Statistical learning

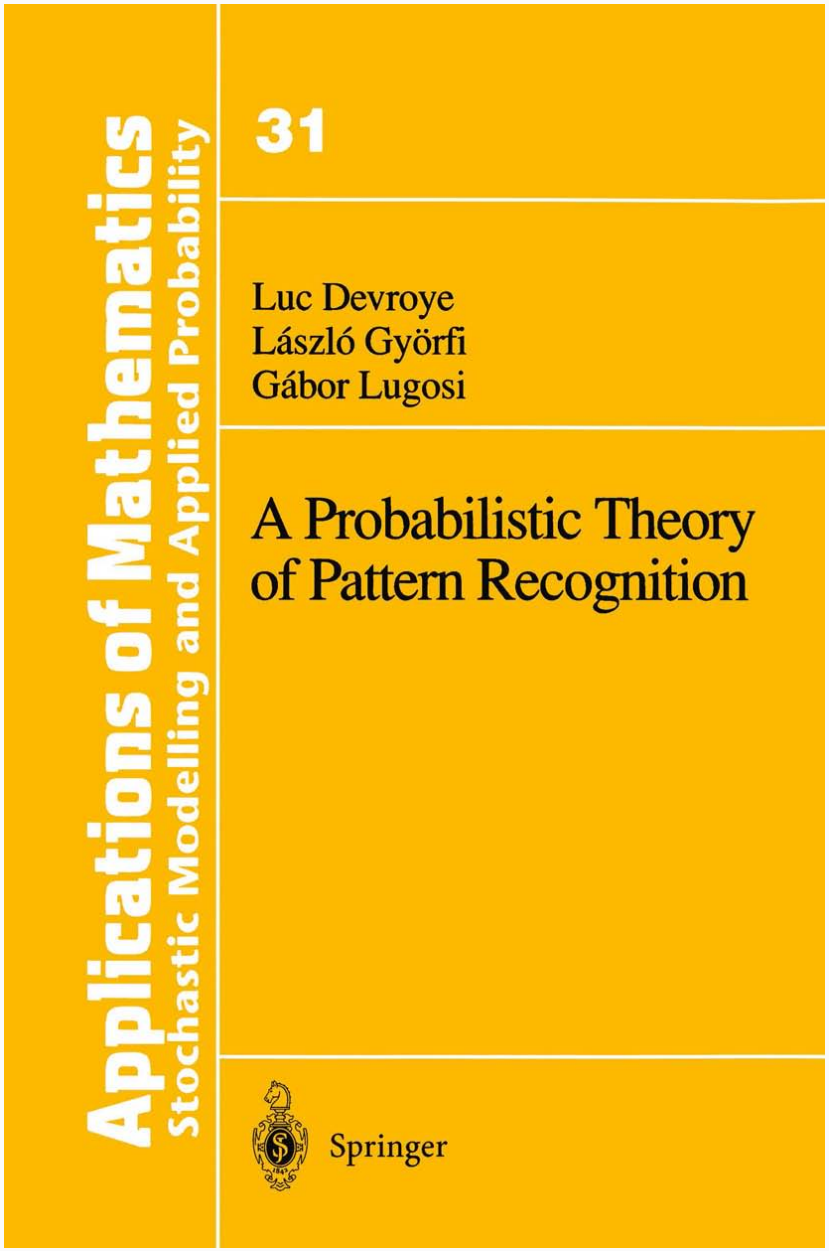
Gérard Biau

Semester 1

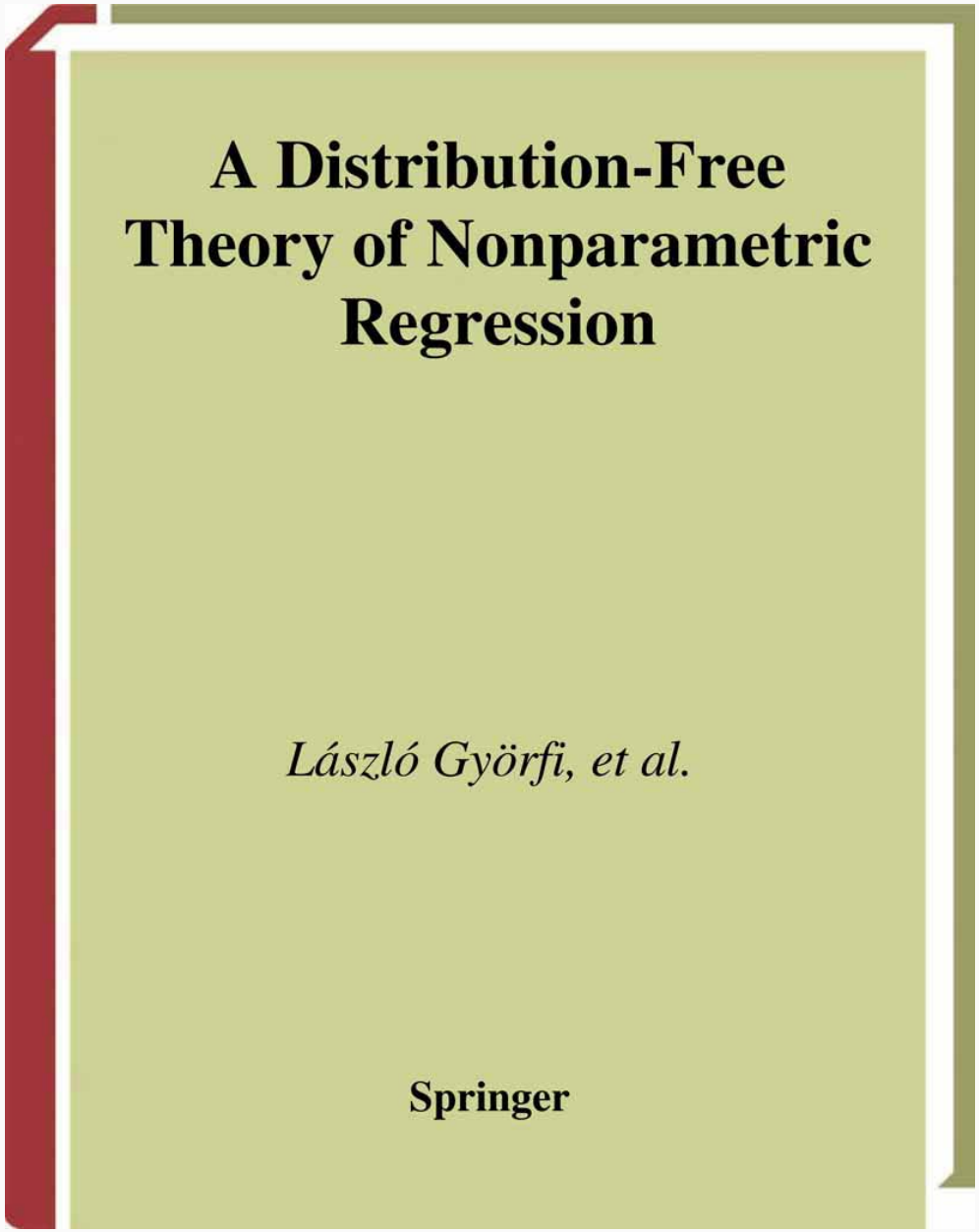
Overview

- Prerequisites: basics of probability and statistics.
- Objectives:
  - ▷ Learn and apply the fundamental concepts of statistical learning;
  - ▷ Understand the basic theory underlying data science, machine learning, and AI;
  - ▷ Be able to read current research books and papers.
- Keywords: classification, prediction, performance guarantees.
- Resources: slides and proof textbook.
- Teacher: Gérard Biau, gerard.biau@sorbonne-universite.fr .
- Office: ✉ 15-25, 2nd floor, ☎ 01 44 27 85 63.

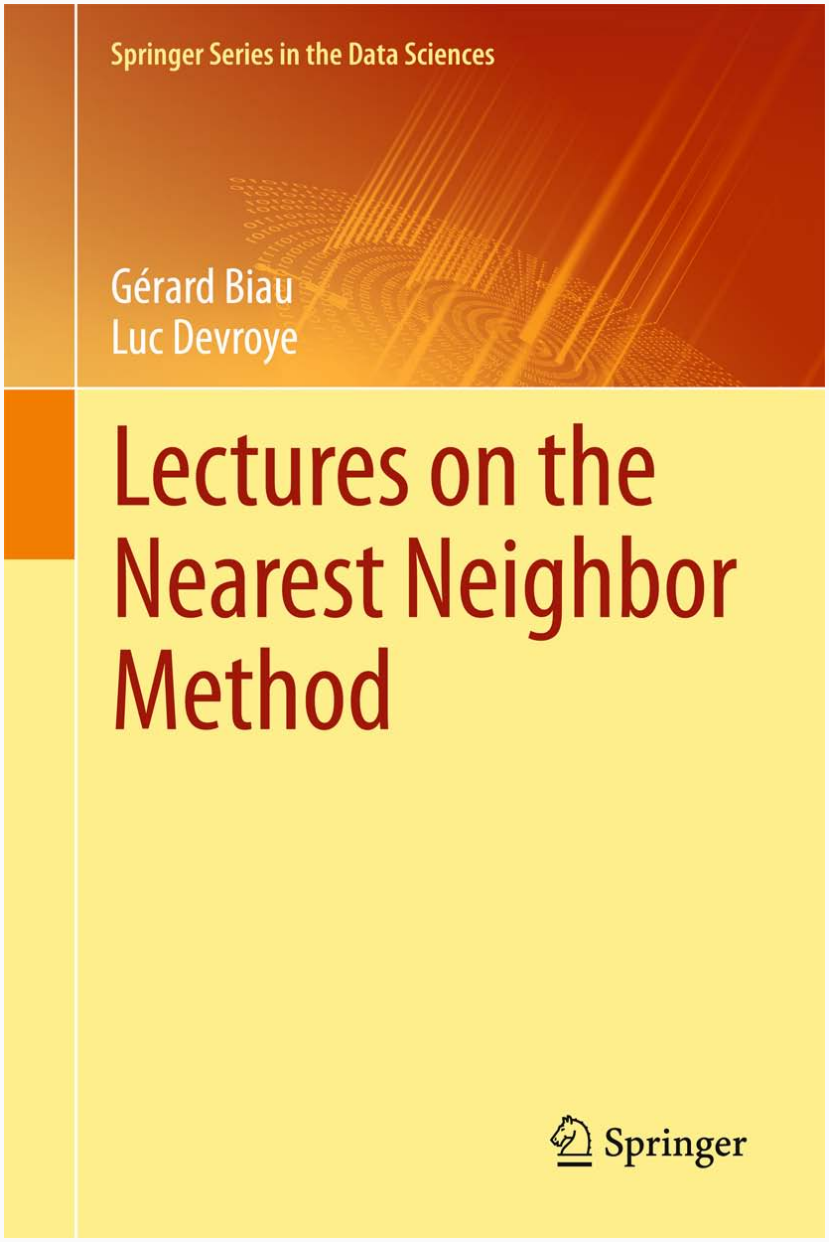
References


Applications of Mathematics
Stochastic Modelling and Applied Probability
31
Luc Devroye
László Györfi
Gábor Lugosi
A Probabilistic Theory of Pattern Recognition
Springer

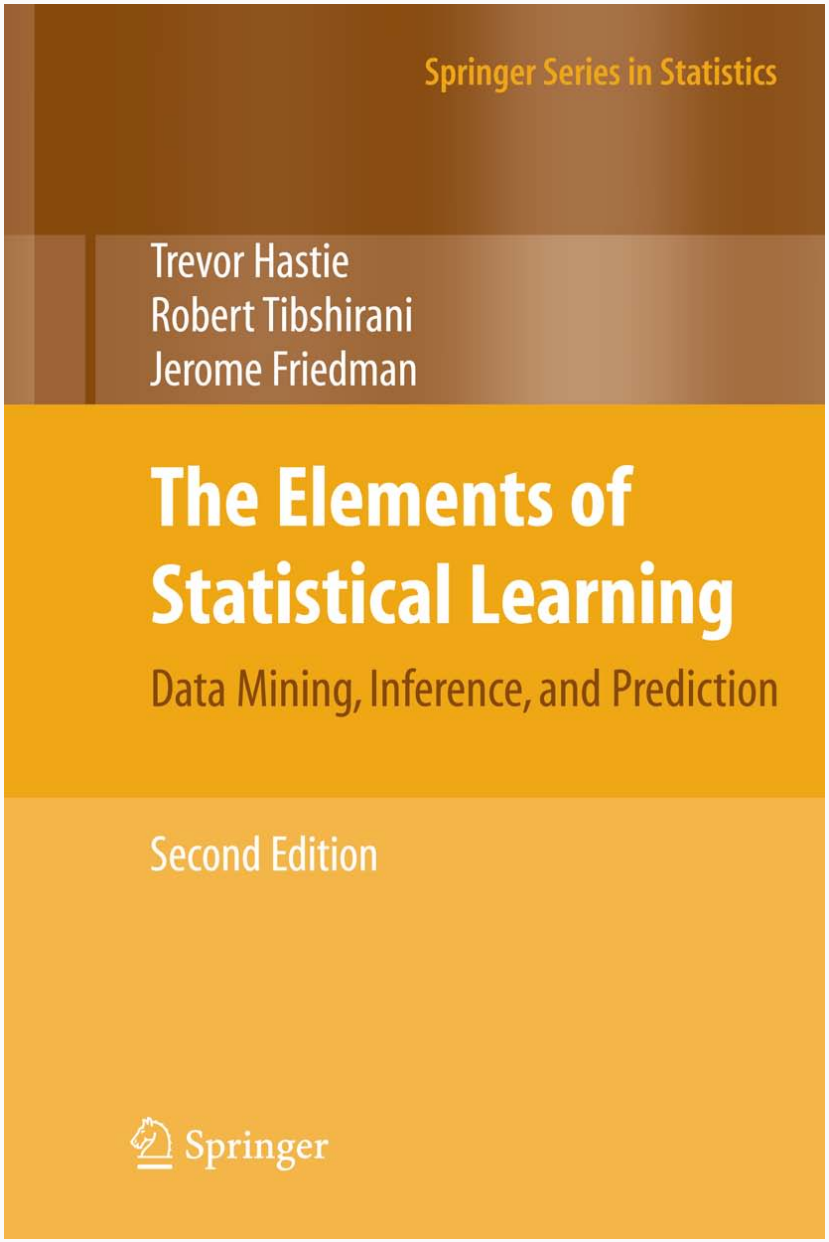
References


A Distribution-Free Theory of Nonparametric Regression
László Györfi, et al.
Springer

References


Springer Series in the Data Sciences
Gérard Biau
Luc Devroye
Lectures on the Nearest Neighbor Method
Springer

References


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

References

ESAIM: PS
November 2005, Vol. 9, p. 323–375
DOI: 10.1051/ps:2005018

ESAIM: Probability and Statistics

THEORY OF CLASSIFICATION: A SURVEY OF SOME RECENT ADVANCES *

Stéphane Boucheron[1], Olivier Bousquet[2] and Gábor Lugosi[3]

Abstract. The last few years have witnessed important new developments in the theory and practice of pattern classification. We intend to survey some of the main new ideas that have led to these recent results.

Mathematics Subject Classification. 62G08, 60E15, 68Q32.

Received June 18, 2004. Accepted September 12, 2005.

1. Introduction

The last few years have witnessed important new developments in the theory and practice of pattern classification. The introduction of new and effective techniques of handling high-dimensional problems – such as boosting and support vector machines – have revolutionized the practice of pattern recognition. At the same time, the better understanding of the application of empirical process theory and concentration inequalities have led to effective new ways of studying these methods and provided a statistical explanation for their success. These new tools have also helped develop new model selection methods that are at the heart of many classification algorithms.

The purpose of this survey is to offer an overview of some of these theoretical tools and give the main ideas of the analysis of some of the important algorithms. This survey does not attempt to be exhaustive. The selection of the topics is largely biased by the personal taste of the authors. We also limit ourselves to describing the key ideas in a simple way, often sacrificing generality. In these cases the reader is pointed to the references for the sharpest and more general results available. References and bibliographical remarks are given at the end of each section, in an attempt to avoid interruptions in the arguments.

2. Basic model

The problem of pattern classification is about guessing or predicting the unknown class of an observation. An observation is often a collection of numerical and/or categorical measurements represented by a d-dimensional vector x but in some cases it may even be a curve or an image. In our model we simply assume that $x \in \mathcal{X}$

Keywords and phrases. Pattern recognition, statistical learning theory, concentration inequalities, empirical processes, model selection.

* *The authors acknowledge support by the PASCAL Network of Excellence under EC grant no. 506778. The work of the third author was supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324.*

[1] Laboratoire Probabilités et Modèles Aléatoires, CNRS & Université Paris VII, Paris, France.

[2] Pertinence SA, 32 rue des Jeûneurs, 75002 Paris, France.

[3] Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain; lugosi@upf.es

References

1

Pattern classification and learning theory

Gábor Lugosi

1.1 A binary classification problem

Pattern recognition (or *classification* or *discrimination*) is about guessing or predicting the unknown class of an observation. An *observation* is a collection of numerical measurements, represented by a d-dimensional vector x. The unknown nature of the observation is called a *class*. It is denoted by y and takes values in the set $\{0,1\}$. (For simplicity, we restrict our attention to binary classification.) In pattern recognition, one creates a function $g(x) : \mathcal{R}^d \to \{0,1\}$ which represents one's guess of y given x. The mapping g is called a *classifier*. A classifier errs on x if $g(x) \neq y$.

To model the learning problem, we introduce a probabilistic setting, and let (X,Y) be an $\mathcal{R}^d \times \{0,1\}$-valued random pair.

The random pair (X,Y) may be described in a variety of ways: for example, it is defined by the pair (μ,η), where μ is the probability measure for X and η is the regression of Y on X. More precisely, for a Borel-measurable set $A \subseteq \mathcal{R}^d$,

$$\mu(A) = \mathbb{P}\{X \in A\},$$

and for any $x \in \mathcal{R}^d$,

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\} = \mathbb{E}\{Y | X = x\}.$$

Thus, $\eta(x)$ is the conditional probability that Y is 1 given $X = x$. The distribution of (X,Y) is determined by (μ,η). The function η is called the *a posteriori probability*.

Any function $g : \mathcal{R}^d \to \{0,1\}$ defines a *classifier*. An *error* occurs if $g(X) \neq Y$, and the *probability of error* for a classifier g is

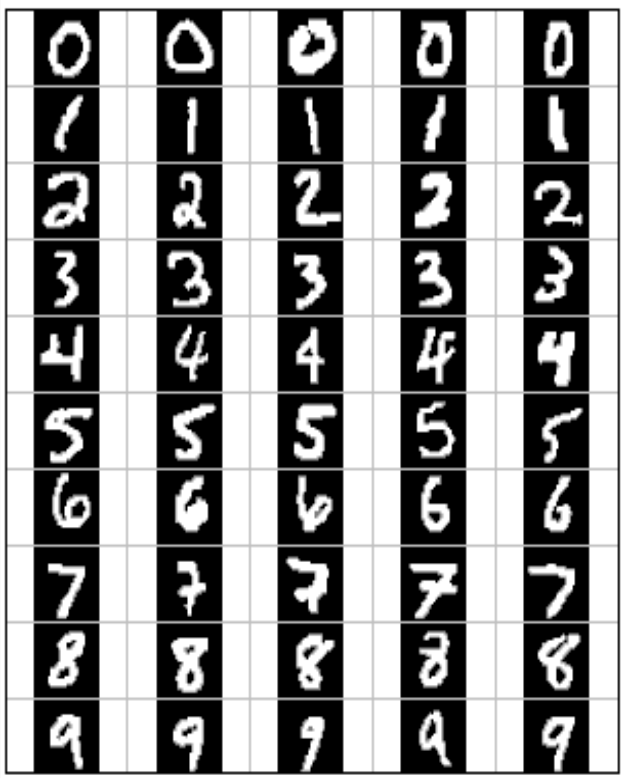
$$L(g) = \mathbb{P}\{g(X) \neq Y\}\ .$$

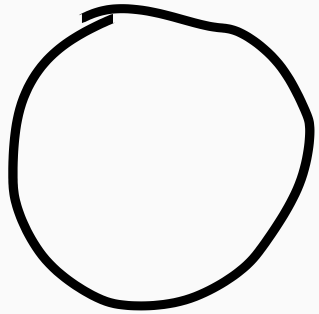
The Bayes classifier given by

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Motivating examples

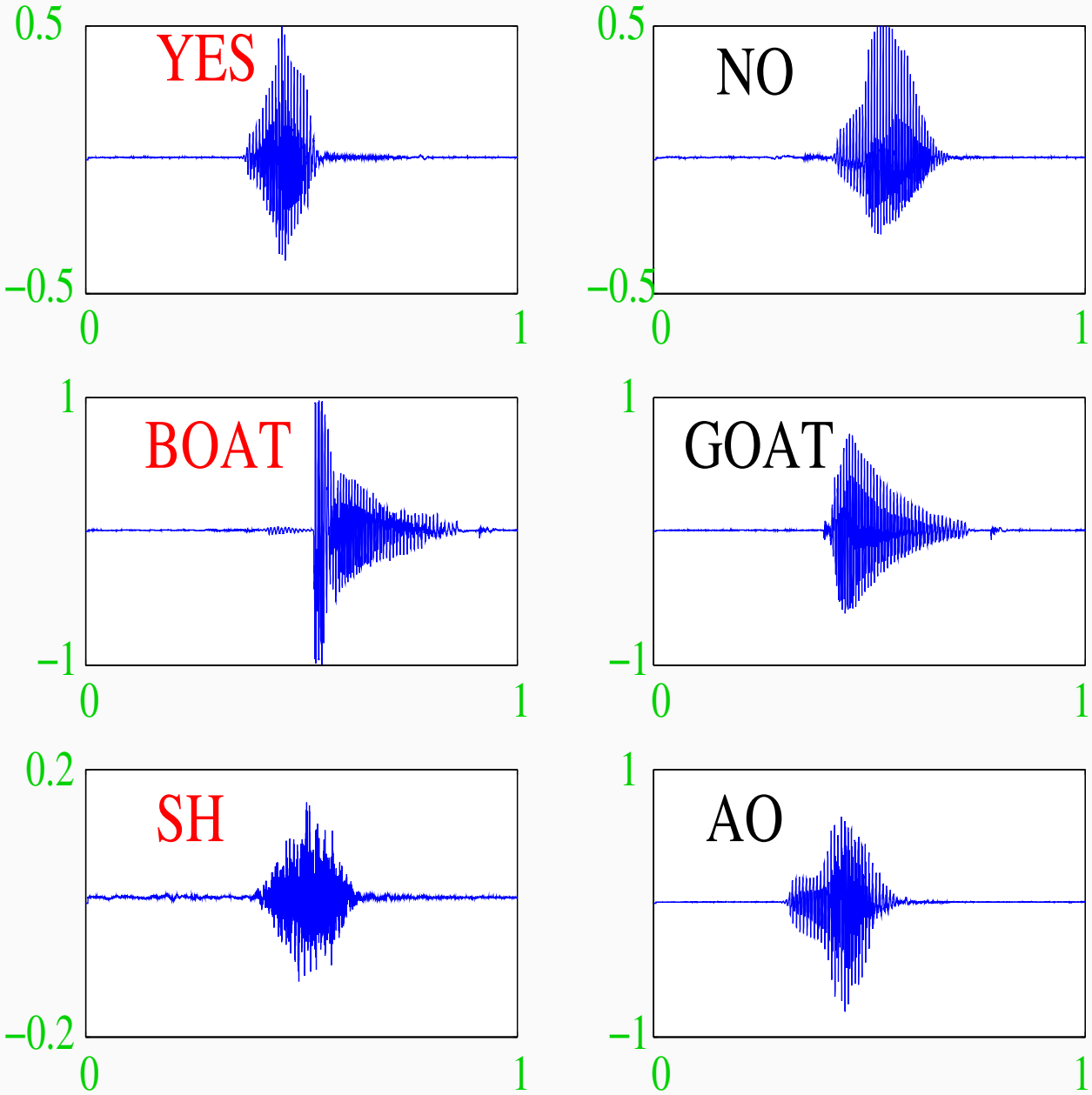
Handwritten digit recognition





Which digit is this? 0, 1, 2,...?

