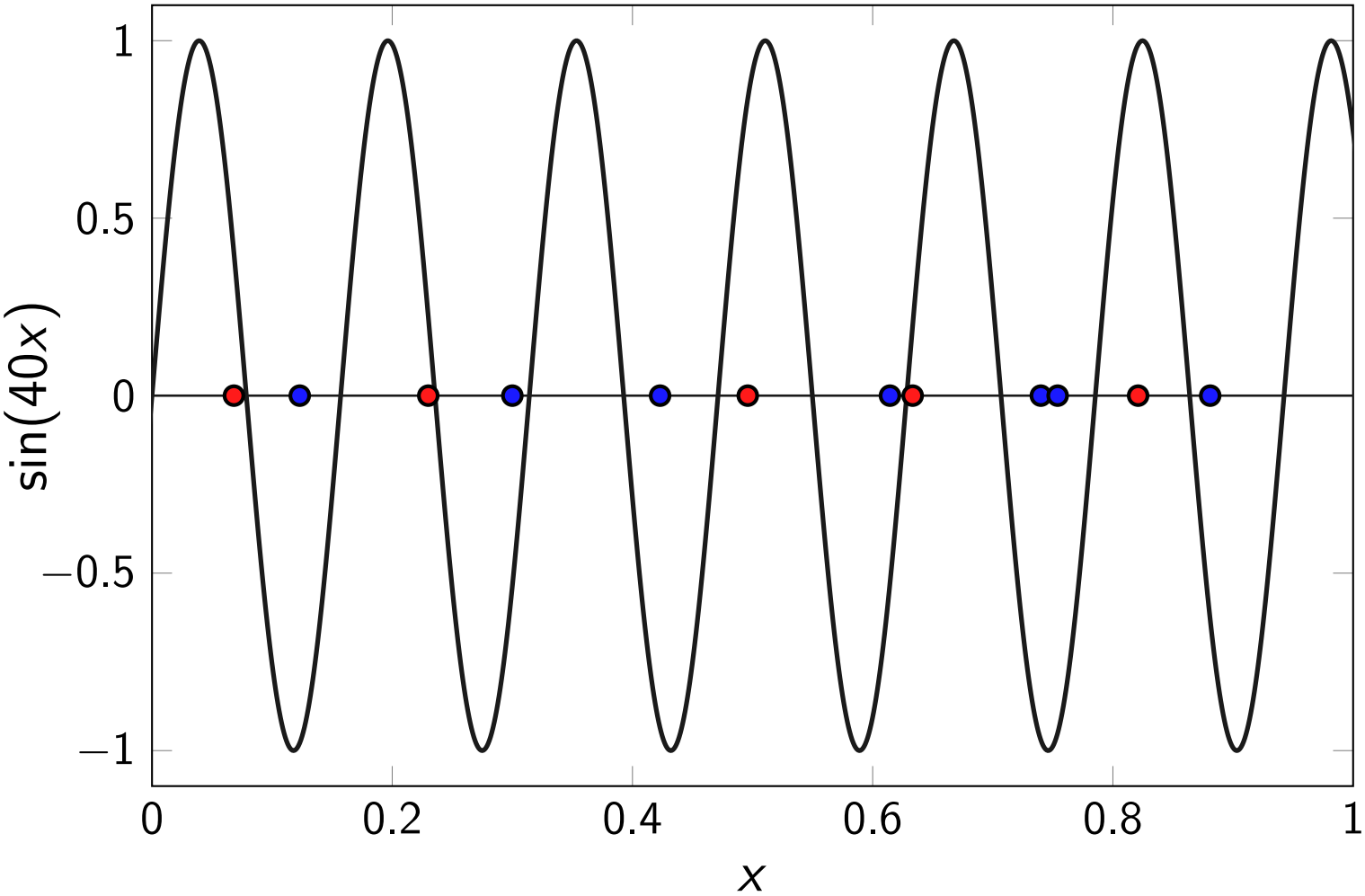
$$f_\omega : x \in \mathbb{R} \mapsto \mathbb{1}_{\sin(\omega x) > 0}, \qquad \text{for any} \quad \omega \in \mathbb{R}\,.$$

Then define the set of predictors

$$S_{\sin} := \Big\{ f_\omega : \ \ \omega \in \mathbb{R} \Big\},$$

This model just has a single parameter ω, but its VC-dimension is $V(S_{\sin}) = +\infty$.

VC-dimension Example (II): Sine function



VC-dimension Example (III): Finite models

Link to last week: consider finite models, i.e. models S with $\text{Card}(S) < +\infty$.

Exercise

Show that the VC dimension of a finite model S has the following upperbound:

$$V(S) \leq \log_2(\text{Card}(S))$$

Remark

Note: the VC-dim of a finite model S might be considerably lower than this upper bound

For instance consider the set of interval classifiers where a and b can only take values $\{1, 2, \ldots, k\}$ for some integer k.

Table of Contents

A preliminary inequality

Before providing the first learning guarantee based on the VC-dim., we need an inequality relating it to the shatter coefficients.

Lemma

Let S be a model with VC-dimension $V(S) = d$. Then, for any $n \geq d$:

$$\mathcal{C}(S, n) \leq \left(\frac{en}{d}\right)^d = O(n^d)$$

Interpretation

For a model S with finite VC-dimension, the number of possible labellings of n points grows at most polynomially with n.

This is much slower than an exponential increase, which we could have typically expected using the combinatorial approach.

Proof (1/2).

The proof uses Sauer's lemma:

Sauer's lemma

Let S be a model with VC-dimension $V(S) = d$. Then, for any $n \geq 1$:

$$\mathcal{C}(S, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

By Sauer's lemma, we have for $1 \leq d \leq n$:

$$\mathcal{C}(S, n) \overset{\text{Sauer's L.}}{\leq} \sum_{i=0}^{d} \binom{n}{i} \underbrace{\leq}_{\substack{\text{Multiplying} \\ \text{by terms} \geq 1}} \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \underbrace{\leq}_{\substack{\text{Adding} \\ \text{positive terms}}} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$

Proof (2/2).

Thus, using Newton's Binomial Theorem and the general inequality $1 + u \leq e^u$ for all $u \in \mathbb{R}$:

$$\begin{aligned} \mathcal{C}(S, n) \leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} &= \left(\frac{n}{d}\right)^{d} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i} \\ &= \left(\frac{n}{d}\right)^{d} \left(1 + \frac{d}{n}\right)^{n} \\ &\leq \left(\frac{n}{d}\right)^{d} e^{d}. \end{aligned}$$

This finishes the proof.