Voice recognition



Weather forecast

Temperature prediction

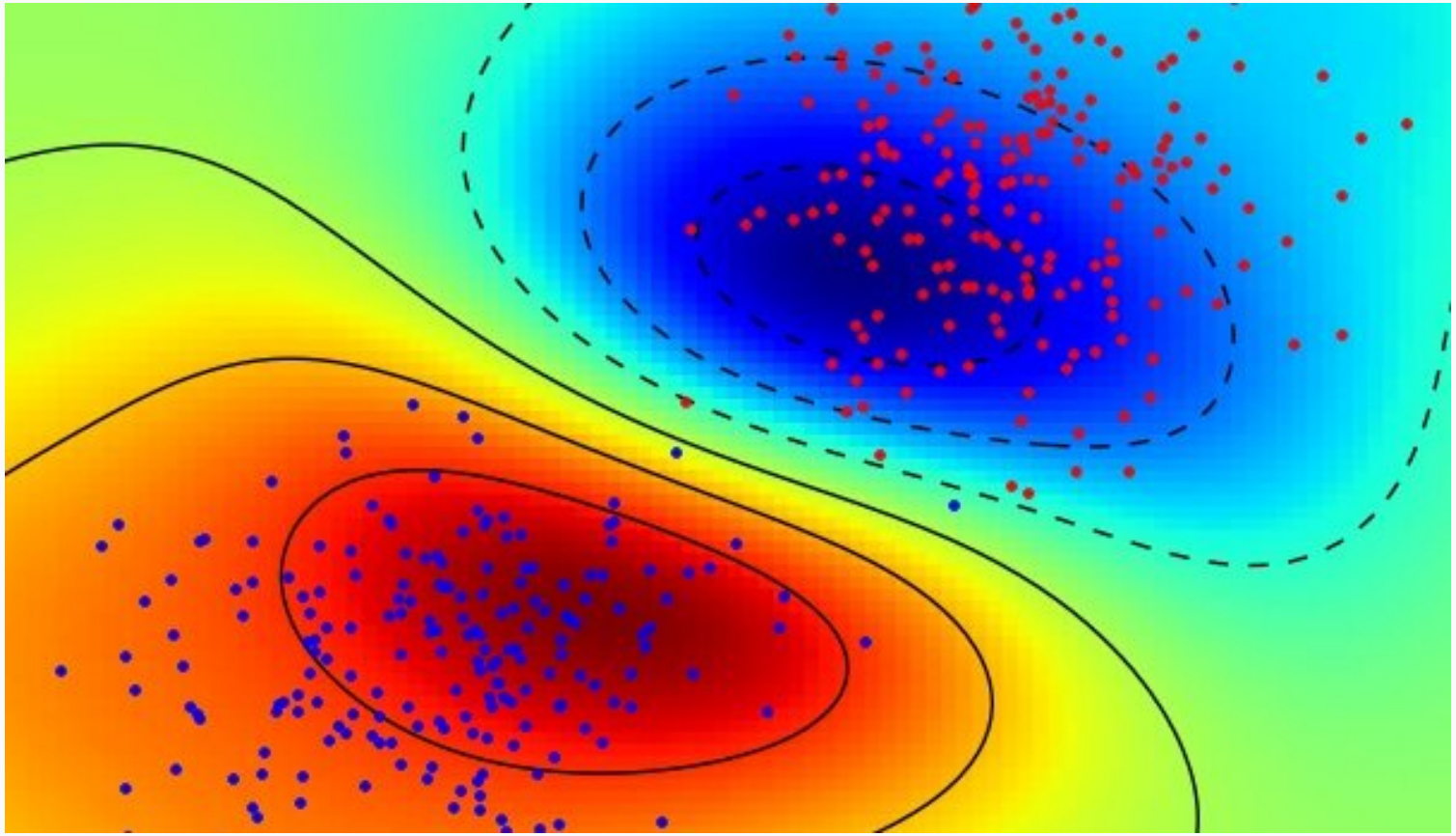


Image segmentation

Outline

1. Supervised classification
2. Empirical risk minimization
3. Vapnik-Chervonenkis theory
4. Margin-based bounds
5. Convex loss functions
6. Model selection
7. Stone's theorem
8. k-nearest neighbor classifiers
9. Partitioning classifiers and trees
10. Neural networks
11. Quantization and clustering

Supervised classification

Basics

- Random pair: $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$.
- X is the observation and Y is the label (or class).
- Variants: $X \in \mathscr{X}$, $Y \in \{-1, 1\}$, $Y \in \{1, \ldots, M\}$, etc.
- The distribution of (X, Y) is described by:
  - ▷ The distribution of X: $\mu(A) = \mathbb{P}(X \in A)$, $A \in \mathscr{B}(\mathbb{R}^d)$;
  - ▷ The a posteriori probability: $\eta(x) = \mathbb{P}(Y = 1 | X = x)$.
- Note: $\eta(x) = \mathbb{E}(Y|X = x) \stackrel{\text{def}}{=} r(x)$.
- Are X and Y independent? Not necessarily.
- Do we have $Y = \varphi(X)$? Not necessarily.
- Objective: find a classifier $g : \mathbb{R}^d \to \{0, 1\}$ such that $g(X) \approx Y$.
- Error probability: $L(g) = \mathbb{P}(g(X) \neq Y)$.
- 0-1 loss: $L(g) = \mathbb{E}\mathbb{1}_{[g(X) \neq Y]}$.

Bayes classifier

Definition

$$g^\star(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 & \text{otherwise,} \end{cases}$$

i.e.,

$$g^\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Lemma

For any classifier g, $L(g^\star) \leq L(g)$. Thus, $g^\star$ is the optimal decision.

- $L^\star \stackrel{\text{def}}{=} L(g^\star)$ is the Bayes risk (or Bayes error).
- One has

$$L^\star = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}(g(X) \neq Y).$$

Bayes risk

- Exercise 1:

$$L^\star = 1 - \mathbb{E}\big[\mathbb{1}_{[\eta(X)>1/2]}\eta(X) + \mathbb{1}_{[\eta(X)\leq 1/2]}(1-\eta(X))\big].$$

- Exercise 2:

$$L^\star = \mathbb{E}\big[\min(\eta(X), 1-\eta(X))\big] = \frac{1}{2} - \frac{1}{2}\mathbb{E}|2\eta(X)-1|.$$

- Exercise 3:

$$L^\star = 0 \Leftrightarrow Y = \varphi(X) \text{ with probability one.}$$

Statistical learning

- Sample: $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. copies of (X, Y).
- (X, Y) and $\mathscr{D}_n$ are independent.
- A classifier: $g_n(x) = g_n(x; \mathscr{D}_n)$ taking values in $\{0, 1\}$.
- Error probability: $L(g_n) = \mathbb{P}(g_n(X) \neq Y | \mathscr{D}_n)$.
- Objective: construct g_n such that $L(g_n) \approx L^\star$.

Statistics = inference. Learning = prediction.

Consistency

Definition (Consistency)

A classifier g_n is consistent for a certain distribution of (X, Y) if

$$\mathbb{E}L(g_n) = \mathbb{P}(g_n(X) \neq Y) \to L^\star \quad \text{as } n \to \infty,$$

and strongly consistent if

$$L(g_n) \to L^\star \quad \text{almost surely.}$$

- ▷ Exercise: consistency $\Leftrightarrow L(g_n) \xrightarrow{L^1} L^\star \Leftrightarrow L(g_n) \xrightarrow{\mathbb{P}} L^\star$.

Definition (Universal consistency)

A classifier is called universally (strongly) consistent if it is (strongly) consistent for any distribution of (X, Y).

Empirical risk minimization

Principle

- A class $\mathscr{C}$ of classifiers $g : \mathbb{R}^d \to \{0, 1\}$.
- Objective: find $g_n^\star \in \mathscr{C}$ such that $L(g_n^\star) \approx \inf_{g \in \mathscr{C}} L(g)$.
- Natural choice: $g_n^\star \in \arg\min_{g \in \mathscr{C}} L_n(g)$, where

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[g(X_i) \neq Y_i]} \quad \text{(empirical risk of } g\text{)}.$$

- Error probability: $L(g_n^\star) = \mathbb{P}(g_n^\star(X) \neq Y | \mathscr{D}_n)$.
- Decomposition:

$$\begin{aligned} L(g_n^\star) - L^\star &= \left[L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g) \right] + \left[\inf_{g \in \mathscr{C}} L(g) - L^\star \right] \\ &= \text{estimation error} + \text{approximation error}. \end{aligned}$$

Estimation = random. Approximation = deterministic.

Overfitting

- Estimation vs. approximation.
- Small $\mathscr{C}$: restrictive. Large $\mathscr{C}$: overfitting.
- Example:
  - $\mathscr{C}$ = all measurable functions;
  - Approximation error = 0;
  - But $L_n(g_n^\star) = 0$ with

$$g_n^\star(x) = \begin{cases} Y_i & \text{for } x = X_i,\ i = 1, \ldots, n \\ 0 & \text{otherwise;} \end{cases}$$

  - No generalization ability.
- Learning strategy: keep $\mathscr{C}$ under control.

Starting point

Lemma

One has

(*i*) $L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g) \leq 2 \sup_{g \in \mathscr{C}} |L_n(g) - L(g)|$,
and

(*ii*) $|L_n(g_n^\star) - L(g_n^\star)| \leq \sup_{g \in \mathscr{C}} |L_n(g) - L(g)|$.

Take-home message: $\sup_{g \in \mathscr{C}} |L_n(g) - L(g)|$ small guarantees:

- ▷ That $L(g_n^\star)$ is not much larger than the best error probability in $\mathscr{C}$;
- ▷ That the empirical estimate $L_n(g_n^\star)$ of $L(g_n^\star)$ is also good.

Concentration

Fact: $nL_n(g) \stackrel{\mathscr{D}}{=} \mathrm{Bin}(n, L(g))$.

☞ We need uniform deviations of binomial r.v. from their means.

Theorem (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that each X_i takes its values in $[a_i, b_i]$ ($a_i < b_i$) with probability one. Then, for all $t > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \geq t\Big) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

and

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \leq -t\Big) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

In particular,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\Big| \geq t\Big) \leq 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

Chernoff's bounding method

Lemma

Let X be a real-valued random variable with $\mathbb{E}X = 0$ and $X \in [a, b]$ $(a < b)$ with probability one. Then, for all $s \geq 0$,

$$\mathbb{E}e^{sX} \leq e^{s^2(b-a)^2/8}.$$

Proof of Hoeffding's inequality:

$$\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i \geq t\Big) &\leq e^{-st}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} \\
&= e^{-st}e^{s^2\sum_{i=1}^n (b_i-a_i)^2/8} \\
&= e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2},
\end{aligned}$$

by choosing $s = 4t/\sum_{i=1}^n (b_i - a_i)^2$.

The case $|\mathscr{C}| < \infty$

Corollary

Let $X \overset{\mathscr{D}}{=} Bin(n, p)$, $n \geq 1$ and $p \in [0, 1]$. Then, for all $t > 0$,

$$\mathbb{P}(|X - np| \geq t) \leq 2e^{-2t^2/n}.$$

Theorem

Assume that $|\mathscr{C}|$ is finite, with $|\mathscr{C}| \leq N$. Then, for all $t > 0$,

$$\mathbb{P}\Big(\sup_{g \in \mathscr{C}} |L_n(g) - L(g)| \geq t \Big) \leq 2Ne^{-2nt^2}.$$

- ▷ The bound is universal.
- ▷ Borel-Cantelli: $\sup_{g \in \mathscr{C}} |L_n(g) - L(g)| \to 0$ almost surely.
- ▷ Consequence: $L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g) \to 0$ almost surely.
- ▷ Bound on $\mathbb{E}L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g)$?

From $\mathbb{P}$ to $\mathbb{E}$

Lemma

Let X be a random variable taking values in $\mathbb{R}_+$. Assume that there exists a constant $C \geq 1$ such that, for all $t > 0$, $\mathbb{P}(X \geq t) \leq Ce^{-2nt^2}$. Then

$$\mathbb{E}X \leq \sqrt{\frac{\log(Ce)}{2n}}.$$

- ▷ Consequence:

$$\mathbb{E} \sup_{g \in \mathscr{C}} |L_n(g) - L(g)| \leq \sqrt{\frac{\log(2eN)}{2n}}$$

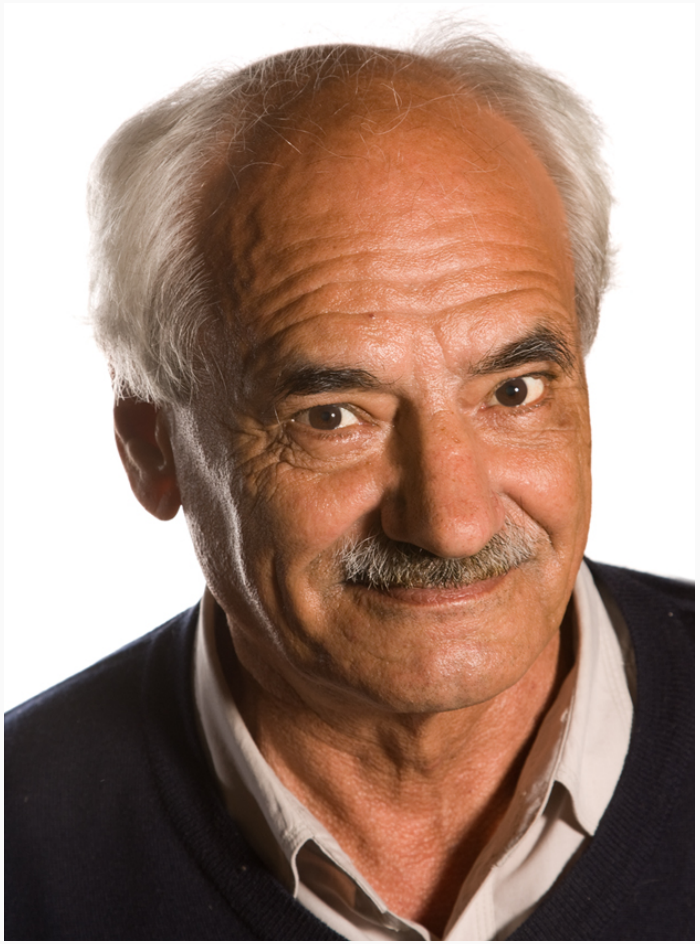
and

$$\mathbb{E}L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g) \leq 2\sqrt{\frac{\log(2eN)}{2n}}.$$

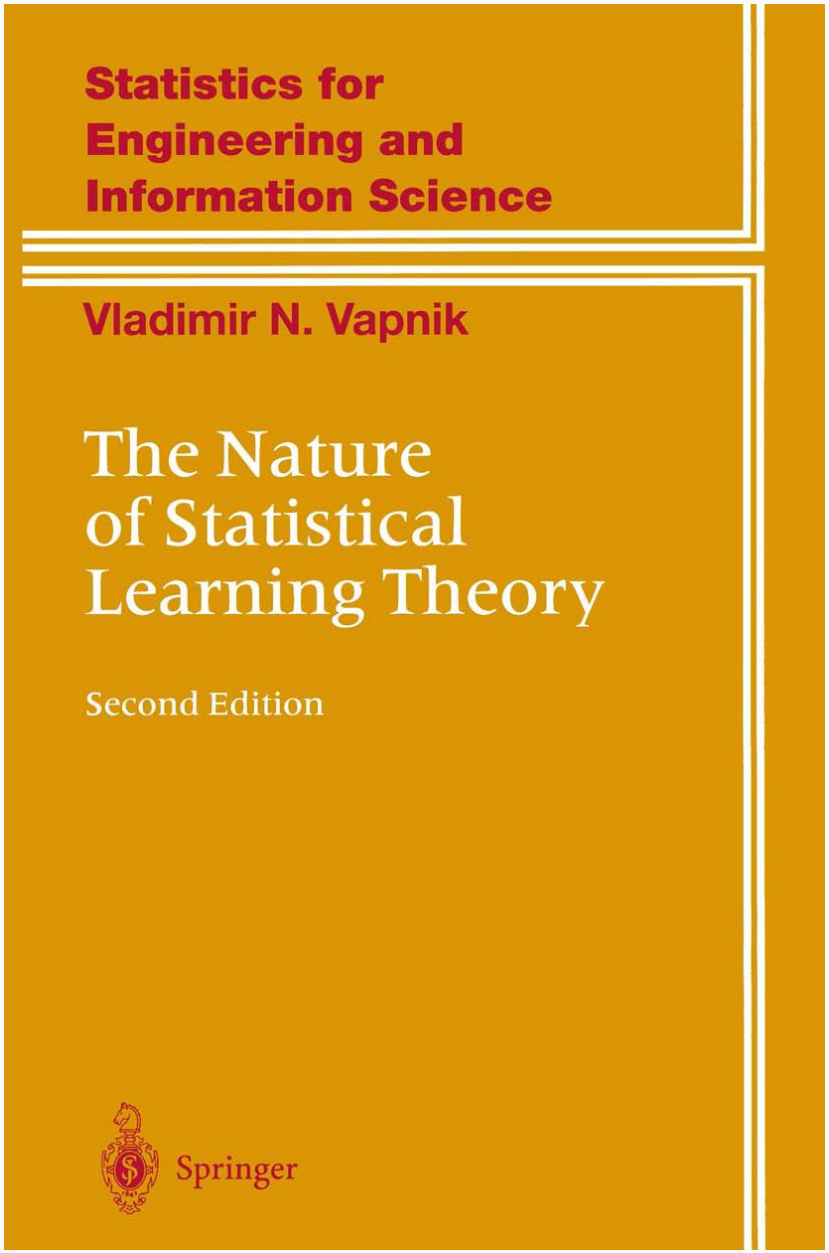
- ▷ Take-home message: estimation error $= \mathrm{O}\big(\sqrt{\frac{\log N}{n}}\big)$.
- ▷ Next objective: handle more complex classes of functions.

Vapnik-Chervonenkis theory

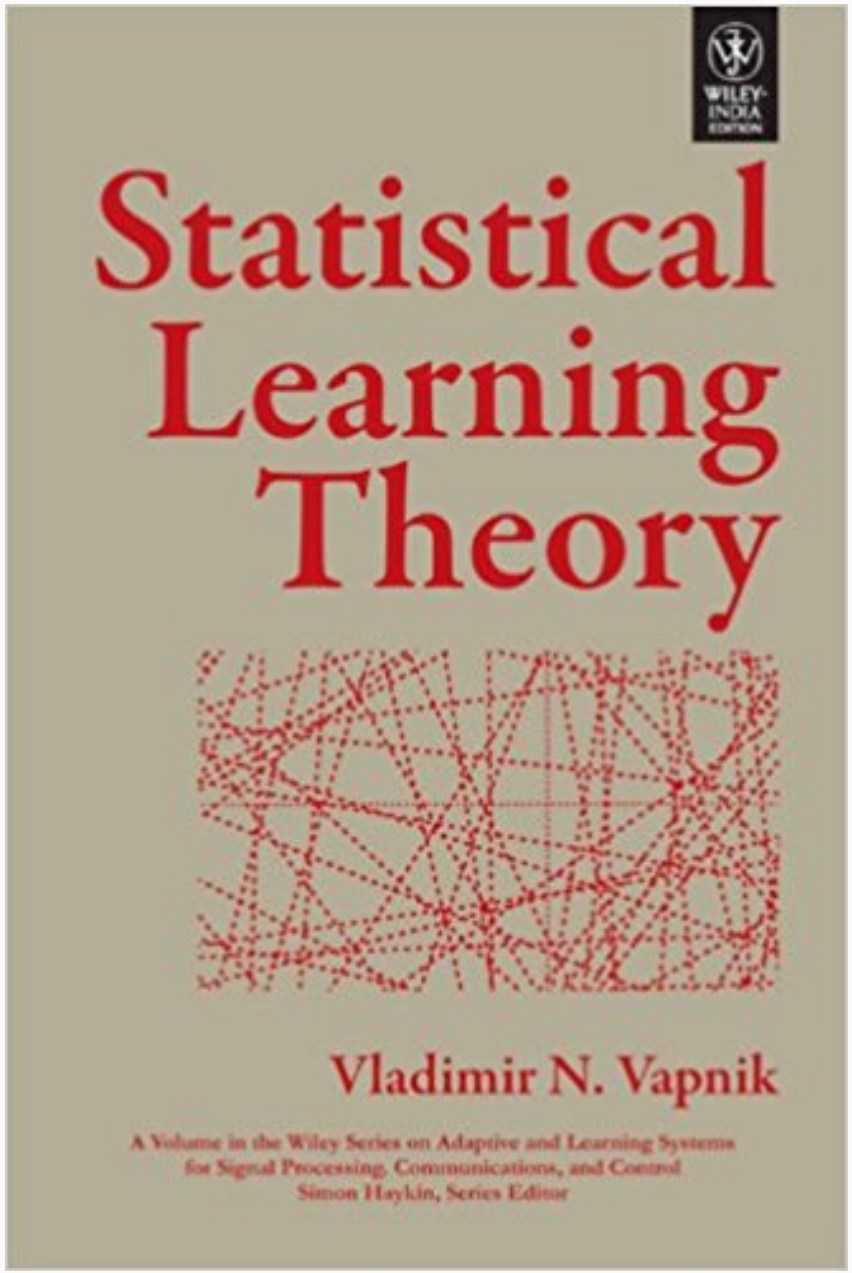
Vladimir Vapnik and Alexey Chervonenkis



References


Statistics for
Engineering and
Information Science
Vladimir N. Vapnik
The Nature
of Statistical
Learning Theory
Second Edition
Springer

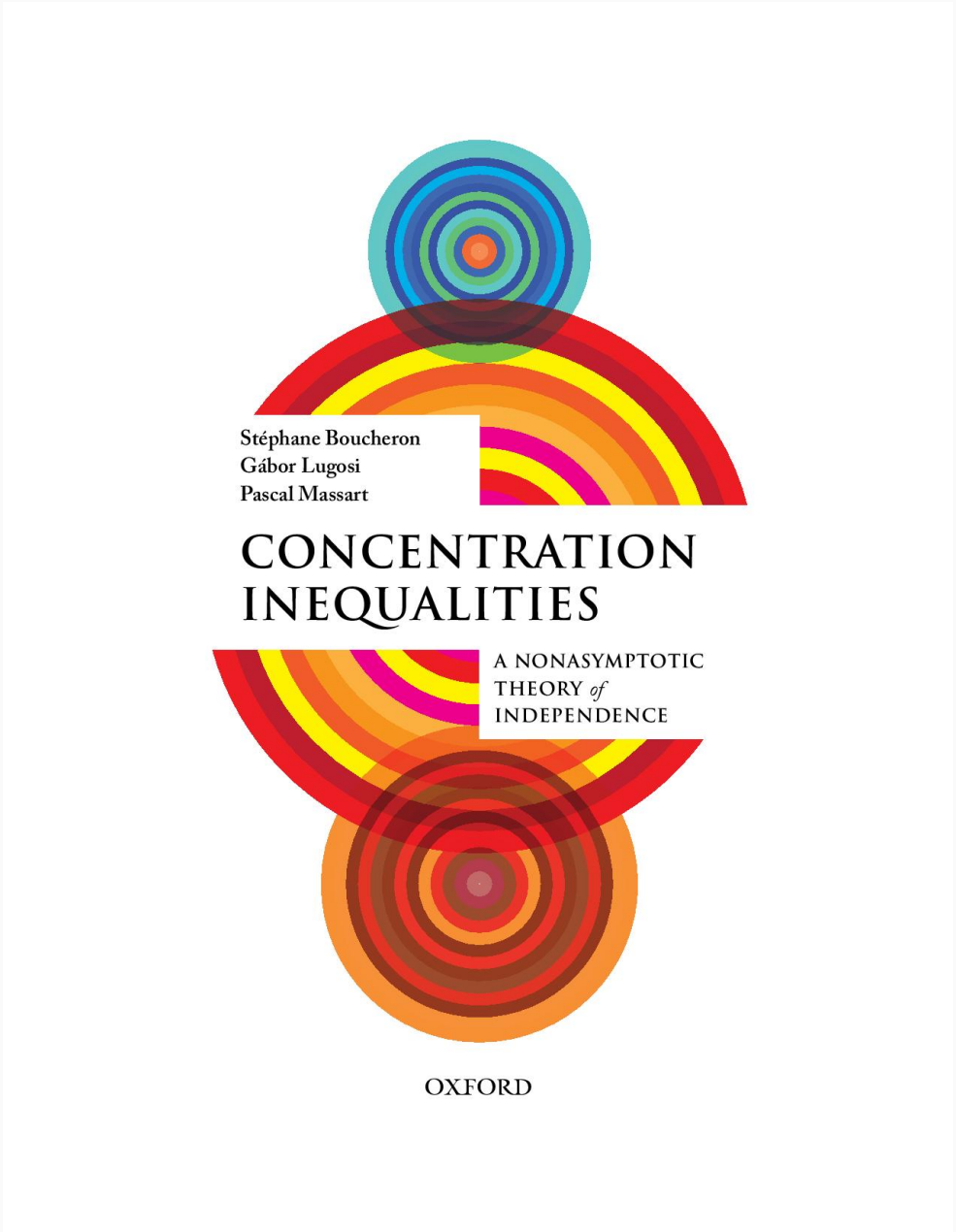
References


WILEY-INDIA EDITION
Statistical Learning Theory
Vladimir N. Vapnik
A Volume in the Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control
Simon Haykin, Series Editor

From finite to infinite classes

- Learning: functions of the form $f = \mathbb{1}_{[(x,y):g(x)\neq y]}$, $g \in \mathscr{C}$.
- General context:
  - $X_1, \ldots, X_n$ = i.i.d. random variables taking values in a set $\mathscr{X}$;
  - $\mathscr{F}$ = a class of bounded functions $f : \mathscr{X} \to [0,1]$ (or [-1,1]).
- Notation: $Pf = \mathbb{E}f(X_1)$ and $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.
- Objective: bound $\sup_{f\in\mathscr{F}} |P_n f - Pf|$.
- Basic tool: concentration inequalities.

Reference


Stéphane Boucheron
Gábor Lugosi
Pascal Massart
CONCENTRATION INEQUALITIES
A NONASYMPTOTIC THEORY of INDEPENDENCE
OXFORD

Bounded difference inequality

Theorem (Bounded difference inequality)

Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathscr{X}$. Assume that $g : \mathscr{X}^n \to \mathbb{R}$ is Borel measurable and satisfies

$$\sup_{\substack{(x_1,\ldots,x_n)\in\mathscr{X}^n \\ x_i'\in\mathscr{X}}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \le c_i, \quad 1 \le i \le n,$$

for some positive constants $c_1, \ldots, c_n$ (bounded difference assumption). Then, for all $t > 0$,

$$\mathbb{P}\big(g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n) \ge t\big) \le e^{-2t^2/\sum_{i=1}^n c_i^2}$$

and

$$\mathbb{P}\big(g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n) \le -t\big) \le e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

In particular,

$$\mathbb{P}\big(\big|g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n)\big| \ge t\big) \le 2e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Concentration

- Natural choice: $g(X_1, \ldots, X_n) = \sup_{f \in \mathscr{F}} |P_n f - Pf|$.
- The bounded difference assumption is satisfied with $c_i = 1/n$.
- Concentration: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathscr{F}} |P_n f - Pf| \leq \mathbb{E} \sup_{f \in \mathscr{F}} |P_n f - Pf| + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

- Consequence: focus on the expected value $\mathbb{E} \sup_{f \in \mathscr{F}} |P_n f - Pf|$.
- Key: Rademacher averages and symmetrization principle.

Rademacher averages

- A random variable σ is Rademacher if $\mathbb{P}(\sigma = \pm 1) = 1/2$.
- Definition:
  - ▷ Let $A \subseteq \mathbb{R}^n$ be a bounded set of vectors $a = (a_1, \ldots, a_n)$;
  - ▷ Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. Rademacher random variables;
  - ▷ Rademacher average associated with A:

$$R_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i a_i \Big|.$$

- $\mathscr{F}(x_1^n)$ = the class of n-vectors $(f(x_1), \ldots, f(x_n))$, $f \in \mathscr{F}$.
- Key quantity:

$$R_n(\mathscr{F}(X_1^n)) = \mathbb{E}\Big(\sup_{f \in \mathscr{F}} \frac{1}{n} \Big| \sum_{i=1}^{n} \sigma_i f(X_i) \Big| \,\Big|\, X_1, \ldots, X_n \Big),$$

where $\sigma_1, \ldots, \sigma_n$ are independent of the X_i.

$R_n(\mathscr{F}(X_1^n)) = \mathbb{E}_\sigma \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right|$ is data-dependent.

Fundamental inequalities

Theorem

One has

$$\mathbb{E}\sup_{f\in\mathscr{F}}|P_nf - Pf| \le 2\mathbb{E}R_n(\mathscr{F}(X_1^n)).$$

In addition, for any $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\sup_{f\in\mathscr{F}}|P_nf - Pf| \le 2\mathbb{E}R_n(\mathscr{F}(X_1^n)) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and, with probability at least $1-\delta$*,*

$$\sup_{f\in\mathscr{F}}|P_nf - Pf| \le 2R_n(\mathscr{F}(X_1^n)) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

- The last inequality is a data-dependent performance bound.
- Sharp estimates of the maximal deviations.
- In fact,

$$\frac{1}{2}\mathbb{E}R_n(\mathscr{F}(X_1^n)) - \frac{1}{2\sqrt{n}} \le \mathbb{E}\sup_{f\in\mathscr{F}}|P_nf - Pf| \le 2\mathbb{E}R_n(\mathscr{F}(X_1^n)).$$

Properties of Rademacher averages

Theorem

Let A, B be bounded subsets of $\mathbb{R}^n$, and let $c \in \mathbb{R}$ be a constant. Then

$$R_n(A \cup B) \leq R_n(A) + R_n(B), \quad R_n(c \cdot A) = |c| R_n(A),$$

and

$$R_n(A \oplus B) \leq R_n(A) + R_n(B),$$

where $c \cdot A = \{ca : a \in A\}$ and $A \oplus B = \{a + b : a \in A, b \in B\}$. In addition, if $A = \{a^{(1)}, \ldots, a^{(N)}\} \subseteq \mathbb{R}^n$ is a finite set, then

$$R_n(A) \leq \max_{1 \leq j \leq N} \|a^{(j)}\| \frac{\sqrt{2 \log(2N)}}{n},$$

where $\|\cdot\|$ denotes Euclidean norm.

Class of indicators

- Learning: functions of the form $f = \mathbb{1}_{[(x,y):g(x)\neq y]}$, $g \in \mathscr{C}$.
- Binary framework: $\mathscr{F} = \{\mathbb{1}_A : A \in \mathscr{A}\}$, with $|\mathscr{A}| \geq 2$.
- In this case, for $x_1^n = (x_1, \ldots, x_n)$, $\mathscr{F}(x_1^n)$ is a finite subset of $\mathbb{R}^n$.
- Combinatorial quantity: $|\mathscr{F}(x_1^n)|$.
- Shatter coefficient: $\mathsf{S}_{\mathscr{A}}(n) = \max_{x_1^n} |\mathscr{F}(x_1^n)|$.
- Properties: $\mathsf{S}_{\mathscr{A}}(1) = 2$, $2 \leq \mathsf{S}_{\mathscr{A}}(n) \leq 2^n$, and

$$\mathsf{S}_{\mathscr{A}}(k) < 2^k \text{ for some } k > 1 \Leftrightarrow \mathsf{S}_{\mathscr{A}}(n) < 2^n \text{ for all } n \geq k.$$

- For all x_1^n,

$$R_n(\mathscr{F}(x_1^n)) \leq \sqrt{\frac{2\log(2\mathsf{S}_{\mathscr{A}}(n))}{n}} \leq 2\sqrt{\frac{\log \mathsf{S}_{\mathscr{A}}(n)}{n}}.$$

- $\log \mathsf{S}_{\mathscr{A}}(n)$ may be bounded in terms of the VC dimension.

VC dimension

Definition (VC dimension)

The VC dimension $V_{\mathscr{A}}$ of $\mathscr{A}$ is the largest integer $n_0 \geq 1$ for which $S_{\mathscr{A}}(n_0) = 2^{n_0}$. If $S_{\mathscr{A}}(n) = 2^n$ for all $n \geq 1$, then $V_{\mathscr{A}} = \infty$.

- Exercise 1: if $|\mathscr{A}| < \infty$, then $S_{\mathscr{A}}(n) \leq |\mathscr{A}|$ and $V_{\mathscr{A}} \leq \log_2 |\mathscr{A}|$.
- Exercise 2: if $\mathscr{A} = \{(-\infty, a] : a \in \mathbb{R}\}$, then $V_{\mathscr{A}} = 1$. If $\mathscr{A} = \{[a, b] : (a, b) \in \mathbb{R}^2\}$, then $V_{\mathscr{A}} = 2$.
- Generalization: if $\mathscr{A} = \{(-\infty, a_1] \times \cdots \times (-\infty, a_p] : (a_1, \ldots, a_p) \in \mathbb{R}^p\}$, then $V_{\mathscr{A}} = p$. If $\mathscr{A} = \{\text{rectangles of } \mathbb{R}^p\}$, then $V_{\mathscr{A}} = 2p$. If $\mathscr{A} = \{\text{convex polygons of } \mathbb{R}^2\}$, then $V_{\mathscr{A}} = \infty$.
- Important: for $\mathscr{G}$ a finite-dimensional vector space of functions $\mathbb{R}^p \to \mathbb{R}$, and for
$$\mathscr{A} = \big\{\{x \in \mathbb{R}^p : g(x) \geq 0\} : g \in \mathscr{G}\big\},$$
then $V_{\mathscr{A}} \leq \dim \mathscr{G}$.
- Consequence: if $\mathscr{A} =$ subsets of $\mathbb{R}^p$ of the form $\{x \in \mathbb{R}^p : a^\top x + b \geq 0 : a \in \mathbb{R}^p, b \in \mathbb{R}\}$, then $V_{\mathscr{A}} \leq p + 1$.

Sauer's lemma

Theorem (Sauer's lemma)

If $V_{\mathscr{A}} < \infty$, then, for all $n \geq 1$, $\mathsf{S}_{\mathscr{A}}(n) \leq \sum_{i=1}^{V_{\mathscr{A}}} \binom{n}{i}$.

- ▷ Exercise: $\mathsf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}$.
- ▷ Only two cases:
  - (i) Either $V_{\mathscr{A}} = \infty \to \mathsf{S}_{\mathscr{A}}(n) = 2^n$ for all $n \geq 1$;
  - (ii) Either $V_{\mathscr{A}} < \infty \to \mathsf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}$ for all $n \geq 1$.
- ▷ Never $\mathsf{S}_{\mathscr{A}}(n) \sim 2^{\sqrt{n}}$.
- ▷ Important consequence: $\log \mathsf{S}_{\mathscr{A}}(n) \leq V_{\mathscr{A}} \log(n+1)$ and

$$R_n(\mathscr{F}(x_1^n)) \leq 2\sqrt{\frac{V_{\mathscr{A}} \log(n+1)}{n}}.$$

Vapnik-Chervonenkis inequality

Theorem (Vapnik-Chervonenkis inequality)

For all distributions, one has

$$\mathbb{E} \sup_{f \in \mathscr{F}} |P_n(f) - P(f)| \leq 4\sqrt{\frac{V_{\mathscr{A}} \log(n+1)}{n}}.$$

- Chaining arguments: $\mathbb{E} \sup_{f \in \mathscr{F}} |P_n(f) - P(f)| \leq c\sqrt{\frac{V_{\mathscr{A}}}{n}}$.
- Variants: $\mathbb{P}(\sup_{f \in \mathscr{F}} |P_n(f) - P(f)| \geq t) \leq c_1 S_{\mathscr{A}}(n) e^{-c_2 n\varepsilon^2}$.
- Take-home message: $\mathbb{E} \sup_{f \in \mathscr{F}} |P_n(f) - P(f)| = \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{A}} \log n}{n}}\right)$.

The VC inequality is valid for all distributions!

Back to learning

- Reminder: $L(g_n^\star) - \inf_{g\in\mathscr{C}} L(g) \le 2\sup_{g\in\mathscr{C}} |L_n(g) - L(g)|$.
- Strategy:
  - ▷ $L(g_n^\star) - \inf_{g\in\mathscr{C}} L(g) \le 2\sup_{f\in\mathscr{F}} |P_n f - Pf|$;
  - ▷ $\mathscr{F}$ = indicator functions of the form $f = \mathbb{1}_{[(x,y):g(x)\neq y]}$, $g \in \mathscr{C}$;
  - ▷ $\mathscr{A} = \{A_g : g \in \mathscr{C}\}$, where $A_g = \{(x,y) \in \mathbb{R}^d \times \{0,1\} : g(x) \neq y\}$.

Proposition

For all $n \ge 1$, $\mathsf{S}_{\bar{\mathscr{A}}}(n) = \mathsf{S}_{\mathscr{A}}(n)$, where

$$\bar{\mathscr{A}} = \big\{\{x \in \mathbb{R}^d : g(x) = 1\} : g \in \mathscr{C}\big\}.$$

In particular, $V_{\bar{\mathscr{A}}} = V_{\mathscr{A}}$.

- ▷ Notation: $V_{\mathscr{C}}$ instead of $V_{\bar{\mathscr{A}}}$.
- ▷ Important consequence:

$$\mathbb{E}L(g_n^\star) - \inf_{g\in\mathscr{C}} L(g) = \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{C}} \log n}{n}}\right).$$

Examples

- Linear classification:

$$g(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{d} a_j x^{(j)} + a_0 > 0 \\ 0 & \text{otherwise,} \end{cases}$$

  where $(a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}$. Thus,

$$\bar{\mathscr{A}} \subseteq \left\{ \{x \in \mathbb{R}^d : a^\top x + a_0 \geq 0\} : a \in \mathbb{R}^d, a_0 \in \mathbb{R} \right\},$$

  and $V_{\mathscr{C}} \leq d + 1$.
- Exercise: $\eta(x) = 0.1\mathbb{1}_{[x^{(1)} \leq 1/2]} + 0.9\mathbb{1}_{[x^{(1)} > 1/2]}$. Compute $g^\star$, $L^\star$, and $\inf_{g \in \mathscr{C}} L(g) - L^\star$.
- Closed balls:

$$\bar{\mathscr{A}} = \left\{ \{x \in \mathbb{R}^d : \sum_{j=1}^{d} (x^{(j)} - a_j)^2 \leq a_0\} : (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1} \right\}.$$

  We see that $\bar{\mathscr{A}} \subseteq \{\{x \in \mathbb{R}^d : g(x) \geq 0\} : g \in \mathscr{G}\}$, where $\mathscr{G}$ is a vector space of dimension $d + 2 \rightarrow V_{\mathscr{C}} \leq d + 2$.

Examples, cont.

- Convex polygons of $\mathbb{R}^2$: bad idea since $V_{\bar{\mathscr{A}}} = \infty$.
- Generalized linear classification: $\psi_1, \ldots, \psi_{d^\star}$ functions from $\mathbb{R}^d \to \mathbb{R}$ and

$$g(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{d^\star} a_j \psi_j(x) + a_0 > 0 \\ 0 & \text{otherwise,} \end{cases}$$

  where $(a_0, a_1, \ldots, a_{d^\star}) \in \mathbb{R}^{d^\star+1}$.
  - ▷ $\psi_j(x) = x^{(j)}$: linear classifiers ✔
  - ▷ ψ_j = coordinates + products of coordinates:

$$\bar{\mathscr{A}} \subseteq \left\{ a_0 + \sum_{j=1}^{d} a_j x^{(j)} + \sum_{j=1}^{d} b_j (x^{(j)})^2 + \sum_{1 \le j_1 < j_2 \le d} c_{j_1 j_2} x^{(j_1)} x^{(j_2)} \ge 0 \right\}.$$

  Conclusion: $V_{\mathscr{C}} \le d^\star + 1$, where $d^\star = 2d + \frac{d(d-1)}{2}$.

Margin-based bounds

Drawbacks of empirical risk minimization

- For classes with finite VC dimension:

$$\mathbb{E}L(g_n^\star) - \inf_{g \in \mathscr{C}} L(g) = \mathrm{O}\left(\sqrt{\frac{V_{\mathscr{C}} \log n}{n}}\right).$$

- However, such classes are nearly always too small.
- $L(g_n^\star)$ is typically far from the Bayes risk $L^\star$.
- Example: for any class $\mathscr{C}$ with finite VC dimension and for all $\varepsilon \in (0, 1/2)$, there exists (X, Y) such that

$$\inf_{g \in \mathscr{C}} L(g) - L^\star > 1/2 - \varepsilon.$$

- Besides, minimizing $L_n(g)$ is often a computationally difficult problem.
- Solution: modify the empirical functional to be minimized.

A different framework

- We consider ± 1-classifiers of the form

$$g_f(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{otherwise,} \end{cases}$$

  where $f : \mathbb{R}^d \to \mathbb{R}$, $f \in \mathscr{F}$ bounded.
- Error probability: $\mathbb{P}(Yf(X) < 0) \leq L(g_f) \leq \mathbb{P}(Yf(X) \leq 0)$.
- Notation: $L(f) = L(g_f)$.
- Empirical error probability:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i f(X_i) < 0]} \leq L_n(f) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i f(X_i) \leq 0]}.$$

- The product $Yf(X)$ is called the margin.

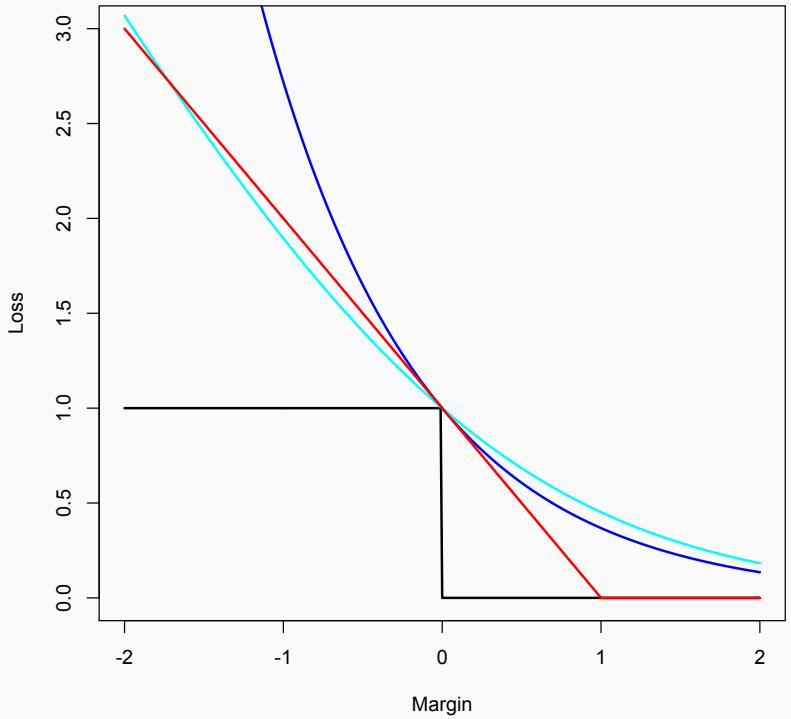
Large margin = good confidence.

Loss functions

- $\phi : \mathbb{R} \to \mathbb{R}_+$ a loss function such that

$$\phi(0) = 1, \quad \mathbb{1}_{[x \leq 0]} \leq \phi(x), \quad \phi \text{ is } L_\phi\text{-Lipschitz on the range of } \mathscr{F}.$$

- Typical choices:



▷ Exponential: $\phi(x) = e^{-x}$

▷ Logit: $\phi(x) = \log_2(1 + e^{-x})$

▷ Hinge: $\phi(x) = \max(1-x, 0)$!!!!

- Risk functional: $A(f) = \mathbb{E}\phi(Yf(X))$.
- Empirical risk functional: $A_n(f) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i))$.
- Clearly: $L(f) \leq A(f)$ and $L_n(f) \leq A_n(f)$.

Contraction principle

Theorem

Let A be a bounded subset of $\mathbb{R}^n$. If

$$\bar{A} = \Big\{ \sum_{j=1}^{N} c_j a^{(j)} : N \in \mathbb{N}, a^{(j)} \in A, \sum_{j=1}^{N} |c_j| = 1 \Big\},$$

then $R_n(\bar{A}) = R_n(A)$. In addition, if $\psi : \mathbb{R} \to \mathbb{R}$ is a function with $\psi(0) = 0$ and Lipschitz constant L_ψ, then

$$R_n(\psi \circ A) \leq 2L_\psi R_n(A) \quad \textit{(contraction principle)},$$

where $\psi \circ A$ is the set of vectors of the form $(\psi(a_1), \ldots, \psi(a_n)) \in \mathbb{R}^n$, $a = (a_1, \ldots, a_n) \in A$.