Learning guarantee of ERM predictors based on VC-dimension

Theorem (VC-dimension generalisation bound)

- Let $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \{0, 1\}$, P any feature/output distribution.
- Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ an iid P-distributed sample.
- Let S be a model with VC-dimension d.
- With c the 0-1 cost, let $\hat{f}_S$ be a predictor minimising the empirical risk over the model S.

Then, for any $\delta > 0$, $n \geq d$,

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + 4\sqrt{\frac{2\ln(\frac{4}{\delta}) + 2d\ln\left(\frac{2en}{d}\right)}{n}}\right) \geq 1 - \delta.$$

Remark

Remark

- The interpretation is similar as in the finite model case:

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + C\sqrt{\frac{2\ln(\frac{2}{\delta}) + 2\ln\left(\mathrm{Card}S\right)}{n}}\right) \geq 1 - \delta.$$

- The main change lies in the term $\ln(\mathrm{Card}S)$ being replaced (up to some constants) by $d\ln(\frac{2en}{d})$ depending on the VC-dimension.

- We are now able to quantify the generalisation risk of the ERM predictor based on the sample size even for infinite models.

Proof.

The proof uses the result of the following lemma:

Lemma: bound in terms of shatter coefficient

Under the conditions of the theorem, for any $\varepsilon > 0$, $n \geq d$,

$$\mathbb{P}\left(\sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)| \geq \varepsilon\right) \leq 4\, \mathcal{C}(S, 2n) \exp\left(-\frac{\varepsilon^2 n}{8}\right),$$

The proof of this lemma is omitted (quite technical).

Combining this lemma with the global upperbound of the estimation error (week 4), and using that $\mathcal{C}(S, n) \leq (en/d)^d$ for $n \geq d$, gives the learning guarantee after some straightforward manipulations (verify yourself!).

Table of Contents

Linear classifiers

- Let us return to the example of linear classifiers from $\mathcal{X} = \mathbb{R}^p$ into $\{0, 1\}$, i.e. to the model $S_{\text{lin}}^{\text{class}}$ composed of all functions of the form

$$f_{\boldsymbol{w},b}^{\text{class}} : x \in \mathbb{R}^p \mapsto \mathbb{1}_{\boldsymbol{w}^\top x + b \geq 0}.$$

- We will show in the coming slides that the VC-dimension of this model is $V(S_{\text{lin}}^{\text{class}}) = p + 1$.
- Let $\varepsilon > 0$ and $\delta > 0$. How large should the sample size n be to guarantee

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + \varepsilon\right) \geq 1 - \delta?$$

- By the theorem, it follows that n must then be such that:

$$4\sqrt{\frac{2\ln\left(\frac{4}{\delta}\right) + 2(p+1)\ln\left(\frac{2en}{p+1}\right)}{n}} \leq \varepsilon$$

Estimation error learning guarantees based on VC-dim

ε-accuracy, with confidence $1 - \delta = 99\%$

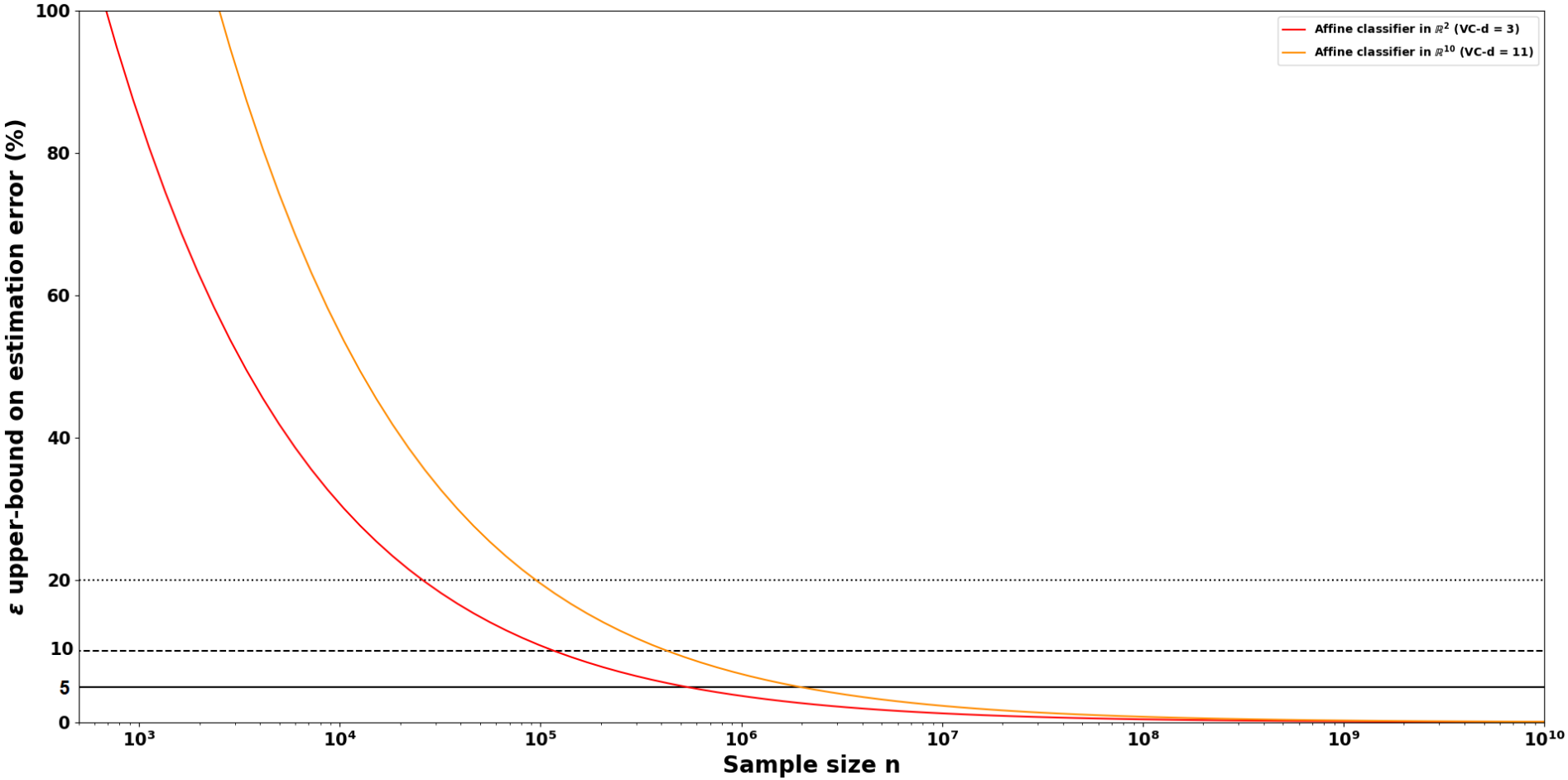


Table of Contents

Boundaries defined by a set of basis functions

- Linear classifiers separate the feature space using hyperplanes.
- A natural extension is to consider classifiers which can separate the space according to more complex boundaries
- More specifically, we can transform the features using m **basis functions** and then define a linear classifier based on this transformed feature vector.