Rademacher averages and performance bounds

Theorem

Let $f_n^\star \in \arg\min_{f \in \mathscr{F}} A_n(f)$, *let* B *denote a uniform upper bound on* $\phi(yf(x))$, *and let* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$,

$$L(f_n^\star) \leq A_n(f_n^\star) + 4L_\phi \mathbb{E} R_n(\mathscr{F}(X_1^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

- Warning: L is a probability and A_n is empirical!
- The Rademacher average of $\mathscr{F}$ bounds the performance of $f_n^\star$.

Illustration 1: Weighted voting schemes

- Class:

$$\mathscr{F}_\lambda = \left\{ f = \sum_{j=1}^{N} c_j g_j : N \in \mathbb{N}, g_1, \ldots, g_N \in \mathscr{C}, \sum_{j=1}^{N} |c_j| = \lambda \right\},$$

  where $\mathscr{C}$ = class of ± 1 base classifiers.

- Rademacher averages:

$$R_n(\mathscr{F}_\lambda(x_1^n)) = \lambda R_n(\mathscr{C}(x_1^n)) \leq 2\lambda \sqrt{\frac{V_{\mathscr{C}} \log(n+1)}{n}}.$$

- Example:
  - $\mathscr{C}$ = classifiers of the form $g(x) = 2\mathbb{1}_{[x \geq a]} - 1$, $a \in \mathbb{R}$;
  - $V_{\mathscr{C}} = 1$ and $\overline{\mathscr{F}_\lambda}$ = all functions of total variation bounded by 2λ.
- To summarize: with probability $1 - \delta$,

$$L(f_n^\star) \leq A_n(f_n^\star) + 8L_\phi \lambda \sqrt{\frac{V_{\mathscr{C}} \log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

- Advantage: only $V_{\mathscr{C}}$ in the bound. Inconvenient: A_n instead of L_n.

Weighted voting schemes, cont.

- γ a positive parameter and

$$\phi(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 0 & \text{if } x \geq \gamma \\ 1 - x/\gamma & \text{otherwise.} \end{cases}$$

- Quick check: $B = 1$, $L_\phi = 1/\gamma$, $\mathbb{1}_{[x \leq 0]} \leq \phi(x) \leq \mathbb{1}_{[x < \gamma]}$.
- Consequence: $A_n(f) \leq L_n^\gamma(f)$, where

$$L_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[Y_i f(X_i) < \gamma]} \quad \text{is the margin error.}$$

- Properties: $L_n(f) \leq L_n^\gamma(f)$ and $L_n^\gamma(f)$ is nondecreasing in γ.
- Interpretation: $L_n^\gamma(f)$ counts the number of misclassified pairs + those which are well classified but only with a small margin.
- Margin-based bound: with probability at least $1 - \delta$,

$$L(f_n^\star) \leq L_n^\gamma(f_n^\star) + \frac{8\lambda}{\gamma} \sqrt{\frac{V_{\mathscr{C}} \log(n+1)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Illustration 2: Kernel methods

- Examples: Support Vector Machines and kernel Fisher discriminant.
- Necessitates a detour through the theory of Hilbert spaces.

Definition (Reproducing kernel)

Let $\mathscr{F}$ be a Hilbert space of functions $f : \mathscr{X} \to \mathbb{R}$ defined on a non-empty set $\mathscr{X}$. A function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a reproducing kernel of $\mathscr{F}$ if

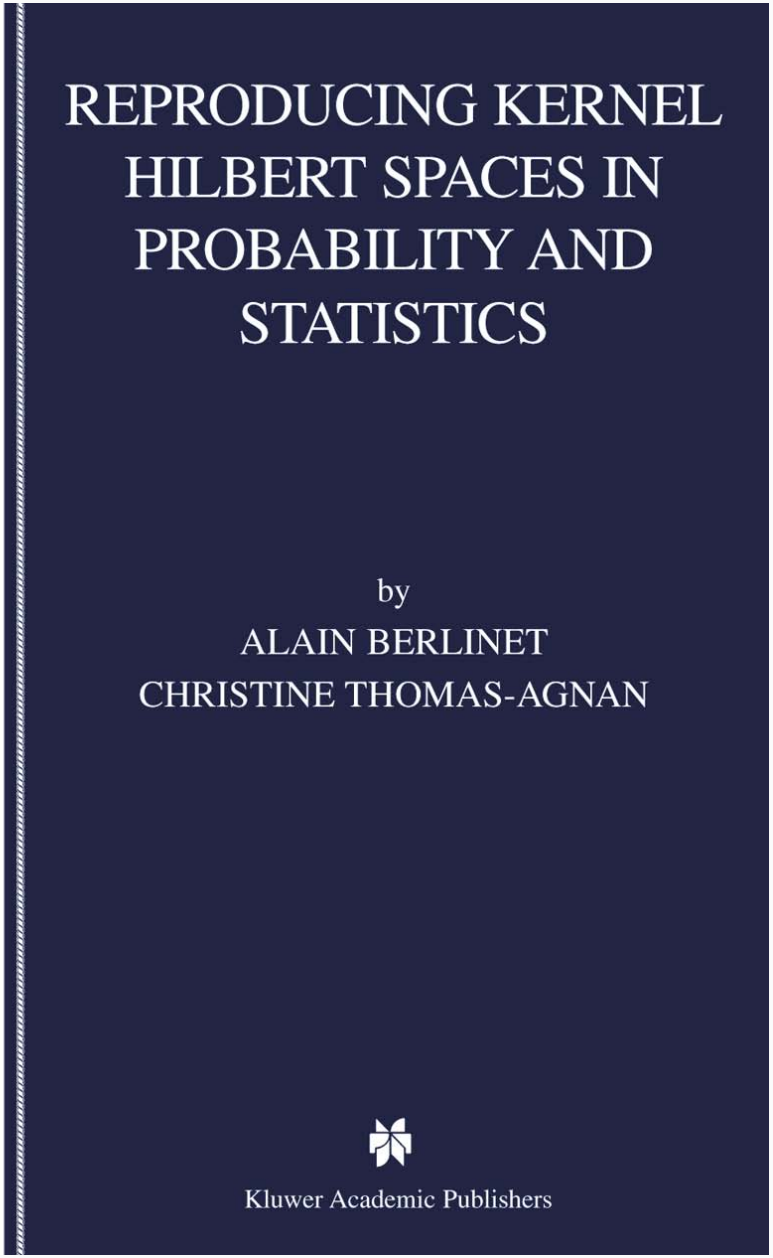
(i) $\forall x \in \mathscr{X}, k(\cdot, x) \in \mathscr{F}$;

(ii) $\forall f \in \mathscr{F}, \langle f, k(\cdot, x) \rangle = f(x)$ *(reproducing property).*

- ▷ $k(x, x') = k(x', x) = \langle k(\cdot, x), k(\cdot, x') \rangle$.
- ▷ Vocabulary: $\mathscr{F}$ is a reproducing kernel Hilbert space (RKHS).

RKHS $\Rightarrow$ Hilbert space, but Hilbert space $\not\Rightarrow$ RKHS.

Reference


REPRODUCING KERNEL HILBERT SPACES IN PROBABILITY AND STATISTICS
by
ALAIN BERLINET
CHRISTINE THOMAS-AGNAN
Kluwer Academic Publishers

Kernel methods, cont.

- Example 1: any finite-dimensional Hilbert space of functions is a RKHS, with $k(x, x') = \sum_{i=1}^{n} e_i(x) e_i(x')$.
- Example 2: the space $L^2(\mathbb{R})$ is not a RKHS.
- Example 3: the space

$$\mathscr{F} = \left\{ f : f(0) = 0, f \text{ is absolutely continuous, } f, f' \in L^2(\mathbb{R}) \right\}$$

  is a RKHS, with $k(x, x') = \frac{1}{2} e^{-|x - x'|}$.

Kernel methods, cont.

- Key: any reproducing kernel is symmetric and positive definite, i.e.,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \quad \forall n \geq 1, \forall (x_1, \ldots, x_n) \in \mathscr{X}^n, \forall (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n.$$

- Reciprocal property? Yes.

Theorem (Moore-Aronszajn theorem)

Let k be a positive definite function on $\mathscr{X} \times \mathscr{X}$. There exists only one Hilbert space $\mathscr{F}$ of functions on $\mathscr{X}$ with k as reproducing kernel. The subspace $\mathscr{F}_0$ of $\mathscr{F}$ spanned by the functions $(k(\cdot, x))_{x \in \mathscr{X}}$ is dense in $\mathscr{F}$, and $\mathscr{F}$ is the set of functions on $\mathscr{X}$ that are pointwise limits of Cauchy sequences in $\mathscr{F}_0$ with the inner product

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j),$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j)$.

Statistical consequences

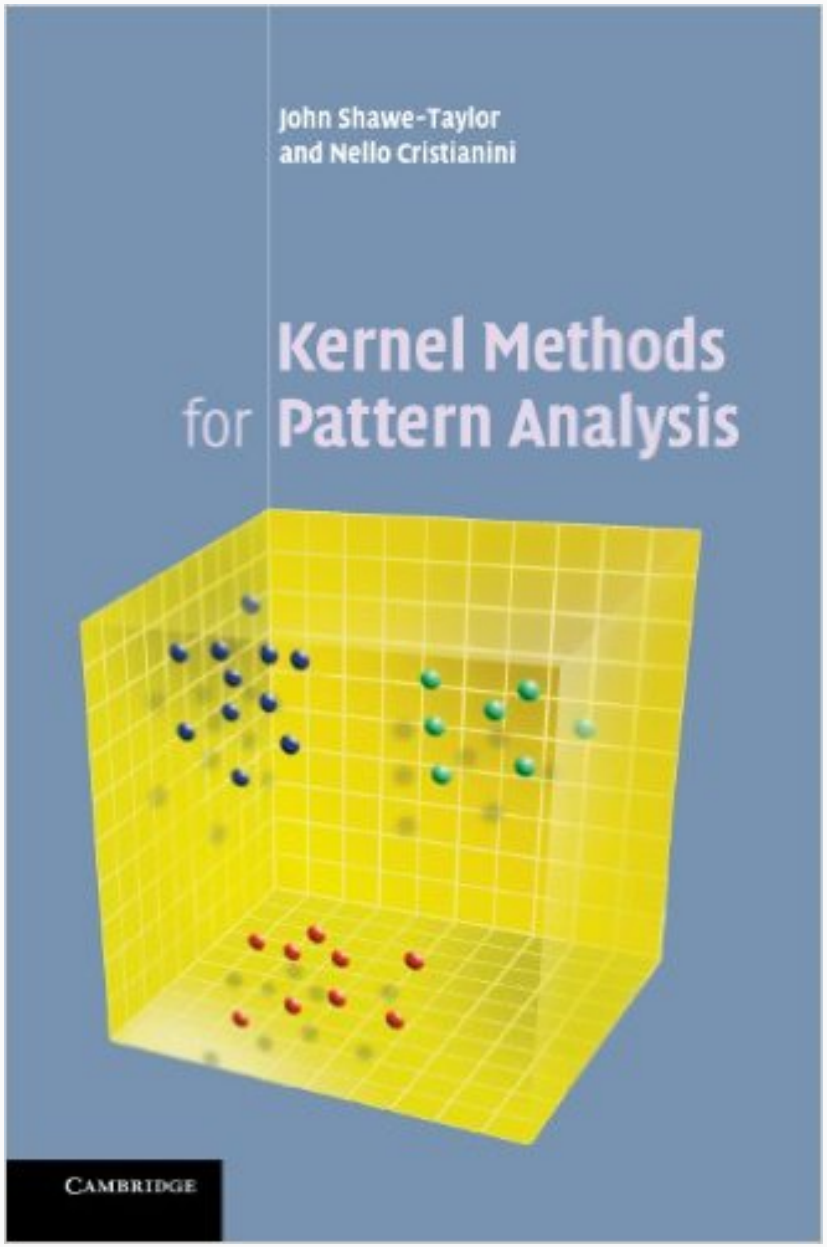
- Non-linear embedding:

$$\begin{array}{rccc} \varphi & : & \mathscr{X} & \to & \mathscr{F} \\ & & x & \mapsto & k(\cdot, x). \end{array}$$

- Rationale: $\forall (x, x') \in \mathscr{X}^2$, $\langle \varphi(x), \varphi(x') \rangle = k(x, x')$.
- Advantage: any linear algorithm based on computing inner products can be extended into a non-linear version by replacing the inner products by a kernel function $\to$ kernel trick.
- Even though the algorithm remains of low complexity, it works in a big class of functions.
- Science of kernel design (vectors, strings, graphs, etc.).

φ is never computed.

Reference


John Shawe-Taylor
and Nello Cristianini
Kernel Methods
for Pattern Analysis
CAMBRIDGE

Optimization in a RKHS

- Kernel algorithms often perform minimization over the ball of a RKHS:

$$\mathscr{F}_\lambda = \left\{ f \in \mathscr{F} : \|f\| \leq \lambda \right\}$$
$$\approx \left\{ f = \sum_{i=1}^{N} c_i k(\cdot, x_i) : N \in \mathbb{N}, x_1, \ldots, x_N \in \mathscr{X}, \sum_{i,j=1}^{N} c_i c_j k(x_i, x_j) \leq \lambda^2 \right\}.$$

- Example: Support Vector Machines (SVM).

Theorem

$$\frac{\lambda}{n\sqrt{2}} \sqrt{\sum_{i=1}^{n} k(X_i, X_i)} \leq R_n(\mathscr{F}_\lambda(X_1^n)) \leq \frac{\lambda}{n} \sqrt{\sum_{i=1}^{n} k(X_i, X_i)}.$$

- ▷ Nice bound, which can be computed very easily from the data.
- ▷ Performance bound in the RKHS: with probability $1 - \delta$,

$$L(f_n^\star) \leq L_n^\gamma(f_n^\star) + \frac{4\lambda}{\gamma n} \sqrt{\sum_{i=1}^{n} k(X_i, X_i)} + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

References

The Annals of Statistics
2008, Vol. 36, No. 2, 489–531
DOI: 10.1214/009053607000000839

STATISTICAL PERFORMANCE OF SUPPORT VECTOR MACHINES

BY GILLES BLANCHARD,[1] OLIVIER BOUSQUET AND PASCAL MASSART

Fraunhofer-Insitute FIRST, *Google and Université Paris-Sud*

The support vector machine (SVM) algorithm is well known to the computer learning community for its very good practical results. The goal of the present paper is to study this algorithm from a statistical perspective, using tools of concentration theory and empirical processes.

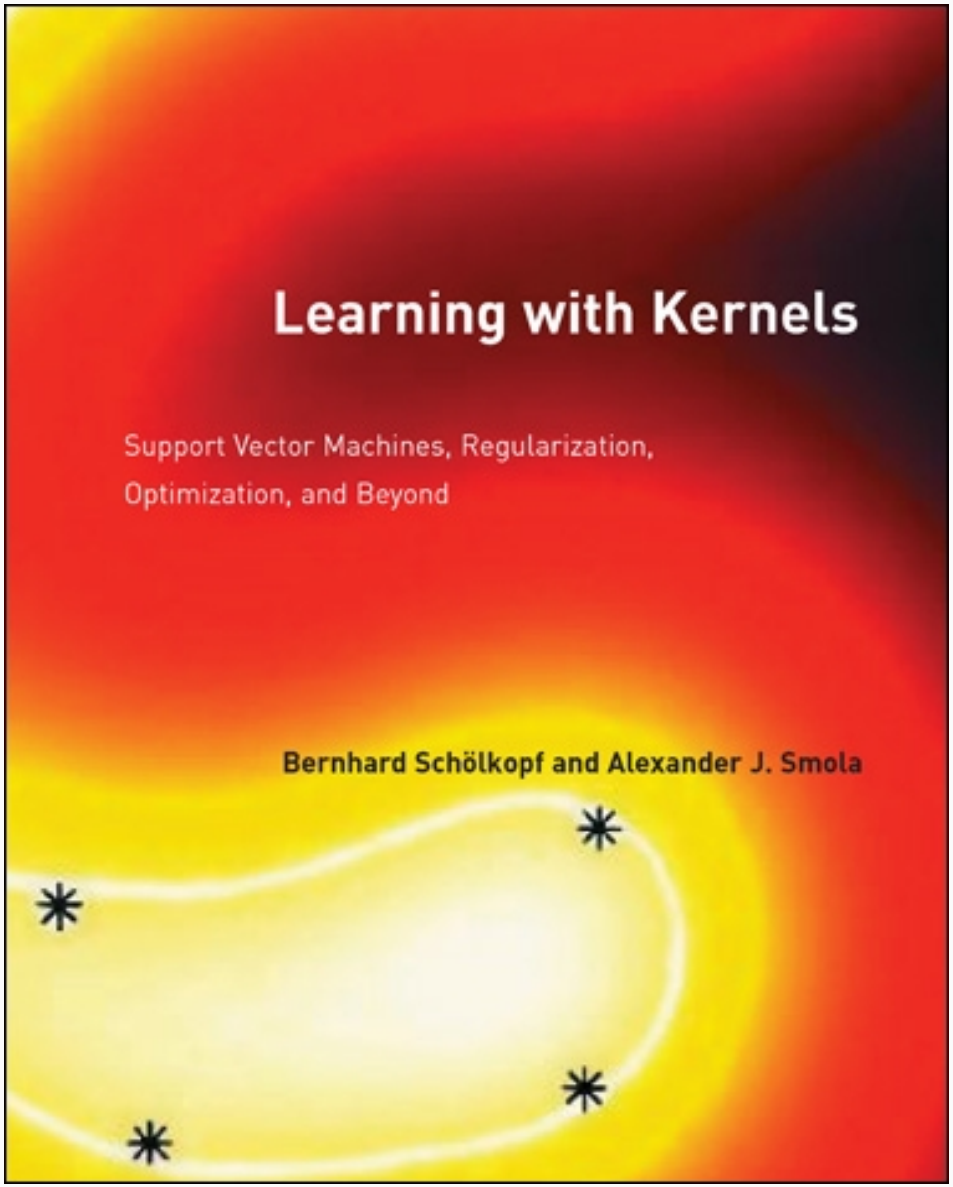
Our main result builds on the observation made by other authors that the SVM can be viewed as a statistical regularization procedure. From this point of view, it can also be interpreted as a model selection principle using a penalized criterion. It is then possible to adapt general methods related to model selection in this framework to study two important points: (1) what is the minimum penalty and how does it compare to the penalty actually used in the SVM algorithm; (2) is it possible to obtain "oracle inequalities" in that setting, for the specific loss function used in the SVM algorithm? We show that the answer to the latter question is positive and provides relevant insight to the former. Our result shows that it is possible to obtain fast rates of convergence for SVMs.

1. Introduction. The success of the support vector machine (SVM) algorithm for pattern recognition is probably mainly due to the number of remarkable experimental results that have been obtained in very diverse domains of application. The algorithm itself can be written as a nice convex optimization problem for which there exists a unique optimum, except in rare degenerate cases. It can also be expressed as the minimization of a regularized functional where the regularizer is the squared norm in a Hilbert space of functions on the input space. Although these are nice mathematical formulations, quite amenable to analysis, the statistical behavior of this algorithm remains only partially understood. Our goal in this work is to investigate the properties of the SVM algorithm in a statistical setting.

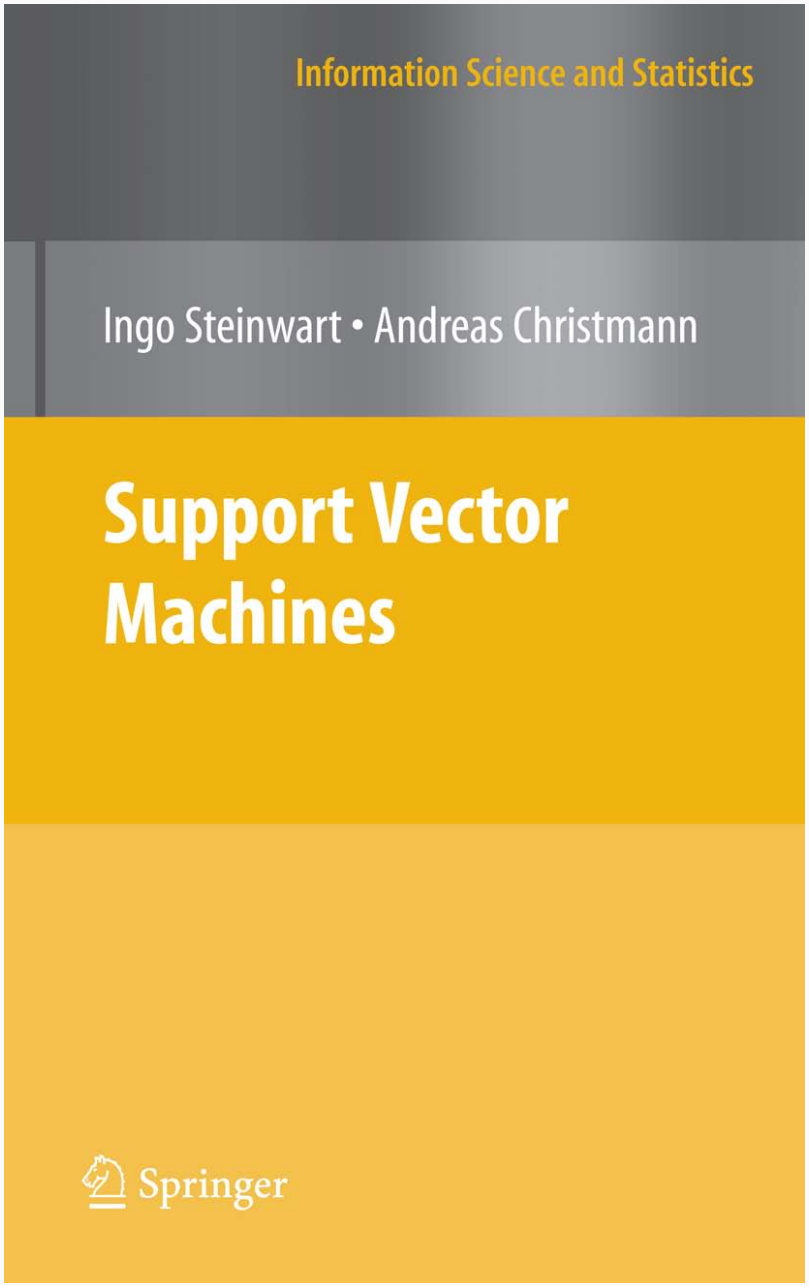
1.1. *The abstract classification problem and convex loss approximation.* We consider a generic (binary) classification problem, defined by the following setting: assume that the product $\mathcal{X} \times \mathcal{Y}$ is a measurable space endowed with an unknown probability measure P, where $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{X}$ is called the input space. The pair (X, Y) denotes a random variable with values in $\mathcal{X} \times \mathcal{Y}$ distributed according to P. We will denote P_X the marginal distribution of variable X. We observe a

Received October 2006; revised April 2007.
[1]Supported in part by a grant of the Humboldt Foundation.
AMS 2000 subject classifications. 62G05, 62G20.
Key words and phrases. Classification, support vector machine, model selection, oracle inequality.

References


Learning with Kernels
Support Vector Machines, Regularization,
Optimization, and Beyond
Bernhard Schölkopf and Alexander J. Smola

References

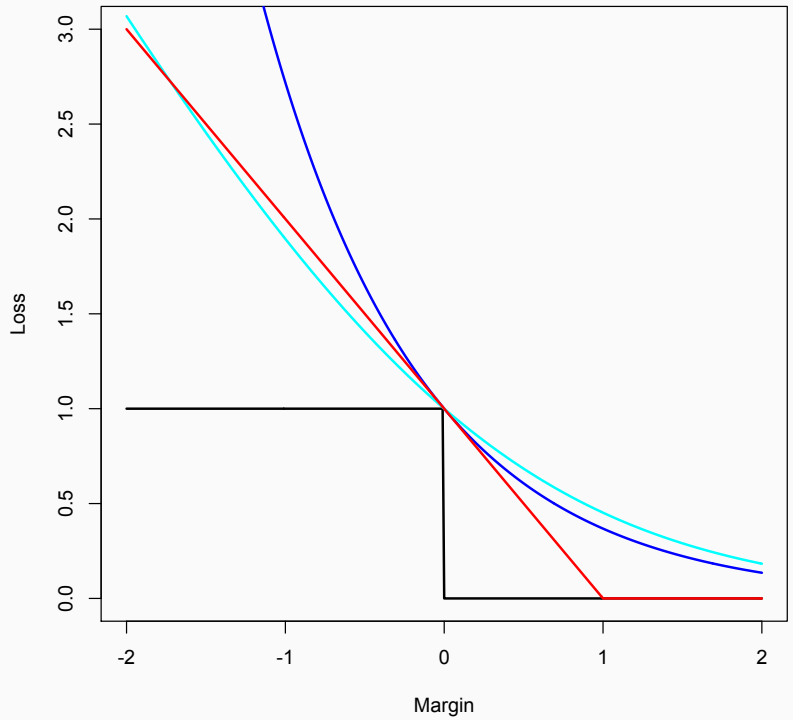

Information Science and Statistics
Ingo Steinwart • Andreas Christmann
Support Vector Machines
Springer

Convex loss functions

Convex surrogates

- Error probability: $\mathbb{E}\mathbb{1}_{[Yf(X)<0]} \leq L(g_f) \leq \mathbb{E}\mathbb{1}_{[Yf(X)\leq 0]}$.
- Idea: smooth the indicator by a convex function of $Yf(X)$.
- $\phi : \mathbb{R} \to \mathbb{R}_+$ a loss function such that ϕ is strictly decreasing, strictly convex, of class C^1, $\phi(0) = 1$, and $\lim_{x\to\infty} \phi(x) = 0$.
- Risk functionals: $A(f) = \mathbb{E}\phi(Yf(X))$ and $A_n(f) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i))$.
- Typical choices:

  - ▷ Exponential: $\phi(x) = e^{-x}$
  - ▷ Logit: $\phi(x) = \log_2(1 + e^{-x})$
  - ▷ Hinge: $\phi(x) = \max(1-x, 0)$!!!!



Convex surrogates, cont.

- What is $f^\star = \arg\min_f A(f)$ for ϕ strictly convex and differentiable?
- Clearly, $\mathbb{E}(\phi(Yf(X)) \,|\, X = x) = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x))$.
- Consequence: $f^\star(x) = \arg\min_\alpha h_{\eta(x)}(\alpha)$, where

$$h_\eta(\alpha) = \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha), \quad \eta \in [0, 1].$$

- Note: h_η is strictly convex and therefore $f^\star$ is well defined.

Convex surrogates, cont.

- The minimum is achieved for $h'_\eta(\alpha) = 0$, that is, when

$$\frac{\eta}{1-\eta} = \frac{\phi'(-\alpha)}{\phi'(\alpha)}.$$

- Since ϕ' is strictly increasing, the solution is positive if and only if $\eta > 1/2$.
- Conclusion: $g^\star(x) \stackrel{\text{def}}{=} 2\mathbb{1}_{[f^\star(x)>0]} - 1$ is the Bayes classifier!
- Examples:
  - $\phi(x) = e^{-x} \to f^\star(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)})$;
  - $\phi(x) = \log_2(1 + e^{-x}) \to f^\star(x) = \log(\frac{\eta(x)}{1-\eta(x)})$.
- For the hinge loss: $f^\star(x) = 2\mathbb{1}_{[\eta(x)>1/2]} - 1 =$ Bayes classifier itself!

Connection between 0-1-risk and ϕ-risk

- Objective: connect $L(f) - L^\star$ with $A(f) - A^\star$.
- Tool: $H : [0,1] \to \mathbb{R}$ defined by $H(\eta) = \inf_\alpha h_\eta(\alpha)$.

Lemma

Let ϕ be a convex loss function such that the following hold:

(i) $f^\star(x) > 0$ *if and only if* $\eta(x) > 1/2$;

(ii) *There exist constants* $c \geq 0$ *and* $s \geq 1$ *satisfying*

$$\left|\frac{1}{2} - \eta\right|^s \leq c^s(1 - H(\eta)), \quad \eta \in [0,1].$$

Then, for any function $f : \mathbb{R}^d \to \mathbb{R}$,

$$L(f) - L^\star \leq 2c(A(f) - A^\star)^{1/s}.$$

- ▷ Exponential: $H(\eta) = 2\sqrt{\eta(1-\eta)}$.
- ▷ Logit: $H(\eta) = -\eta \log_2 \eta - (1-\eta)\log_2(1-\eta)$.
- ▷ In both cases: $c = 1/\sqrt{2}$ and $s = 2$.
- ▷ Hinge: $H(\eta) = 2\min(\eta, 1-\eta) \to c = 1/2$ and $s = 1$.

Excess risk of convex minimizers

Context: $\mathscr{C}$ = a class of ± 1 base classifiers, and

$$\mathscr{F}_\lambda = \Big\{ f = \sum_{j=1}^{N} c_j g_j : N \in \mathbb{N}, g_1, \ldots, g_N \in \mathscr{C}, \sum_{j=1}^{N} |c_j| = \lambda \Big\}.$$

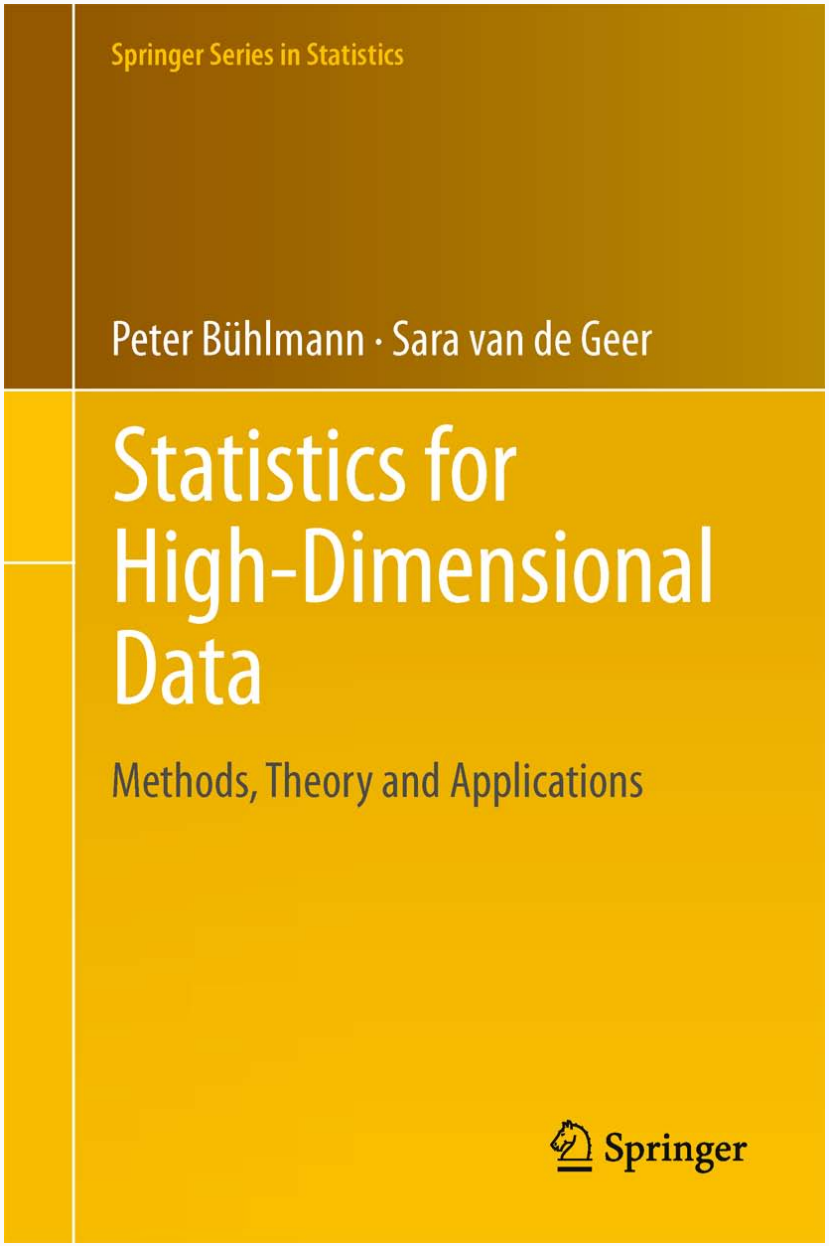
Theorem

Let $f_n^\star \in \arg\min_{f \in \mathscr{F}_\lambda} A_n(f)$, using either the exponential or the logit loss function, and let $\delta \in (0,1)$. Then, with probability at least $1-\delta$,

$$L(f_n^\star) - L^\star \leq 2\left(8 L_\phi \lambda \sqrt{\frac{V_{\mathscr{C}} \log(n+1)}{n}} + B \sqrt{\frac{\log(1/\delta)}{2n}} \right)^{1/2} + \sqrt{2} \big(\inf_{f \in \mathscr{F}_\lambda} A(f) - A^\star \big)^{1/2}.$$

- Exponential: $B = e^\lambda$ and $L_\phi = e^\lambda$.
- Logit: $B = \log_2(1 + e^\lambda)$ and $L_\phi = 1/\log 2$.
- If $\inf_{f \in \mathscr{F}_\lambda} A(f) - A^\star = 0$, then $L(f_n^\star) - L^\star = \mathrm{O}\big(\big(\frac{\log n}{n} \big)^{1/4} \big)$.
- The exponent in the rate of convergence is dimension-free!
- Convex optimization $\rightarrow$ boosting algorithms.

Reference


Springer Series in Statistics
Peter Bühlmann · Sara van de Geer
Statistics for High-Dimensional Data
Methods, Theory and Applications
Springer

Model selection

Oracle inequalities

- Choosing the right set $\mathscr{C}$ of possible classifiers is a key to success.
- If $\mathscr{C}$ is too large: overfitting risk.
- If $\mathscr{C}$ is too small: unable to approximate complex decisions.
- Welcome to the domain of model selection.
- A possibly infinite collection of models $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_k, \ldots$
- For each model: empirical risk minimizer $g^\star_{n,k} \in \mathscr{C}_k$.
- Model selection problem: select the best among the $(g^\star_{n,k})_k$.
- Oracle inequality:

$$\mathbb{E}L(g^\star_{n,\hat{k}}) - L^\star \leq C \inf_k \left(\mathbb{E}L(g^\star_{n,k}) - L^\star_k + L^\star_k - L^\star + \gamma(k,n) \right),$$

where $L^\star_k = \inf_{g \in \mathscr{C}_k} L(g)$.

Data-driven penalization

- Penalty-based model selection: choose $\hat{k}$ that minimizes

$$L_n(g^\star_{n,k}) + \text{pen}(n,k).$$

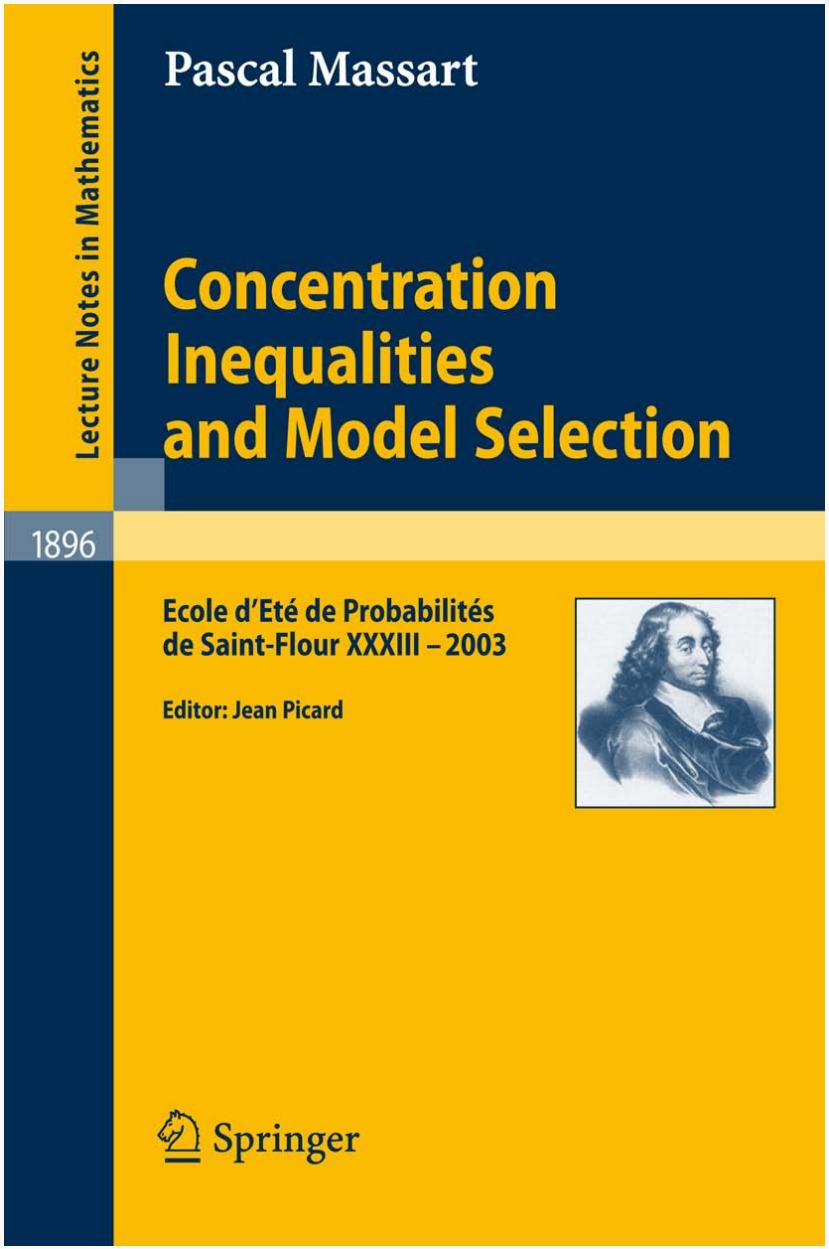
- The penalty should estimate the amount of overfitting

$$\mathbb{E} \sup_{g \in \mathscr{C}_k} |L_n(g) - L(g)|.$$

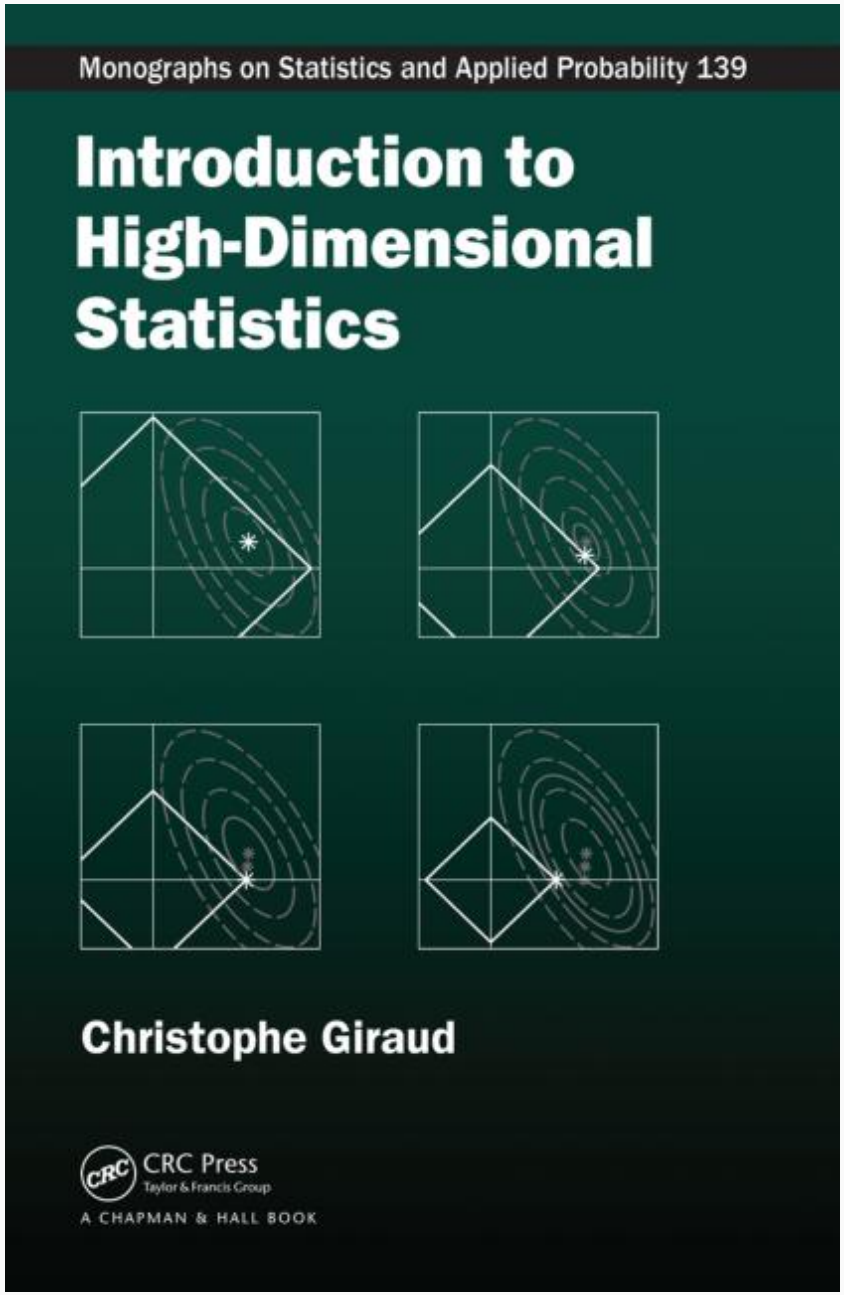
- Distribution-free penalties $\rightarrow$ VC dimension $V_{\mathscr{C}_k} \rightarrow$ too conservative.
- Data-driven penalties $\rightarrow$ Rademacher averages.

Model selection is a science.

References


Lecture Notes in Mathematics
Pascal Massart
Concentration Inequalities and Model Selection
1896
Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003
Editor: Jean Picard
Springer

References


Monographs on Statistics and Applied Probability 139
Introduction to High-Dimensional Statistics
Christophe Giraud
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Oracle inequality

Theorem

Let $\mathrm{pen}(n,k)$ *be defined by*

$$\mathrm{pen}(n,k) = 2R_n(\mathscr{F}_k(X_1^n)) + 4\sqrt{\frac{\log k}{n}}.$$

Then

$$\mathbb{E}L(g^\star_{n,\hat{k}}) - L^\star \leq \inf_k \left(4\mathbb{E}R_n(\mathscr{F}_k(X_1^n)) + L_k^\star - L^\star + 4\sqrt{\frac{\log k}{n}} \right) + \frac{\sqrt{2}\pi^{5/2}}{3\sqrt{n}}.$$

- ▷ Optimal constant in front of the infimum.
- ▷ Satisfactory in the general case.

Stone's theorem

Plug-in principle

- Starting point:

$$g^\star(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

- Idea: estimate $r(x)$ from the training data $\mathscr{D}_n \rightsquigarrow r_n(x)$.
- Plug-in classifier:

$$g_n(x) = \begin{cases} 1 & \text{if } r_n(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

- Question 1: connection $r_n \leftrightarrow L(g_n)$?
- Question 2: which choice for r_n?

Plug-in = regression estimation problem.

Connection $r_n \leftrightarrow L(g_n)$

Theorem (Classification and regression)

Let r_n be a regression function estimate of r, and let g_n be the corresponding plug-in classifier. Then

$$0 \leq L(g_n) - L^\star \leq 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu(\mathrm{d}x).$$

In particular, for all $p \geq 1$,

$$0 \leq L(g_n) - L^\star \leq 2 \Big(\int_{\mathbb{R}^d} |r_n(x) - r(x)|^p \mu(\mathrm{d}x) \Big)^{1/p},$$

and

$$0 \leq \mathbb{E} L(g_n) - L^\star \leq 2\, \mathbb{E}^{1/p} |r_n(X) - r(X)|^p.$$

- Take-home message:

$$\mathbb{E} \int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(\mathrm{d}x) \to 0$$

implies that the corresponding plug-in classifier g_n is consistent.