Model defined by a set of basis functions

Let $\mathcal{X} = \mathbb{R}^p$ be a feature space and let $\psi_1, \ldots, \psi_m$ be $m \geq 1$ functions from $\mathbb{R}^p$ into $\mathbb{R}$. The set of predictors

$$S_\psi = \left\{ f : \boldsymbol{x} \mapsto \mathbb{1}_{a_1\psi_1(\boldsymbol{x})+\ldots+a_m\psi_m(\boldsymbol{x})>0} : \quad a_1, \ldots, a_m \in \mathbb{R} \right\},$$

is the model of classifiers with basis functions $\psi_1, \ldots, \psi_m$.

Common examples of basis functions

- The basis functions $\psi_i(\boldsymbol{x})$ are assumed to be fixed (i.e. not learned from the data).
- Common examples:
  - The projection on the input components for $\mathcal{X} = \mathbb{R}^p$: $m = p+1$ and

$$\psi_i(\boldsymbol{x}) = x_i \quad \text{for} \quad i = 1, \ldots, p \quad \text{and} \quad \psi_{p+1}(\boldsymbol{x}) = 1$$

    which brings us back to the model of linear classifiers.
  - i-power map for $\mathcal{X} = \mathbb{R}$:

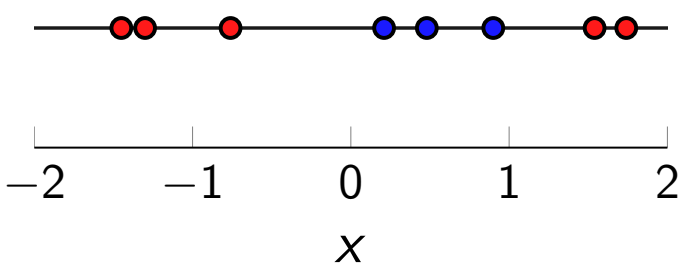
$$\psi_i(x) = x^i \text{ for} \quad i = 1, \ldots, m$$

  - Gaussian basis functions:

$$\psi_i(\boldsymbol{x}) = \exp\left(-(\boldsymbol{x} - \mu_i)^\top \Sigma_i^{-1} (\boldsymbol{x} - \mu_i)\right) \text{ for} \quad i = 1, \ldots, m$$

Simple Example of basis function classifier

- Consider $\mathcal{X} = \mathbb{R}$, and say we have the following sample:

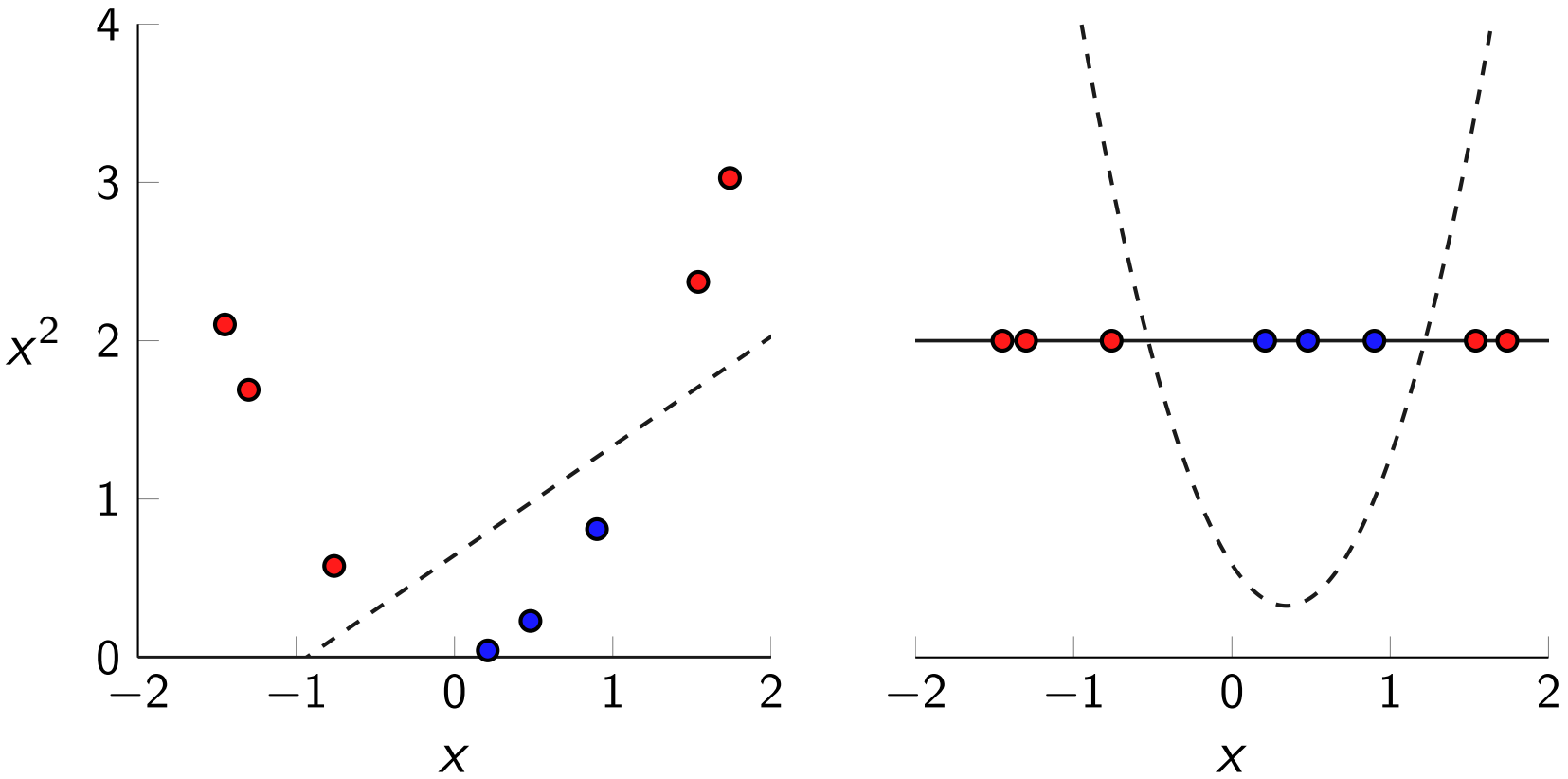


- The model of linear classifiers of the form $f_{a,b} : x \mapsto \mathbb{1}_{ax+b>0}$, for any $a, b \in \mathbb{R}$, does not contain a predictor that correctly predicts the labels of these points.
- Consider instead a basis function classifier with $m = 3$, and i-power map basis functions:

$$\psi_1(x) = 1, \quad \psi_2(x) = x \quad \text{and} \quad \psi_3(x) = x^2.$$

Simple Example of basis function classifier

Now we can separate the two classes:



Boundaries defined by a set of basis functions

Proposition: upperbound on VC-dimension of basis function classifiers

- Let $\mathcal{X} = \mathbb{R}^p$ be a feature space, and $\mathcal{Y} = \{0, 1\}$.
- Let $\psi_1, \ldots, \psi_m$ be m functions from $\mathbb{R}^p$ into $\mathbb{R}$, and S_ψ be the associated basis function classification model.
- Denote by Ψ the vector space spanned by the basis functions, i.e. $\Psi = \{a_1\psi_1 + \ldots + a_m\psi_m : \ a_1, \ldots, a_m \in \mathbb{R}\}$, and let $r = \dim(\Psi)$ (i.e. $r = \#$ linearly independent functions).

Then, the VC-dimension of model S_ψ satisfies:

$$V(S_\psi) \leq r.$$

Proof (1/2). [Optional reading]

It is sufficient to show that the model S_ψ cannot shatter sets of $k = r + 1$ points.

Fix k points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^p$ and consider the linear mapping $L : \Psi \mapsto \mathbb{R}^k$:

$$L(\psi) = \big(\psi(\boldsymbol{x}_1), \ldots, \psi(\boldsymbol{x}_k)\big)'.$$

The image of Ψ by L, $L(\Psi) \subset \mathbb{R}^k$, is a vector space of dimension smaller than $\min\{r, k\} = r = k - 1$. Thus, this image is necessarily contained in a certain hyperplane of $\mathbb{R}^k$. Let $(\gamma_1, \ldots, \gamma_k) \in \mathbb{R}^k$ be a non-zero vector orthogonal to this hyperplane.

Then, for any $\psi \in \Psi$:

$$\sum_{i=1}^{k} \gamma_i \psi(\boldsymbol{x}_i) = 0.$$

Proof (2/2). [Optional reading]

Because the vector $(\gamma_1, \ldots, \gamma_k)$ is non-zero, one can assume without loss of generality that some coordinates are positive and write:

$$\sum_{i:\gamma_i>0} \gamma_i \psi(\boldsymbol{x}_i) = \sum_{i:\gamma_i \leq 0} -\gamma_i \psi(\boldsymbol{x}_i).$$

We conclude with an argument by contradiction: if the model S_ψ **did** shatter the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, then by definition, there would be a function $\psi \in \Psi$ such that:

$$\text{for all } i \text{ such that } \gamma_i > 0, \quad \mathbb{1}_{\psi(\boldsymbol{x}_i)>0} = 1 \iff \psi(\boldsymbol{x}_i) > 0,$$

$$\text{for all } i \text{ such that } \gamma_i \leq 0, \quad \mathbb{1}_{\psi(\boldsymbol{x}_i)>0} = 0 \iff \psi(\boldsymbol{x}_i) \leq 0.$$

The left-hand side of the equation at the top then would be positive while the right-hand side would be non-positive: contradiction.

Therefore, S_ψ cannot shatter $k = r + 1$ points.

Example: Classifier with circular boundaries

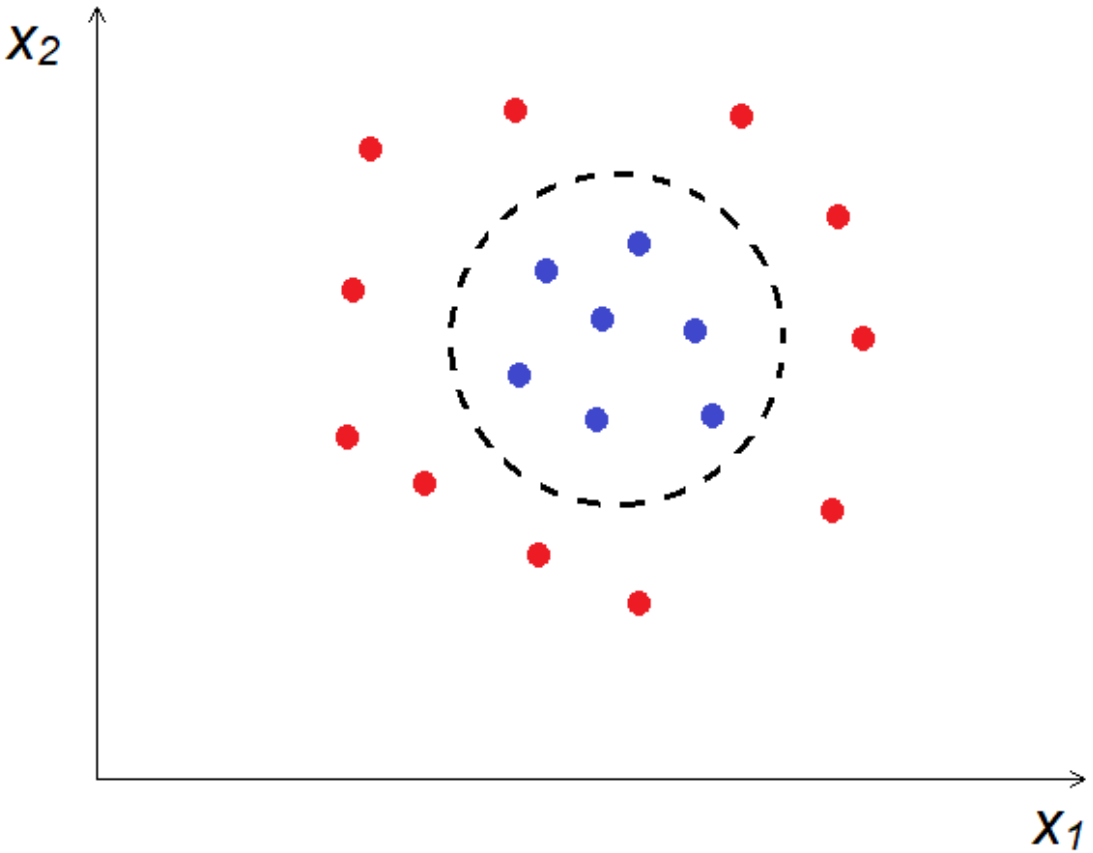
Classifier with circular boundary

A classifier from $\mathbb{R}^p$ into $\{0, 1\}$ with circular boundary is of the form:

$$f : \boldsymbol{x} = (x_1, \ldots, x_p) \mapsto \mathbb{1}_{\psi(\boldsymbol{x})>0}, \text{ with } \psi(\boldsymbol{x}) = a \sum_{j=1}^{p} (x_j - c_j)^2 - ab,$$

with $a, c_1, \ldots, c_p \in \mathbb{R}$, $b > 0$.