Local averaging estimates

- Definition: $r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$.
- Weight vector: $(W_{n1}(x), \ldots, W_{nn}(x))$.
- Interpretation: X_i "close" to x should provide more information.
- Often (but not always) $(W_{n1}(x), \ldots, W_{nn}(x))$ is a probability vector.
- Important: each $W_{ni}(x)$ is a function of x and $X_1, \ldots, X_n$ (not $Y_1, \ldots, Y_n$).
- Equivalently: $r_n(x) = \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{[Y_i=1]}$.
- Companion plug-in classifier:

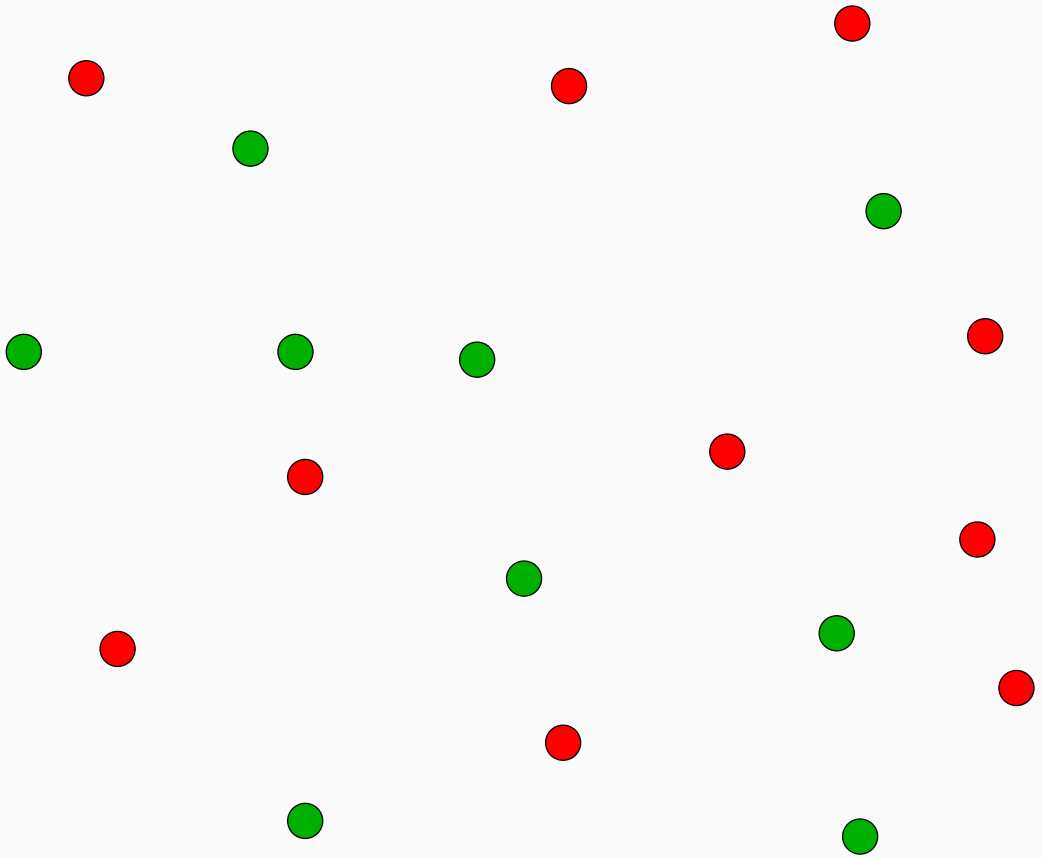
$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n W_{ni}(x) Y_i > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

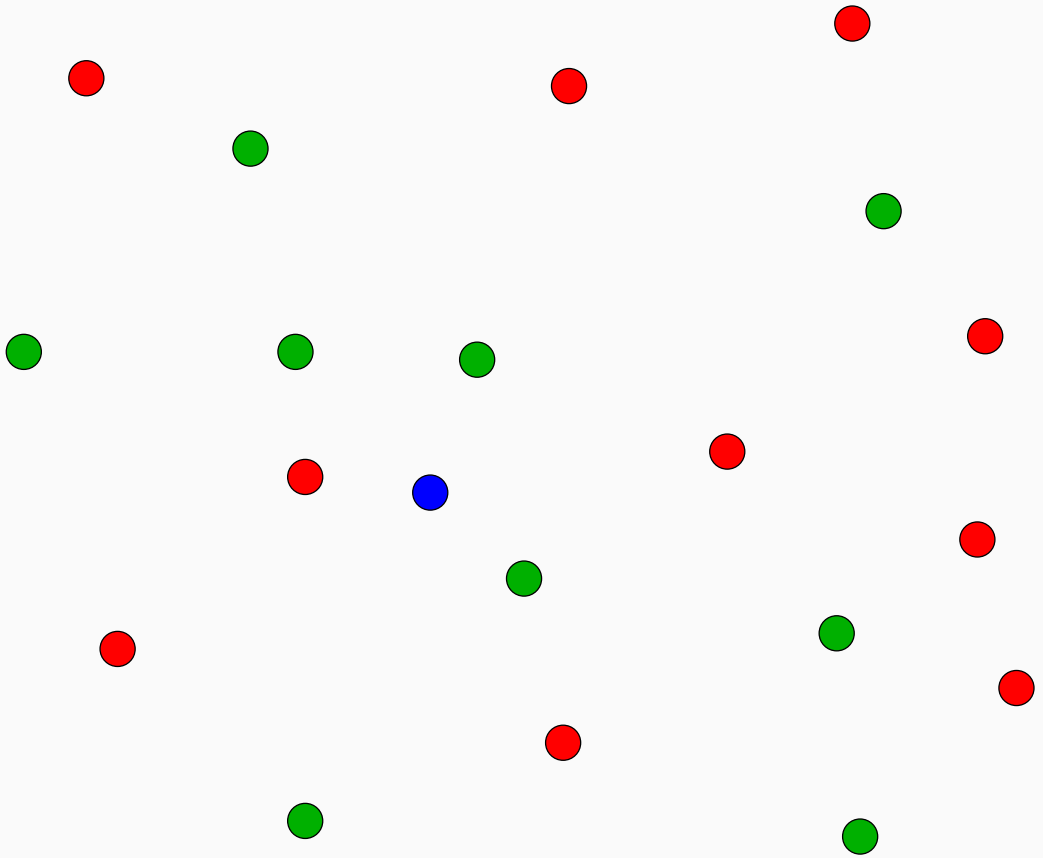
- Whenever $\sum_{i=1}^n W_{ni}(x) = 1$:

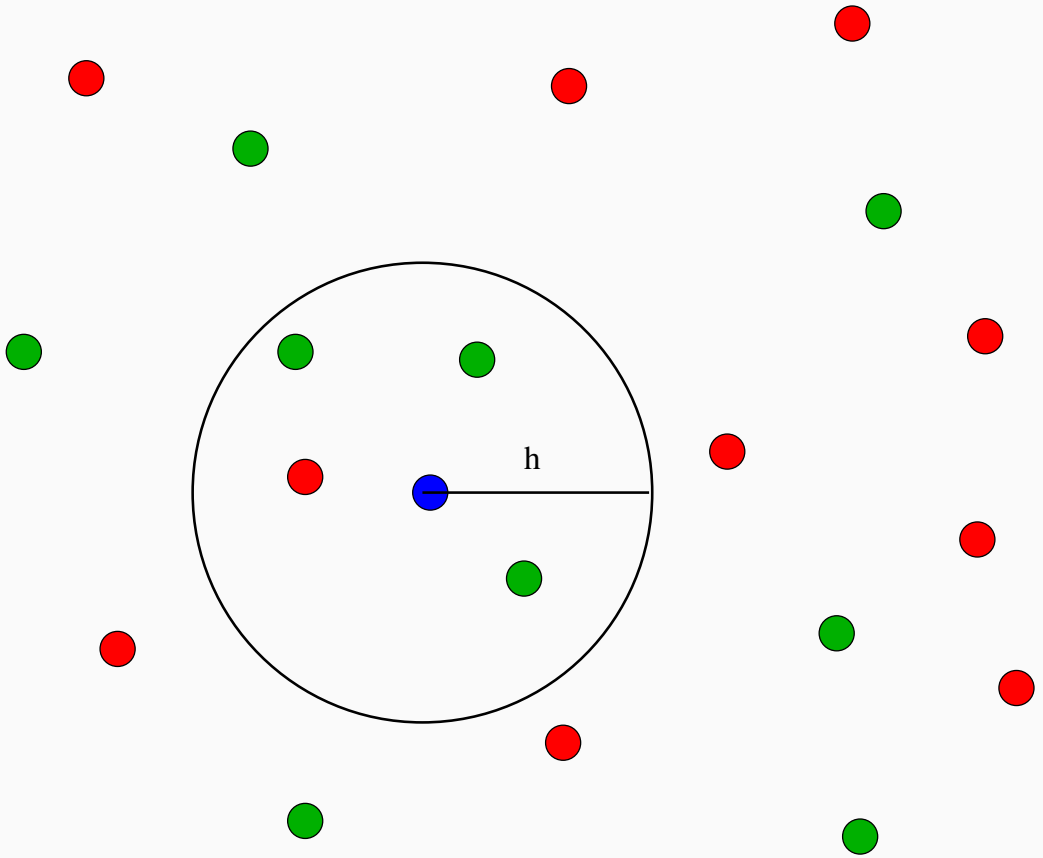
$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{[Y_i=1]} > \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{[Y_i=0]} \\ 0 & \text{otherwise.} \end{cases}$$

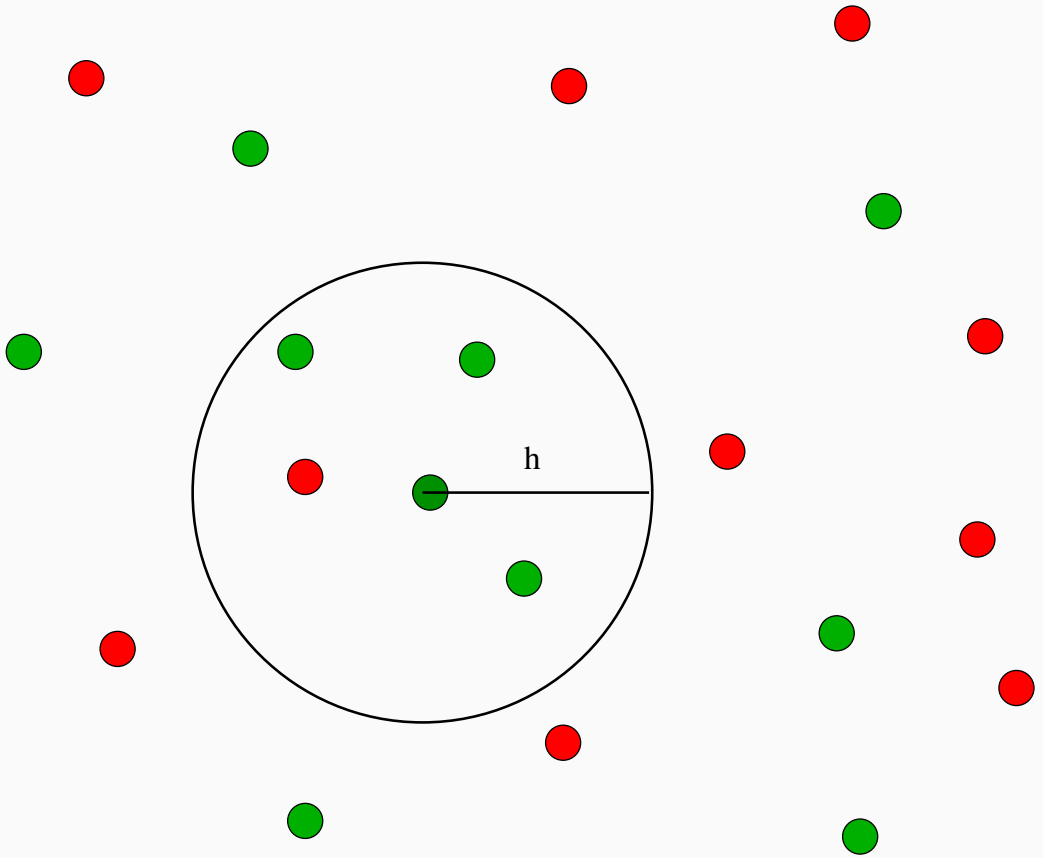
Example 1: kernel estimate

- Definition:
$$r_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}.$$
- Kernel K: a nonnegative real-valued function on $\mathbb{R}^d$.
- Bandwidth h: a positive number (may depend on n).
- Weights:
$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}.$$
- If both denominator and numerator are zero: $r_n(x) = \frac{1}{n}\sum_{i=1}^n Y_i$.
- Kernels:
  - ▷ Naive: $K(z) = \mathbb{1}_{[\|z\| \le 1]}$,
$$r_n(x) = \frac{\sum_{i=1}^n \mathbb{1}_{[\|x-X_i\| \le h]} Y_i}{\sum_{j=1}^n \mathbb{1}_{[\|x-X_j\| \le h]}};$$
  - ▷ Epanechnikov: $K(z) = (1 - \|z\|^2)1_{[\|z\| \le 1]}$;
  - ▷ Gaussian: $K(z) = e^{-\|z\|^2}$.






h


h

Example 2: nearest neighbor (NN) estimate

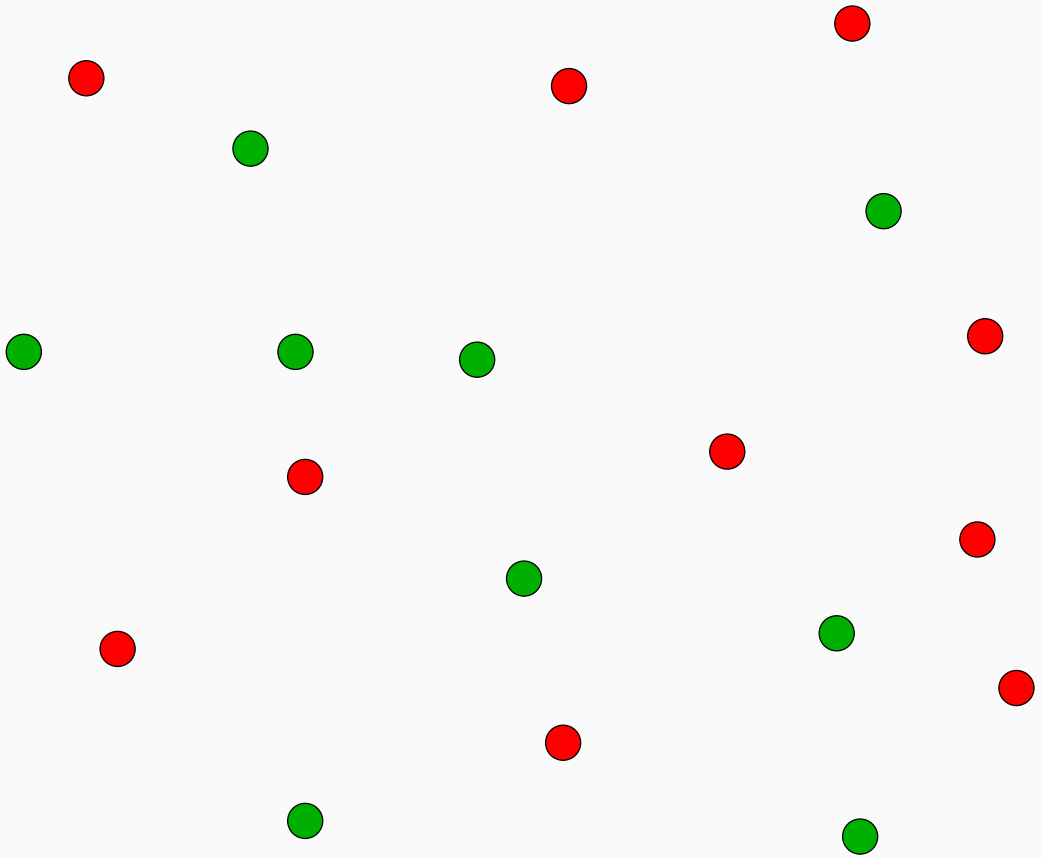
- Definition:
  - ▷ $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ reordering of $\mathscr{D}_n$ according to

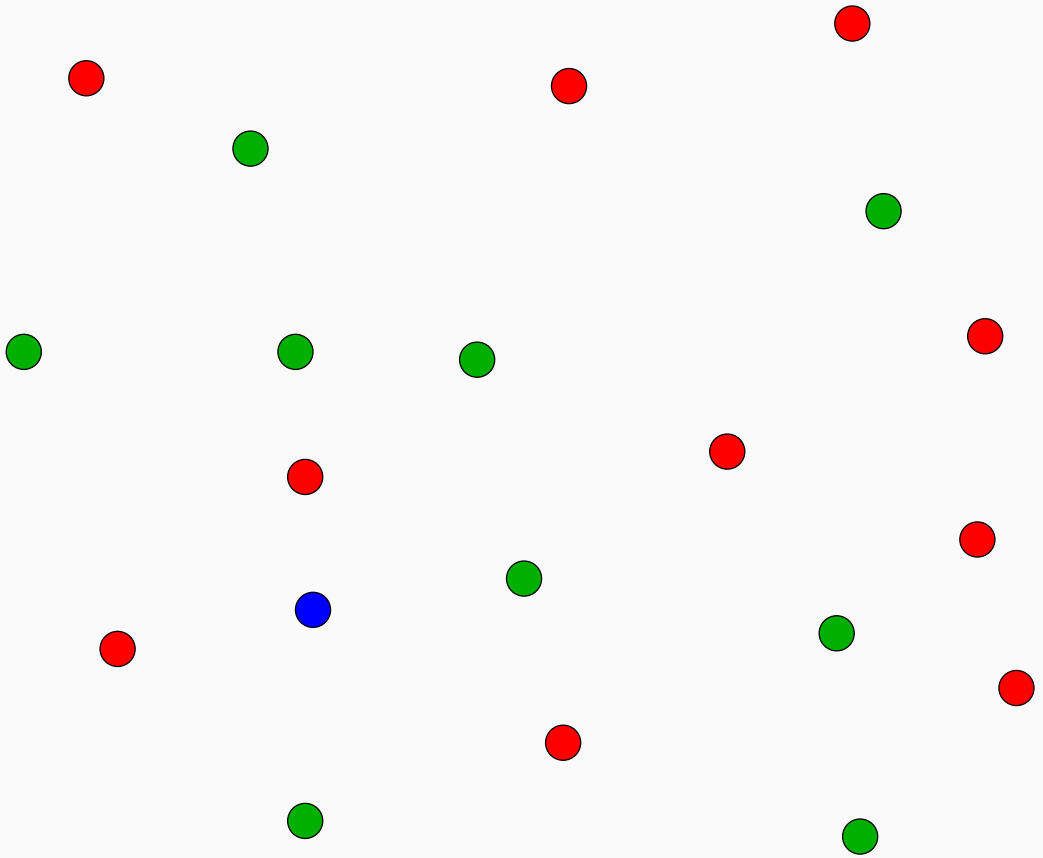
$$\|X_{(1)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|;$$

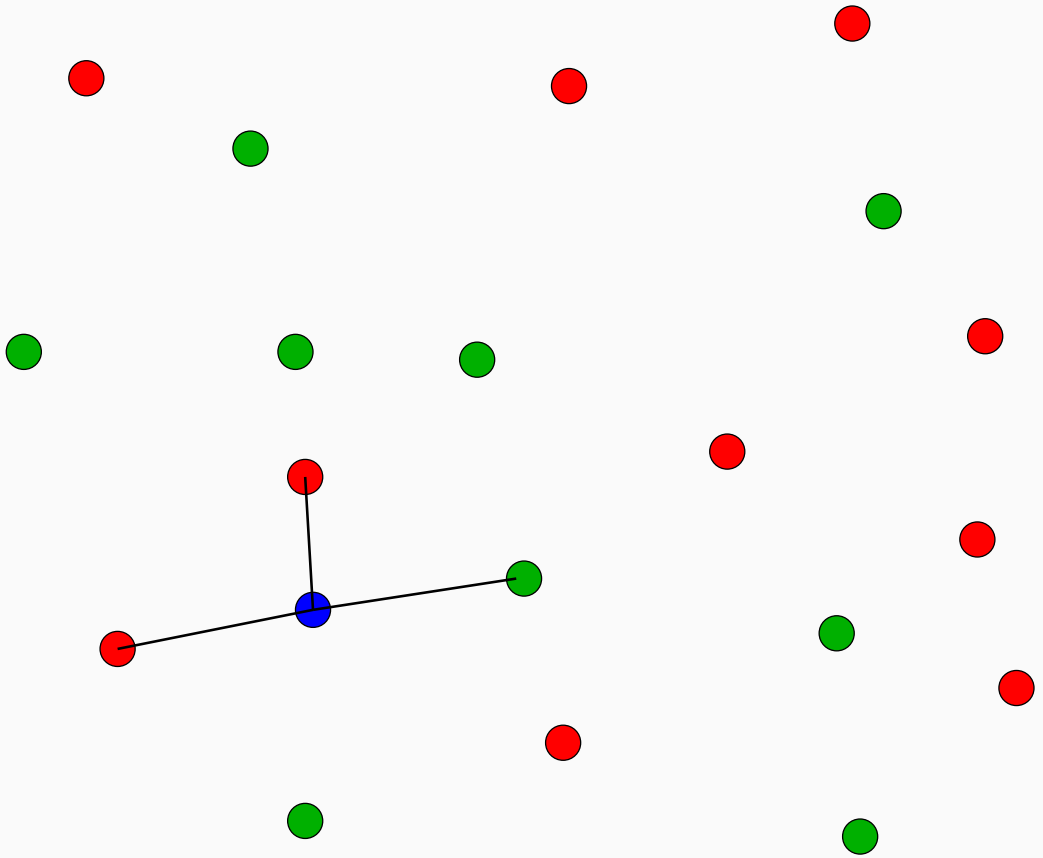
  - ▷ Whenever $\|X_i - x\| = \|X_j - x\|$ and $i < j$, we declare X_i closer to x;
  - ▷ NN estimate: $r_n(x) = \sum_{i=1}^n v_{ni} Y_{(i)}(x)$, where $\sum_{i=1}^n v_{ni} = 1$.
- $(\Sigma_1, \ldots, \Sigma_n)$: permutation of $(1, \ldots, n)$ such that X_i is the Σ_i-th nearest neighbor of x for all i.
- Local averaging: $r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$, where $W_{ni}(x) = v_{n\Sigma_i}$.
- k-NN estimate:

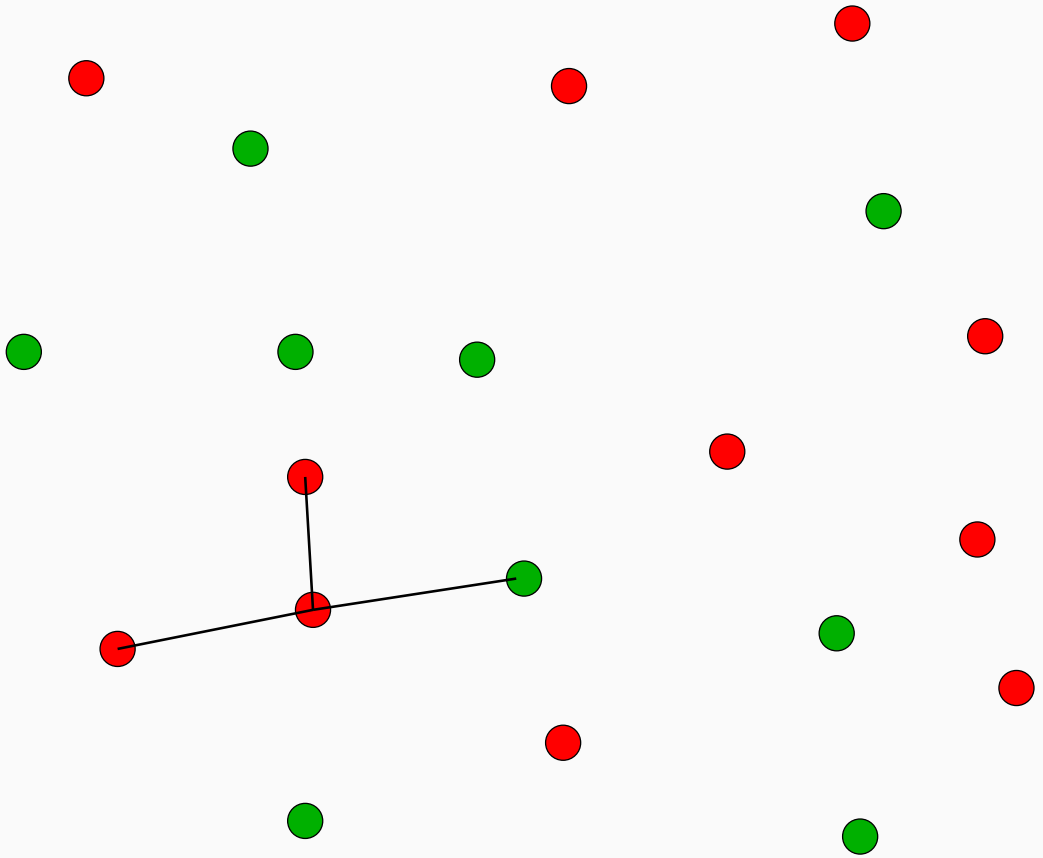
$$v_{ni} = \begin{cases} \frac{1}{k} & \text{for } 1 \leq i \leq k \\ 0 & \text{for } k < i \leq n. \end{cases}$$

- To keep in mind: $r_n(x) = \frac{1}{k} \sum_{i=1}^k Y_{(i)}(x)$.









Stone's theorem

$r_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$, with $(W_{n1}(x), \ldots, W_{nn}(x))$ a probability vector.

Theorem (Stone's theorem)

Assume that for any distribution of X, the weights satisfy the following conditions:

(*i*) *There is a constant C such that, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|f(X)| < \infty$,*

$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)|f(X_i)|\Big) \le C\mathbb{E}|f(X)| \quad \text{for all } n \ge 1.$$

(*ii*) *For all $a > 0$,*

$$\mathbb{E}\Big(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{[\|X_i - X\| > a]}\Big) \to 0.$$

(*iii*) *One has*

$$\mathbb{E} \max_{1 \le i \le n} W_{ni}(X) \to 0.$$

Then the corresponding plug-in classifier g_n is universally consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$ for all distributions of (X, Y).

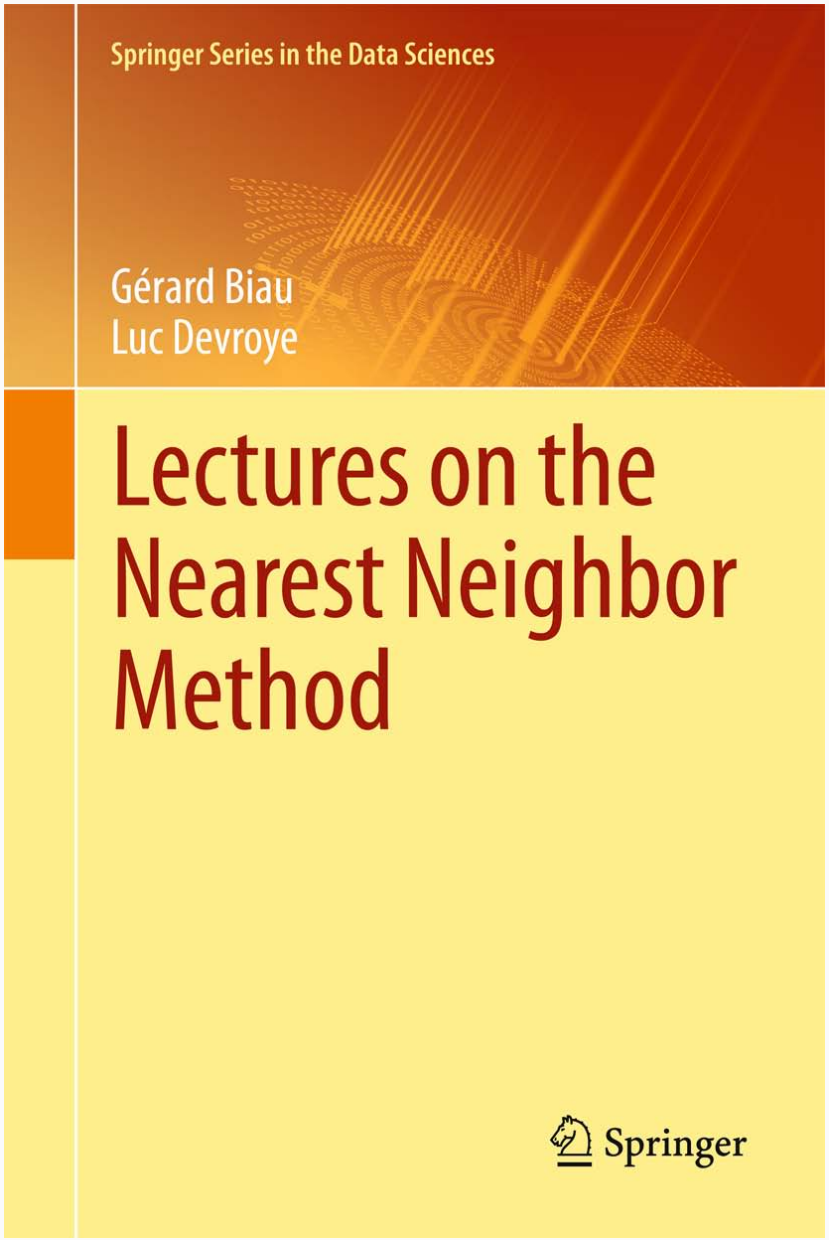
Discussion

- Condition (i) is merely technical.
- Condition (ii) ensures that $r_n(X)$ is asymptotically mostly influenced by the data points close to X.
- Condition (iii) states that asymptotically all weights become small.
- No single observation has a too large contribution to the estimate.
- The number of points in the averaging must tend to infinity.

Charles Stone

k-nearest neighbor classifiers

Reference


Springer Series in the Data Sciences
Gérard Biau
Luc Devroye
Lectures on the Nearest Neighbor Method
Springer

Reminder

- Reordering $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ according to

$$\|X_{(1)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|.$$

- Whenever $\|X_i - x\| = \|X_j - x\|$ and $i < j$, we declare X_i closer to x.
- k-NN regression function estimate: $r_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x)$.
- k-NN classifier:

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{k} \mathbb{1}_{[Y_{(i)}(x)=1]} > \sum_{i=1}^{k} \mathbb{1}_{[Y_{(i)}(x)=0]} \\ 0 & \text{otherwise.} \end{cases}$$

If X has a density, then there is no distance tie.

Universal consistency

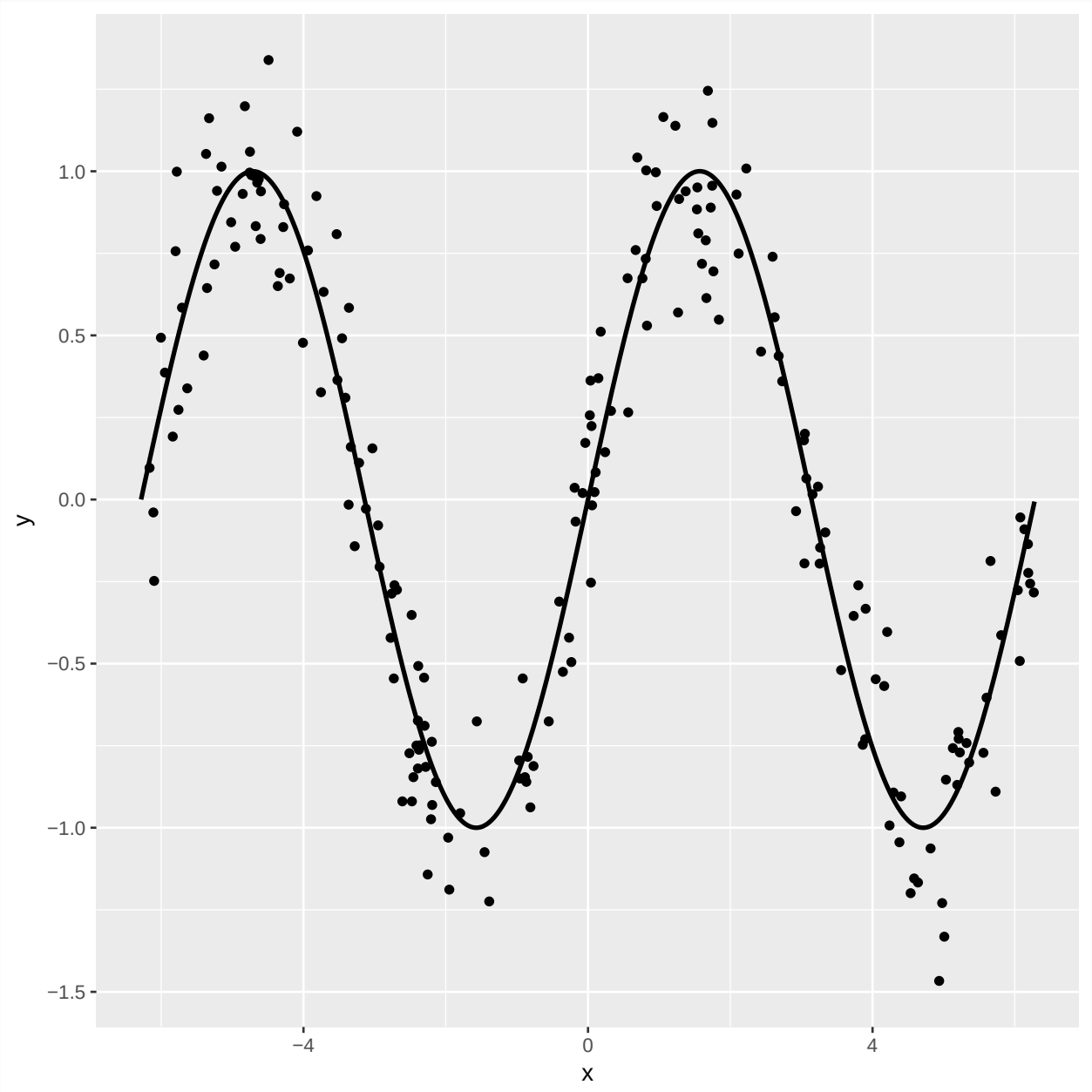
Theorem

Assume that $k \to \infty$ and $k/n \to 0$. Then the k-NN classifier is universally consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$ for all distributions of (X, Y).

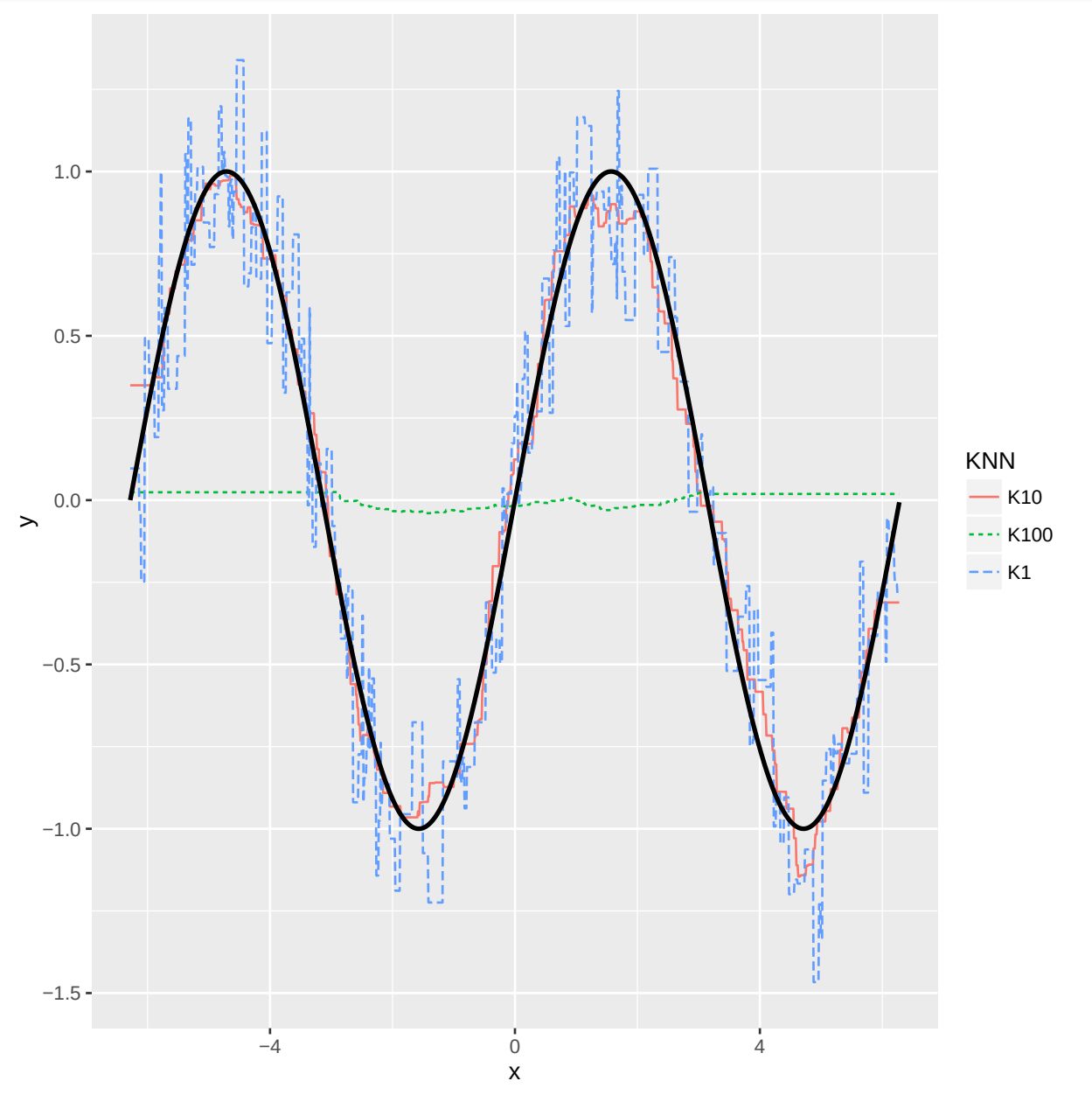
- ▷ Proof's agenda: verify Stone's conditions (i)-(iii).
- ▷ Simplification: distance ties $\|X_i - X\| = \|X_j - X\|$ occur with zero probability.

k is large but small with respect to n: bias/variance compromise.

Choice of k



Choice of k



Properties of the kth nearest neighbor

- The support of μ is defined by

$$\mathrm{supp}(\mu) = \left\{x \in \mathbb{R}^d : \mu(B(x, \rho)) > 0 \text{ for all } \rho > 0\right\}.$$

- Properties:
  - $\mathrm{supp}(\mu)$ is a closed set;
  - $\mathrm{supp}(\mu)$ is the smallest closed subset of $\mathbb{R}^d$ of μ-measure one;
  - One has $\mathbb{P}(X \in \mathrm{supp}(\mu)) = 1$.

Lemma

If $x \in \mathrm{supp}(\mu)$ and $k/n \to 0$, then

$$\|X_{(k)}(x) - x\| \to 0 \quad \textit{almost surely.}$$

Combinatorial lemmas

Lemma

Let ν be a probability measure on $\mathbb{R}^d$. Fix $x' \in \mathbb{R}^d$ and let, for $a \geq 0$,

$$B_a(x') = \Big\{ x \in \mathbb{R}^d : \nu\big(B(x, \|x' - x\|)\big) \leq a \Big\}.$$

Then

$$\nu(B_a(x')) \leq \gamma_d a,$$

where γ_d is a positive constant depending only upon d.

Corollary

If distance ties occur with zero probability, then

$$\sum_{i=1}^{n} \mathbb{1}_{[X \text{ is among the } k\text{-NN of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}]} \leq k\gamma_d,$$

with probability one.

Stone's lemma

Lemma (Stone's lemma)

Assume that distance ties occur with zero probability. Then, for every Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ *such that* $\mathbb{E}|f(X)| < \infty$,

$$\sum_{i=1}^{k} \mathbb{E}\left|f(X_{(i)}(X))\right| \leq k\gamma_d \mathbb{E}|f(X)|,$$

where γ_d *is a positive constant depending only upon* d.

Consistency of the k-NN classifier

- To do: verify Stone's conditions with $W_{ni}(x) = 1/k$ if X_i is among the k nearest neighbors of x and $W_{ni}(x) = 0$ otherwise.
- Condition (*iii*) is clear since $k \to \infty$.
- Condition (*ii*): note that

$$\mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{[\|X_i - X\|]>a}\Big) = \mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{[\|X_{(i)}(X) - X\|]>a}\Big).$$

So,

$$\mathbb{P}(\|X_{(k)}(X) - X\| > a) \to 0 \Rightarrow \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{[\|X_i - X\|]>a}\Big) \to 0.$$

But, for all $a > 0$,

$$\mathbb{P}(\|X_{(k)}(X) - X\| > a) = \int_{\mathbb{R}^d} \mathbb{P}(\|X_{(k)}(x) - x\| > a)\mu(\mathrm{d}x).$$

Assuming that $k/n \to 0$, the conclusion follows by the Lebesgue dominated convergence theorem.

Consistency of the k-NN classifier, cont.

- Condition (i): take f such that $\mathbb{E}|f(X)| < \infty$; we have to show that for some constant C

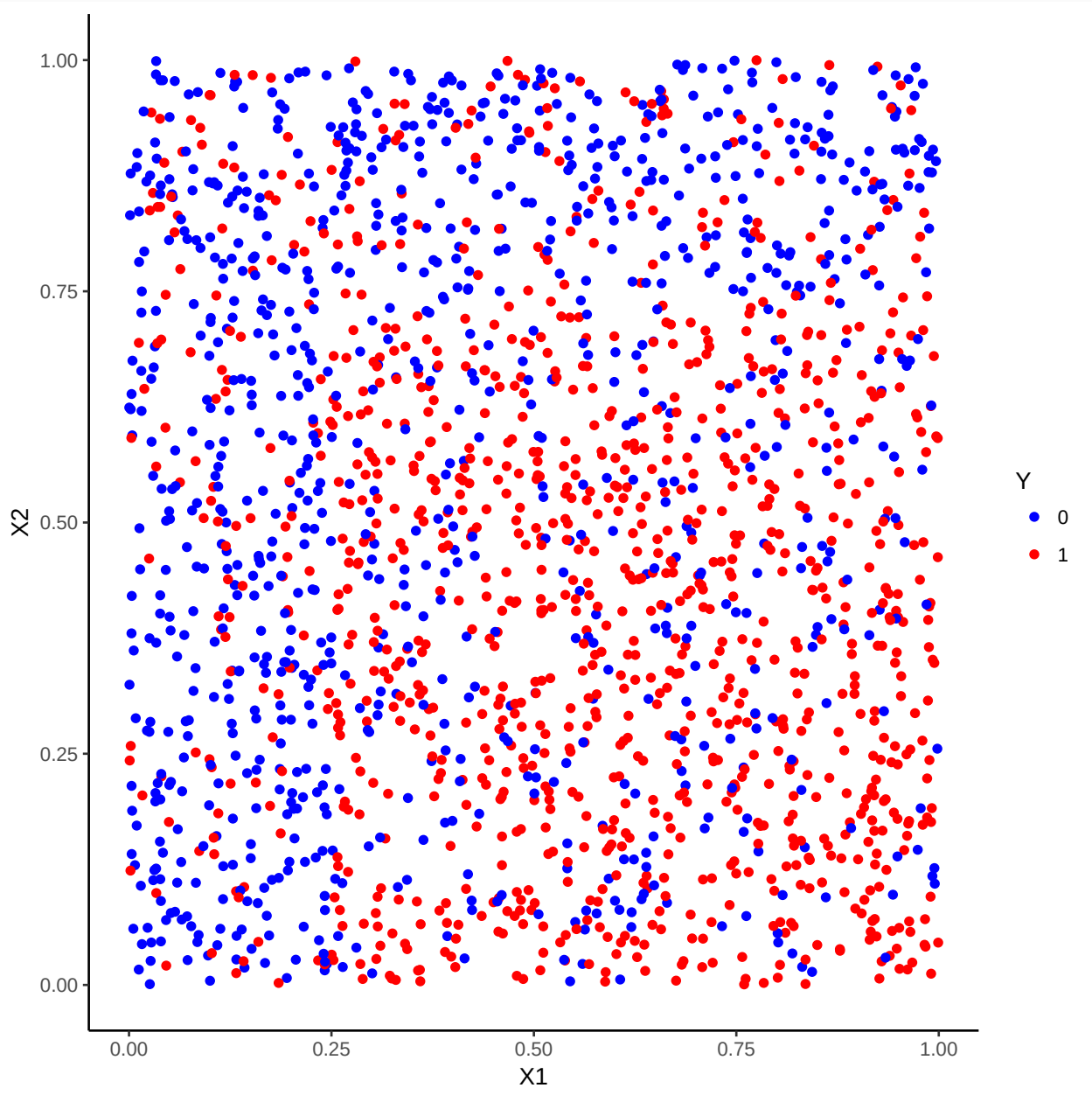
$$\mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{n}|f(X_i)|\mathbb{1}_{[X_i \text{ is among the } k\text{-NN of } X]}\Big) \leq C\mathbb{E}|f(X)|.$$