Circular boundary classifier in dimension 2



Example: Classifiers with circular boundaries

Classifier with circular boundary

- The set of classifier from $\mathbb{R}^p$ into $\{0, 1\}$ with circular boundaries can be rewritten as a linear combination of appropriate basis functions:

$$S_{\text{circ}} = \{\boldsymbol{x} \mapsto \mathbb{1}_{a_0\psi_0(\boldsymbol{x})+\ldots+a_{p+1}\psi_{p+1}(\boldsymbol{x})>0} : \quad a_0, \ldots, a_{p+1} \in \mathbb{R}\},$$

with
$\psi_0(\boldsymbol{x}) = 1$, $\psi_j(\boldsymbol{x}) = x_j$ for $j = 1, \ldots, p$, $\psi_{p+1}(\boldsymbol{x}) = \sum_{j=1}^{p} x_j^2$.

- Since the dimension of the space spanned by the basis functions is $p + 2$, we have:

$$V(S_{\text{circ}}) \leq p + 2.$$

Example: Classifiers with quadric boundaries

Classifier with quadric boundary

- The set of classifier from $\mathbb{R}^p$ into $\{0,1\}$ with quadric boundaries can be expressed as

$$S_{\text{conic}} = \left\{ \boldsymbol{x} \mapsto \mathbb{1}_{a_0 + \sum_{j=1}^{p} a_j x_j + \sum_{1 \leq i \leq j \leq p} b_{ij} x_i x_j > 0} : \quad a_j, b_{i,j} \in \mathbb{R} \right\}.$$

- In this case:

$$V(S_{\text{conic}}) \leq \frac{p(p+1)}{2} + p + 1.$$

Illustration: parabola boundary classifier

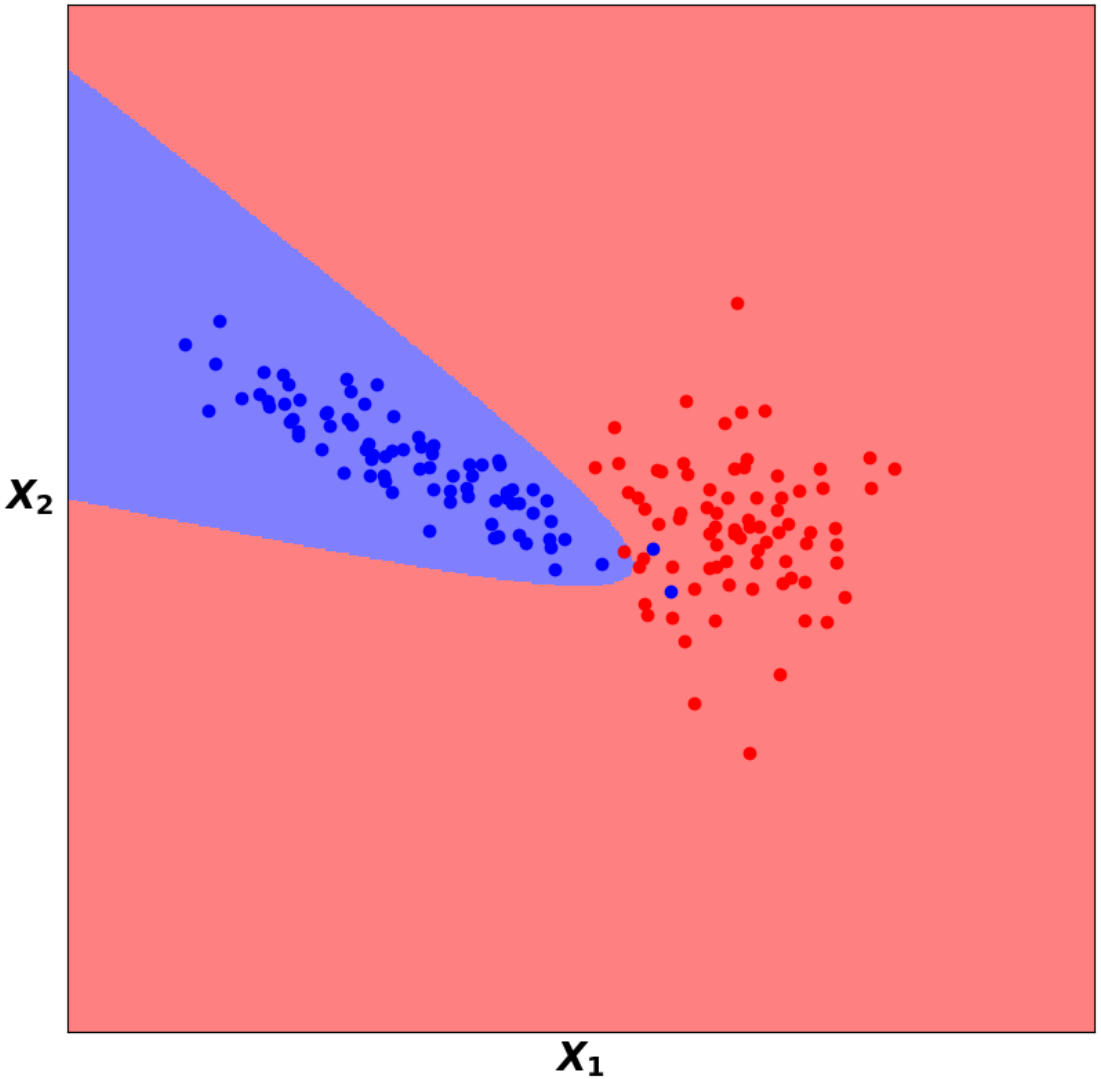
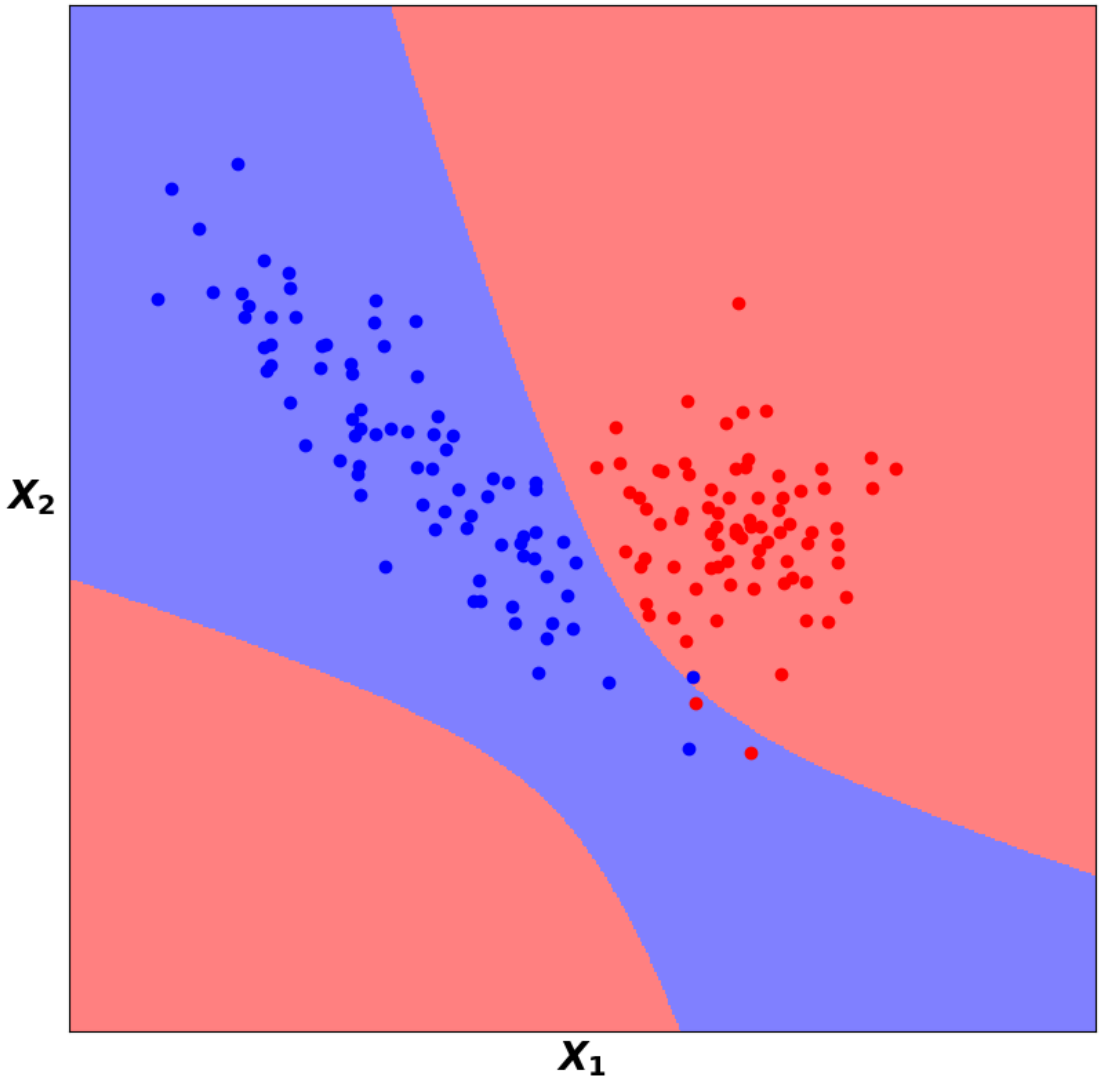
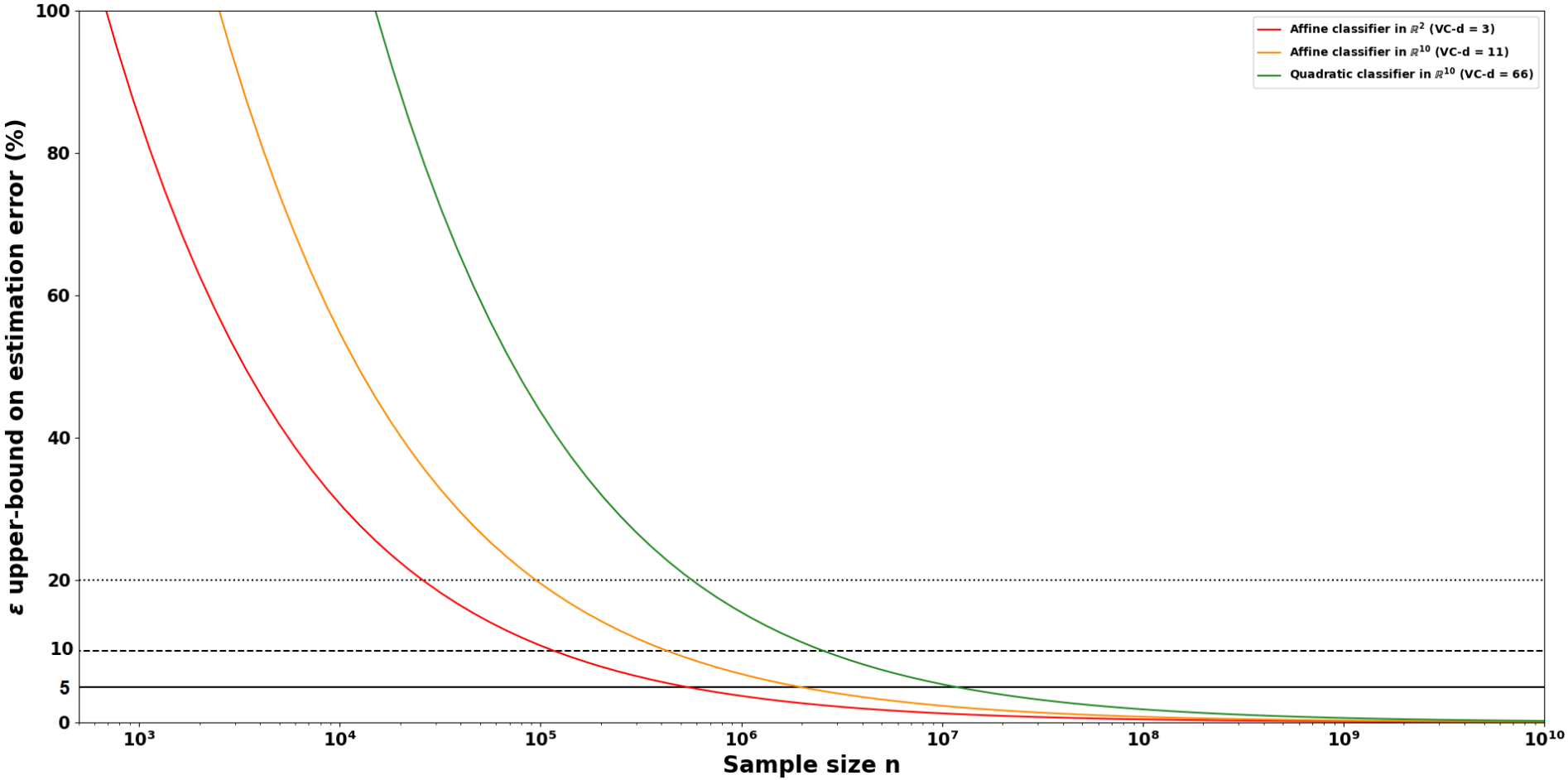


Illustration: hyperbola boundary classifier



Estimation error learning guarantees based on VC-dim

ε-accuracy, with confidence $1 - \delta = 99\%$



VC-dimension of linear classifiers

Proposition: VC-dimension of linear classifiers

- The set of linear classifiers from $\mathbb{R}^p$ into $\{0,1\}$:

$$S_{\text{lin}} = \left\{ \boldsymbol{x} \mapsto \mathbb{1}_{w_0 + \sum_{j=1}^{p} w_j x_j > 0} \; : \quad w_0, \ldots, w_p \in \mathbb{R} \right\}.$$

  is of VC-dimension:

$$V(S_{\text{lin}}) = p + 1.$$

Proof (1/4).

The model $S_{\text{lin}}^{\text{class}}$ is the model of classifiers with basis functions $\psi_0 : \boldsymbol{x} \mapsto 1$, $\psi_j : \boldsymbol{x} \mapsto x_j$, for all $j = 1, \ldots, p$.

Since the vector space $\Psi = \{a_0\psi_0 + \ldots + a_p\psi_p : \quad a_0, \ldots, a_p \in \mathbb{R}\}$ is of dimension $p+1$, we know from the previous proposition that

$$V(S_{\text{lin}}^{\text{class}}) \leq p + 1.$$

Let us now show that $V(S_{\text{lin}}^{\text{class}}) \geq p + 1$. It suffices to find $p+1$ points in $\mathbb{R}^p$ which are shattered by the model $S_{\text{lin}}^{\text{class}}$.

Proof (2/4).

For any $j = 1, \ldots, p$, let $X_j \in \mathbb{R}^p$ be the vector with a 1 on the j^{th} component and zeroes elsewhere [$(X_1, \ldots, X_p)$ is thus the canonical basis of $\mathbb{R}^p$] and let in addition $X_0 = 0 \in \mathbb{R}^p$.

We will show that $S_{\text{lin}}^{\text{class}}$ shatters the points $(X_0, \ldots, X_p)$, i.e. that for any labelling $(Y_0, \ldots, Y_p) \in \{0,1\}^{p+1}$, we can find $f \in S_{\text{lin}}^{\text{class}}$ such that:

$$\forall i \in \{0, \ldots, p+1\}, \quad f(X_i) = Y_i.$$

It is sufficient [check it] to find $w_0, \ldots, w_p \in \mathbb{R}$ such that

$$\forall i \in \{0, \ldots, p+1\}, \quad w_0 + \sum_{j=1}^{p} w_j X_{i,j} = 2Y_i - 1,$$

for any $(Y_0, \ldots, Y_p) \in \{0,1\}^{p+1}$, where $X_i = (X_{i,1}, \ldots, X_{i,p})$.

Proof (3/4).

The latter system of equations is equivalent to

$$\boldsymbol{X}\boldsymbol{w} = 2\boldsymbol{Y} - 1,$$

$$\text{with } \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 1 & 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}, 1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{p+1},$$

$\boldsymbol{Y}^\top = (Y_0, \ldots, Y_p)$, and $\boldsymbol{w}^\top = (w_0, \ldots, w_p)$.

Since $\boldsymbol{X}$ is lower-triangular with non-zero diagonal elements, it is invertible.

Proof (4/4).

Hence, for any labelling $(Y_0, \ldots, Y_p) \in \{0,1\}^{p+1}$ of the points $X_0, \ldots, X_p$, there is a predictor $f \in S_{\text{lin}}^{\text{class}}$, namely,

$$f_{\boldsymbol{w}} : x \mapsto \mathbb{1}_{\boldsymbol{w}^\top x > 0}$$

with $\boldsymbol{w} = \boldsymbol{X}^{-1}(2\boldsymbol{Y} - 1)$, such that

$$f_{\boldsymbol{w}}(X_i) = Y_i, \qquad \text{for all } i \in \{0, \ldots, p\}.$$

This shows that $S_{\text{lin}}^{\text{class}}$ shatters the $p+1$ points $X_0, \ldots, X_p$, and therefore:

$$V(S_{\text{lin}}^{\text{class}}) \geq p + 1,$$

which concludes the proof.

Table of Contents

Example: VC-dimension of Neural Network classifiers

VC-dimension of single layer Neural Networks

- Let $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \{0, 1\}$.
- Let S_k be the model of single layer neural networks from $\mathcal{X}$ into $\mathcal{Y}$, with k units in the hidden layer, and with threshold sigmoid activation functions, i.e. each $f \in S_k$ is such that:

$$f(x) = \sum_{j=1}^{k} \beta_j \mathbb{1}_{\left\{ \boldsymbol{w}_j^\top x + b_j > 0 \right\}}, \qquad \text{for } x \in \mathbb{R}^p$$

Then, for any $k \geq 1$, $n \geq 1$, we have:

$$\mathcal{C}(S_k, n) \leq (ne)^{kp+2k+1},$$

and the VC-dimension is lower- and upper-bounded as:

$$2\left\lfloor \frac{k}{2} \right\rfloor p \leq V(S_k) \leq (2kp + 4k + 2) \log_2 \big(e(kp + 2k + 1)\big).$$

For those interested: the proof is in Devroye et al. (1996), Chapter 30, Section 4, Theorems 30.5 and 30.6, p.540-542

Further results on Neural networks

VC-dimension of deep neural networks

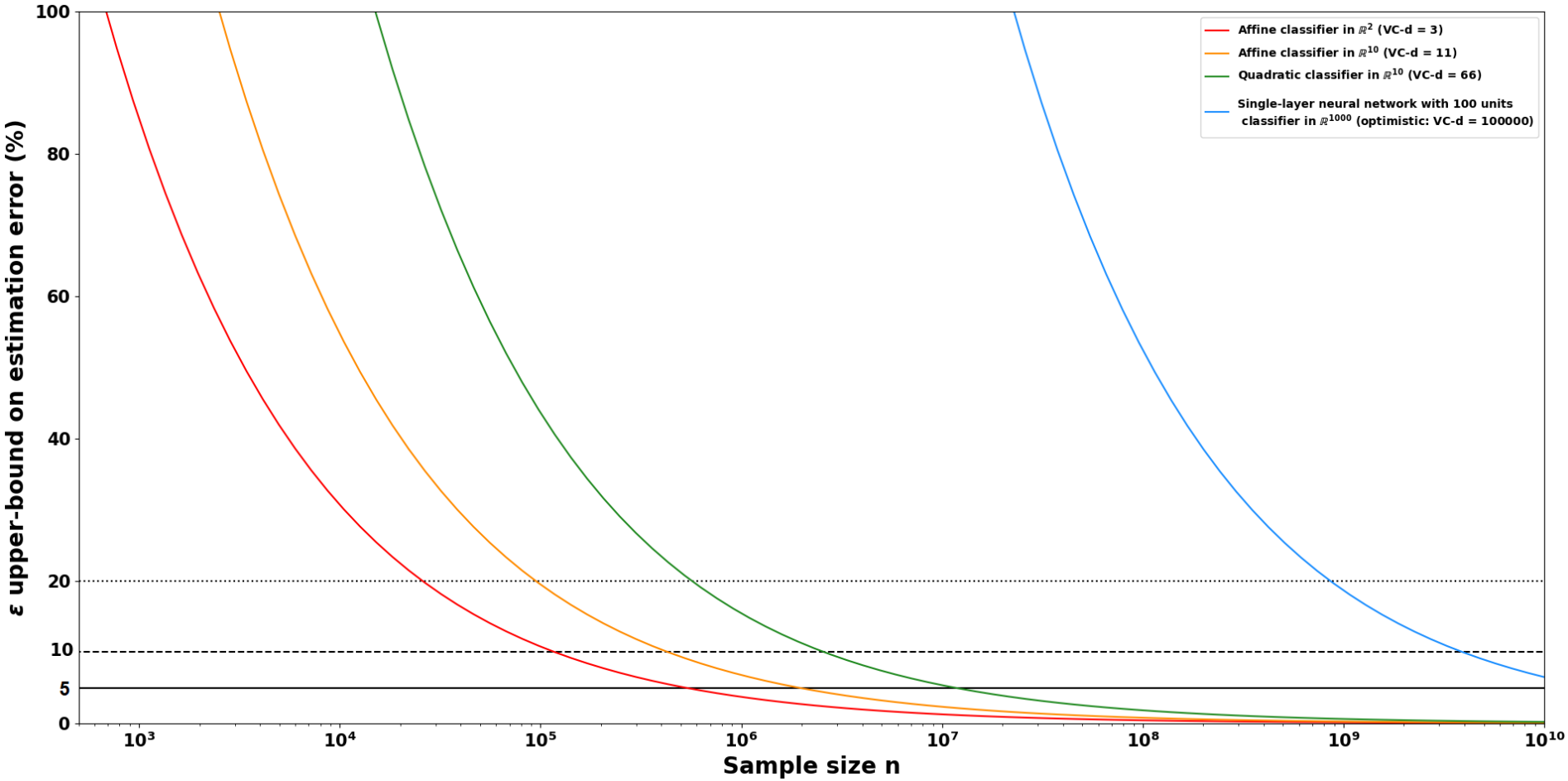
The VC-dimension of deep neural networks, i.e. with a large number of hidden layers and units, is still an ongoing research topic, see for instance:

Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks

by Bartlett, Harvey, Liaw, and Mehrabian
in *Journal of Machine Learning Research* 2019.

Estimation error learning guarantees based on VC-dim

ε-accuracy, with confidence $1 - \delta = 99\%$



Note: Recall that these learning guarantees hold *for any distribution P.*

Generalisation ability Neural Networks

Generalisation ability of deep Neural Networks

The paper

Understanding deep learning requires rethinking generalization

by Zhang, Bengio, Hardt, Recht, and Vinyals in 2017,

critiques traditional view of generalization by showing that measures as VC-dimension cannot distinguish between NNs with radically different generalization performance.

- NNs can perfectly fit any labels if big enough (i.e. can memorize training data)
- By experimenting they show how complexity measures as the VC-dimension fail to explain why some NNs generalize well in practice

Table of Contents

Learning guarantees for Empirical Convexified Risk Minimisation

- ERM for binary classification is computationally hard, and therefore usually not used in practice
- Recall from Week 3: a feasible alternative is Empirical Convexified Risk Minimisation (ECRM), where a convex surrogate loss function replaces the 0-1 cost
- There are learning guarantees for the ECRM in terms of the VC dimension $\rightarrow$ we will see one in this week's tutorial (based on Lugosi and Vayatis (2004))

Table of Contents

Extension to regression framework

- Everything we saw this week concerned binary classification framework
- The learning guarantee for infinite models that we saw does not apply to regression problems, so e.g. if $\mathcal{Y} = \mathbb{R}$
- However, the notion of VC-dimension can be extended to apply to the regression framework: the so-called pseudo-dimension
- This measure of complexity allows for the derivation of learning guarantees for infinite models in the regression framework (but those are beyond the scope of this course)

Pseudo-dimension [not exam material]

Definition: Shattering in regression framework

Let S be a model. Then a set $x_1, \ldots, x_n \subseteq \mathcal{X}$ is said to be *shattered* by S, if there exist $y_1, \ldots, y_n \in \mathbb{R}$ such that

$$\text{Card}\left\{ \left(\text{sgn}(f(x_1) - y_1), \ldots, \text{sgn}(f(x_n) - y_n)\right) : \ f \in S \right\} = 2^n \,,$$

where $\text{sgn}(x)$ denotes the 'sign' function.

Definition: Pseudo-dimension

Let S be a model. Then the pseudo-dimension of S, denoted by $\text{Pdim}(S)$, is the size of the largest set shattered by S. Equivalently:

$$\text{Pdim}(S) := \text{V}\left(\{(x, y) \mapsto \text{sgn}(f(x) - y) : \ f \in S\}\right)$$

Wrap-up

We saw how to quantify complexity of infinite models:

- The n-th shatter coefficient $\mathcal{C}(S, n)$
- The VC-dimension: highest n for which there exist $x_1, \ldots, x_n \in \mathcal{X}$ that the model shatters.

We saw a learning guarantee applicable to infinite models, based on the VC-dimension

We studied the VC dimension of some common models

Assignment

- Question 3 of Assignment Part II concerns the material of this week (and recall questions 1 and 2 concern week 4)
- I encourage you to already start working on it (deadline of Part II is March 18th at 23:59)