  Since

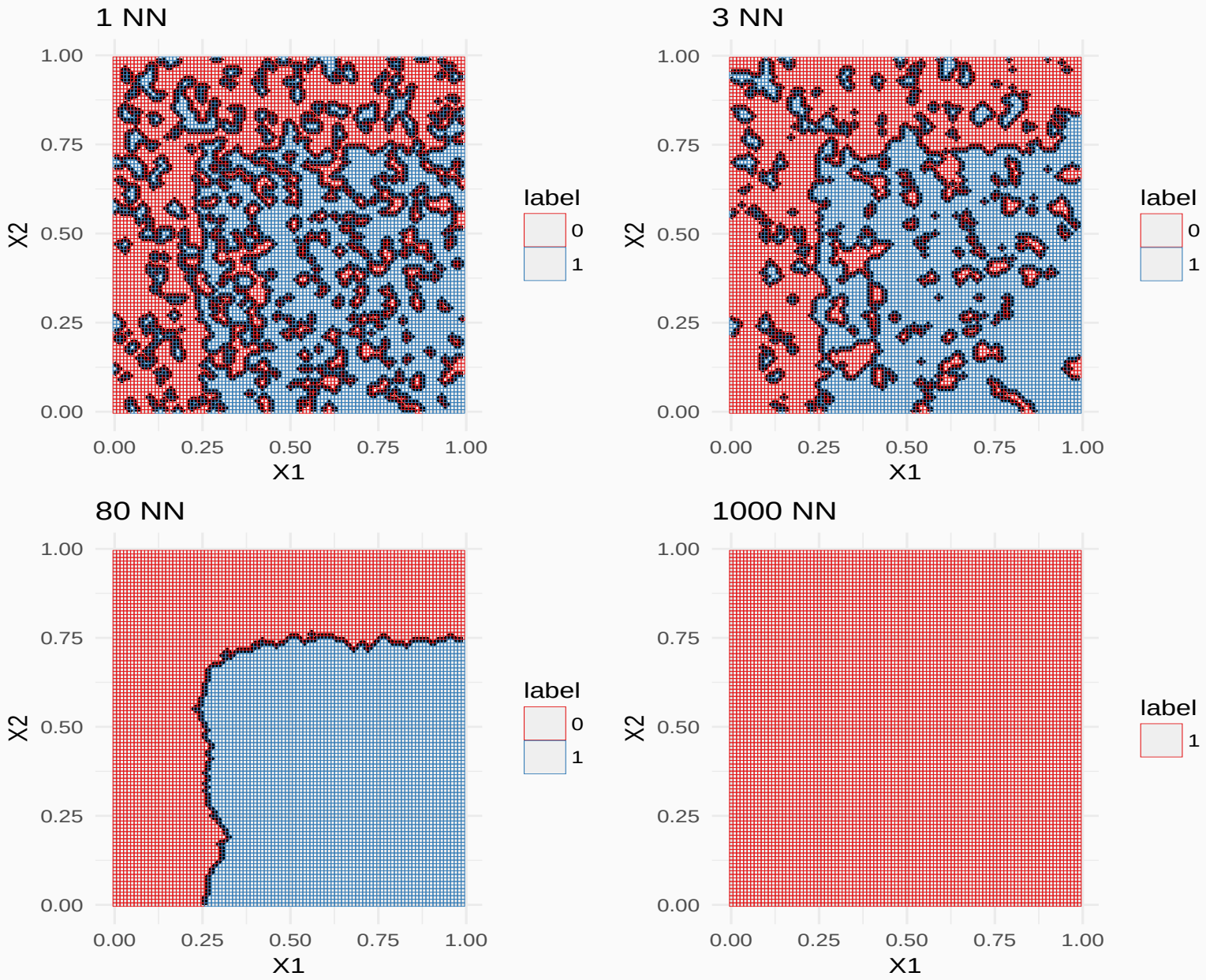
$$\mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{n}|f(X_i)|\mathbb{1}_{[X_i \text{ is among the } k\text{-NN of } X]}\Big) = \mathbb{E}\Big(\frac{1}{k}\sum_{i=1}^{k}\big|f(X_{(i)}(X))\big|\Big),$$

  this is precisely the statement of Stone's lemma. ✔

Example



Bias/Variance compromise



Choice of k by data splitting

- Choosing k: minimizing the empirical error is not a good idea.
- Data splitting:
  - ▷ A training set $\mathscr{D}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$;
  - ▷ A test set $\mathscr{D}_\ell = \{(X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)\}$, with $m + \ell = n$.
- Candidates: $\mathscr{C}_m = \{g_k : 1 \leq k \leq m\} \to k$-NN classifiers using $\mathscr{D}_m$.
- Strategy: choose $g_n^\star \in \mathscr{C}_m$ such that

$$g_n^\star \in \arg\min_{g_k \in \mathscr{C}_m} \frac{1}{\ell} \sum_{i=m+1}^{n} \mathbb{1}_{[g_k(X_i) \neq Y_i]}.$$

Theorem

One has

$$\mathbb{E}\big(L(g_n^\star) - \inf_{g_k \in \mathscr{C}_m} L(g_k)\big) \leq 2\sqrt{\frac{\log(2em)}{2\ell}}.$$

- ▷ The classifier $g_n^\star$ is universally consistent provided

$$\lim_{n\to\infty} m = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{\ell}{\log m} = \infty.$$

Partitioning classifiers and trees

Partitioning classifiers

- Principle: partition $\mathbb{R}^d$ into disjoint cells $A_1, A_2, \ldots$
- Classification by a majority vote in each cell.
- Classifier:

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbb{1}_{[X_i \in A(x)]} \mathbb{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbb{1}_{[X_i \in A(x)]} \mathbb{1}_{[Y_i=0]} \\ 0 & \text{otherwise,} \end{cases}$$

  where $A(x)$ = cell containing x.
- X-property: the partitions depend only on $X_1, \ldots, X_n$ (not $Y_1, \ldots, Y_n$).
- Notation: $\text{diam}(A) = \sup_{(x,y) \in A^2} \|x-y\|$ and $N(x) = \sum_{i=1}^n \mathbb{1}_{[X_i \in A(x)]}$.

Theorem

Let g_n be a partitioning classifier with the X-property. If

(*i*) *$\text{diam}(A(X)) \to 0$ in probability,*

and

(*ii*) *$N(X) \to \infty$ in probability,*

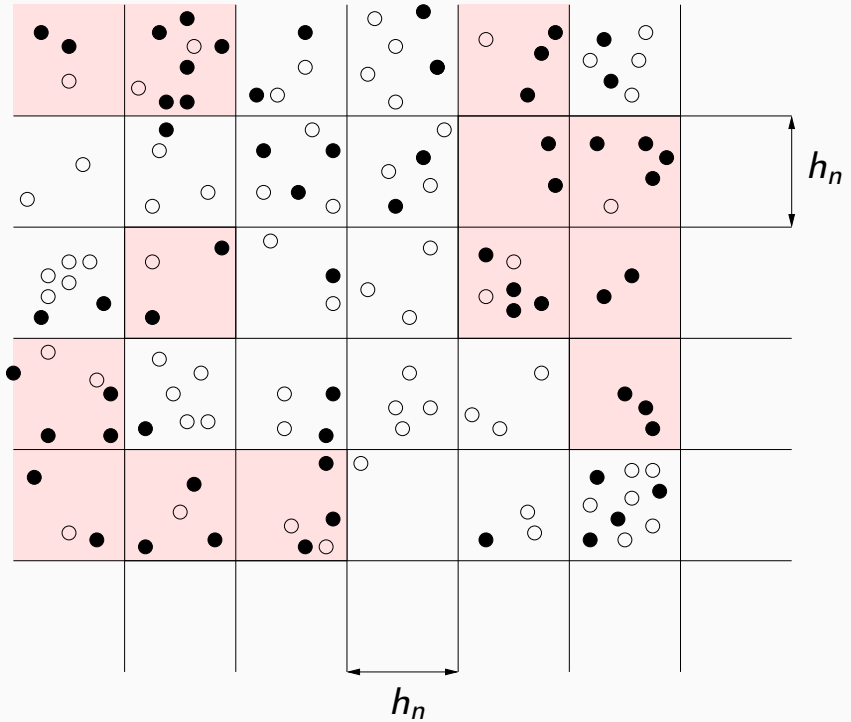
then $\mathbb{E}L(g_n) \to L^\star$.

Example 1: cubic histogram classifier

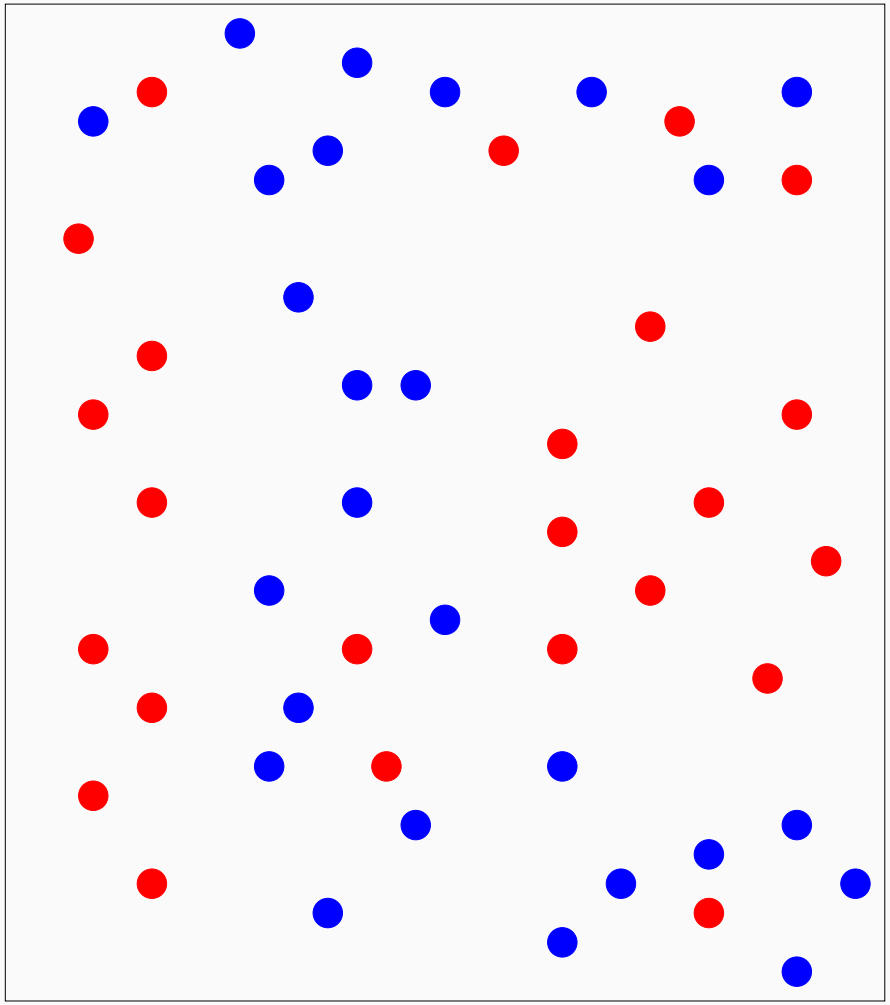
- Definition: $A_{n1}, A_{n2}, \ldots$ a partition of $\mathbb{R}^d$ into cubes of size h.
- So, each cell $= \prod_{j=1}^{d} [k_j h, (k_j + 1)h)$, where the k_j are integers.

Theorem

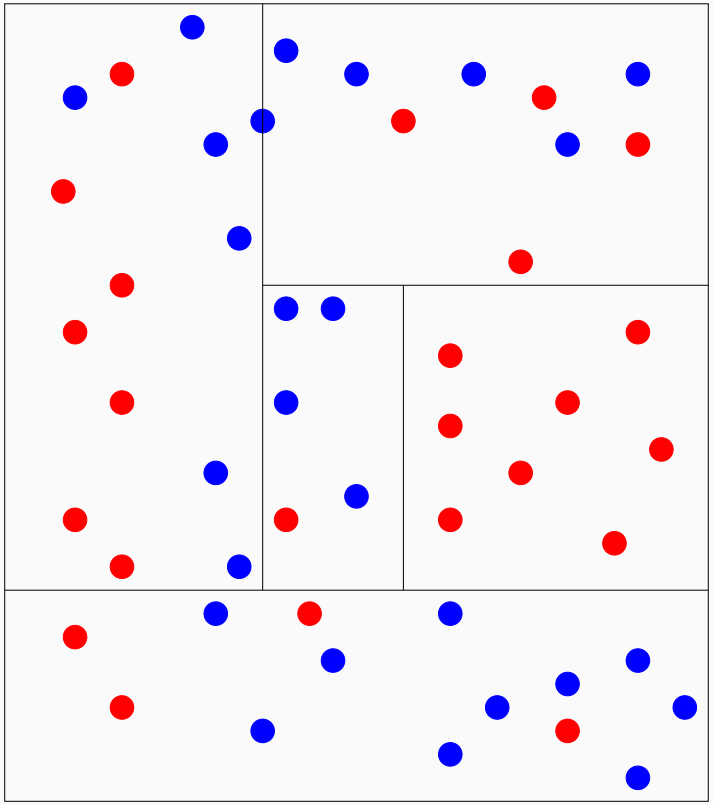
Assume that $h \to 0$ and $nh^d \to \infty$. Then the cubic histogram classifier is universally consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$ for all distributions of (X, Y).

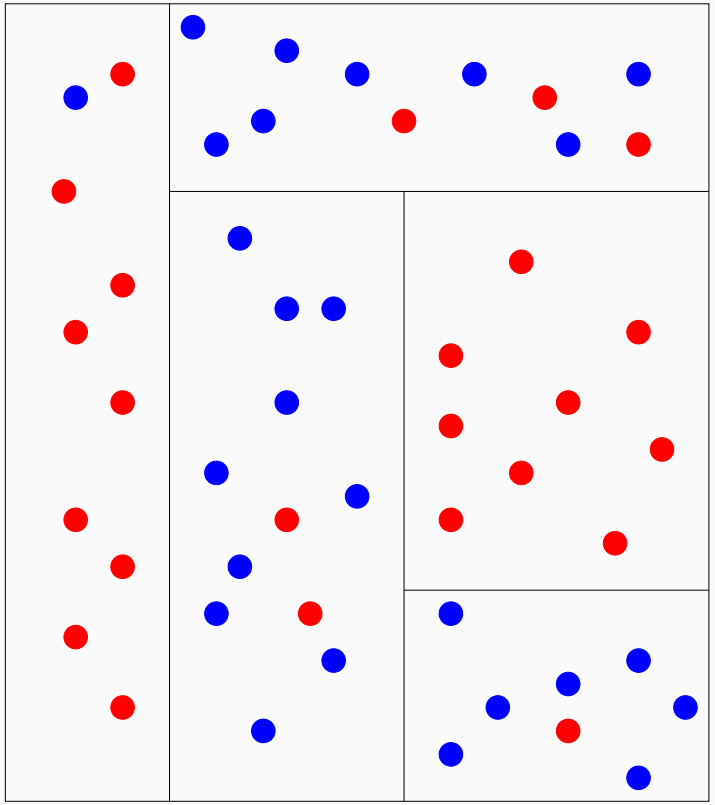


Example 2: tree classifiers

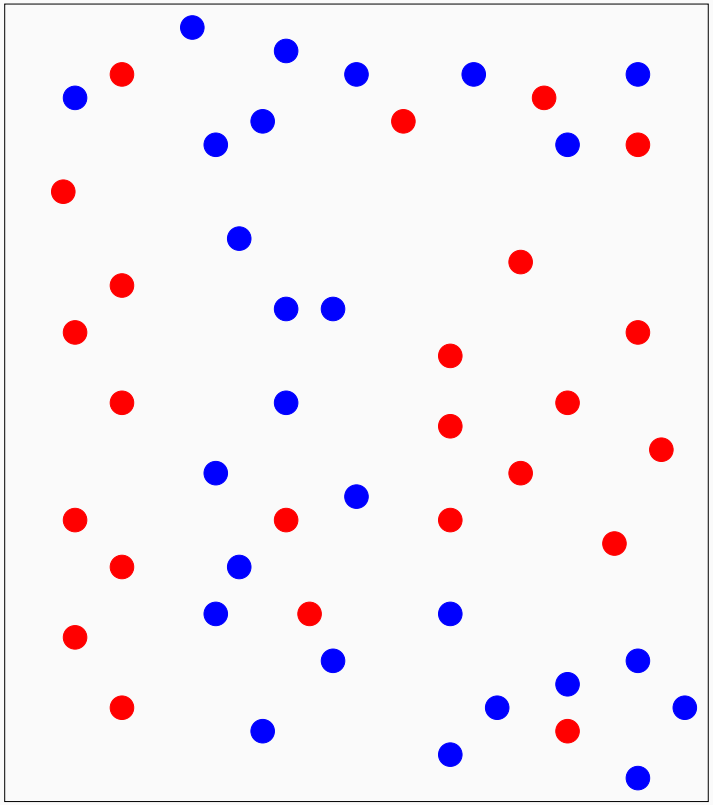


Example 2: tree classifiers

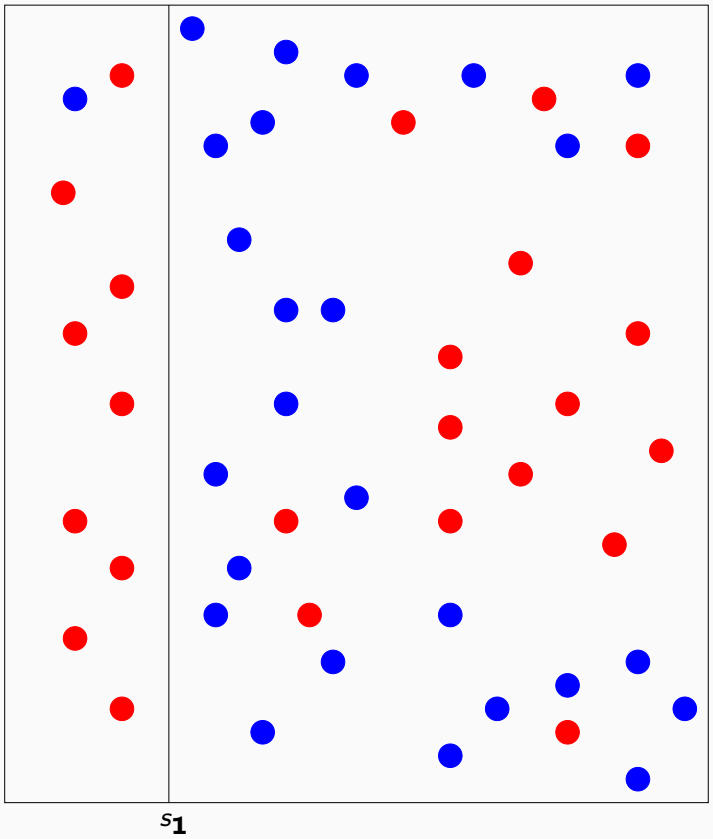




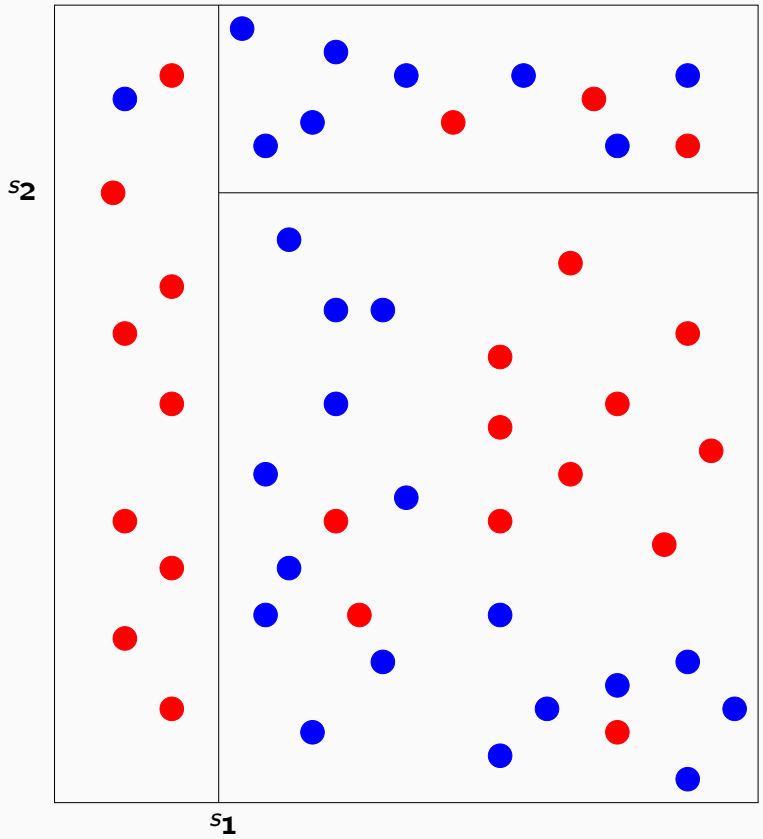
Example 2: tree classifiers



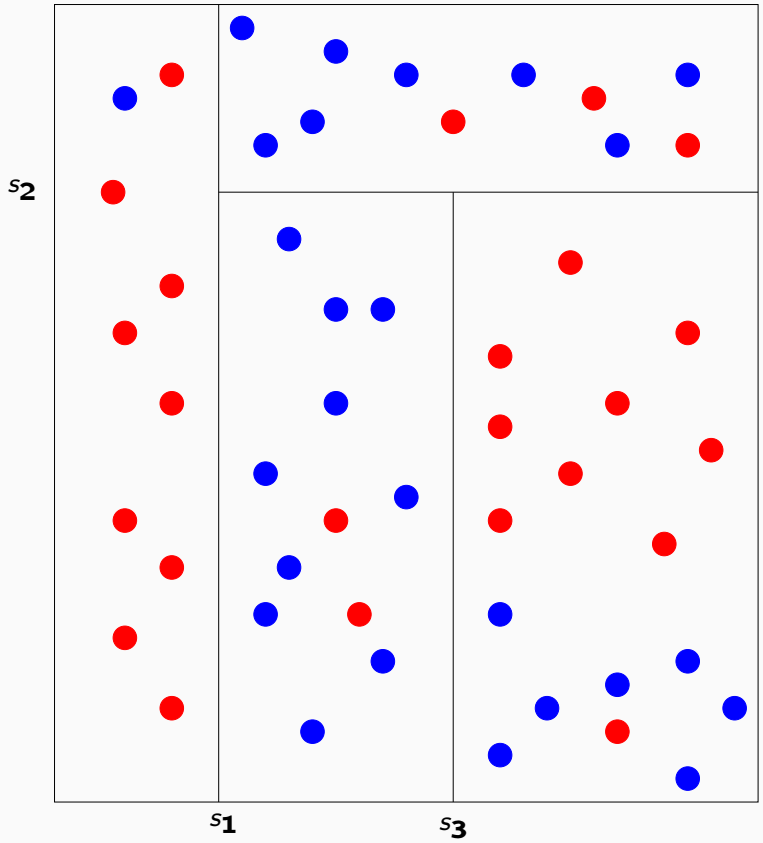
Example 2: tree classifiers



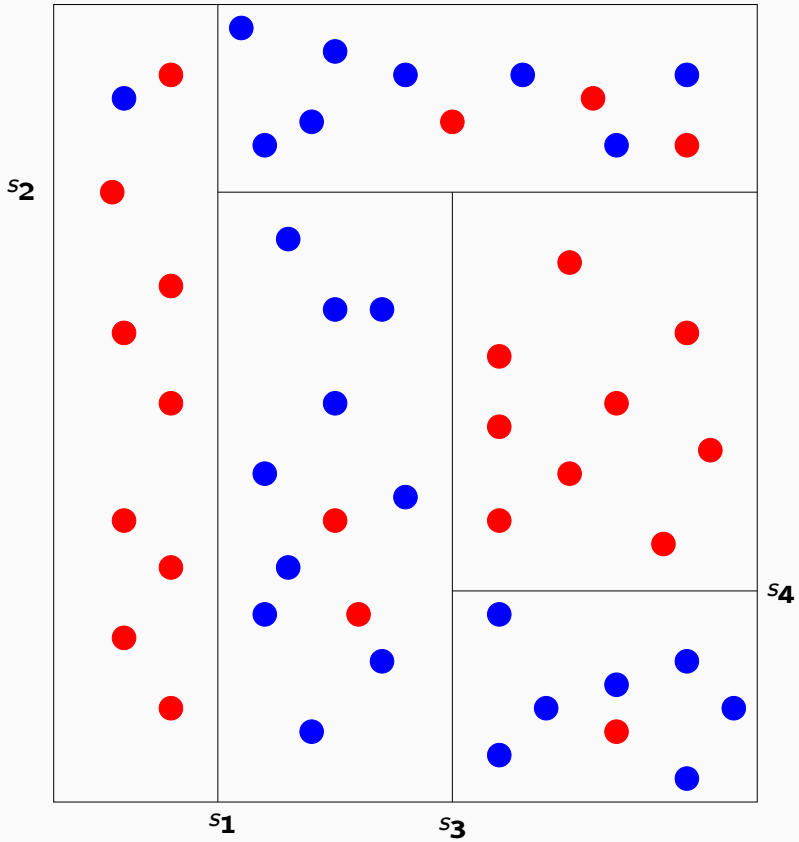
Example 2: tree classifiers



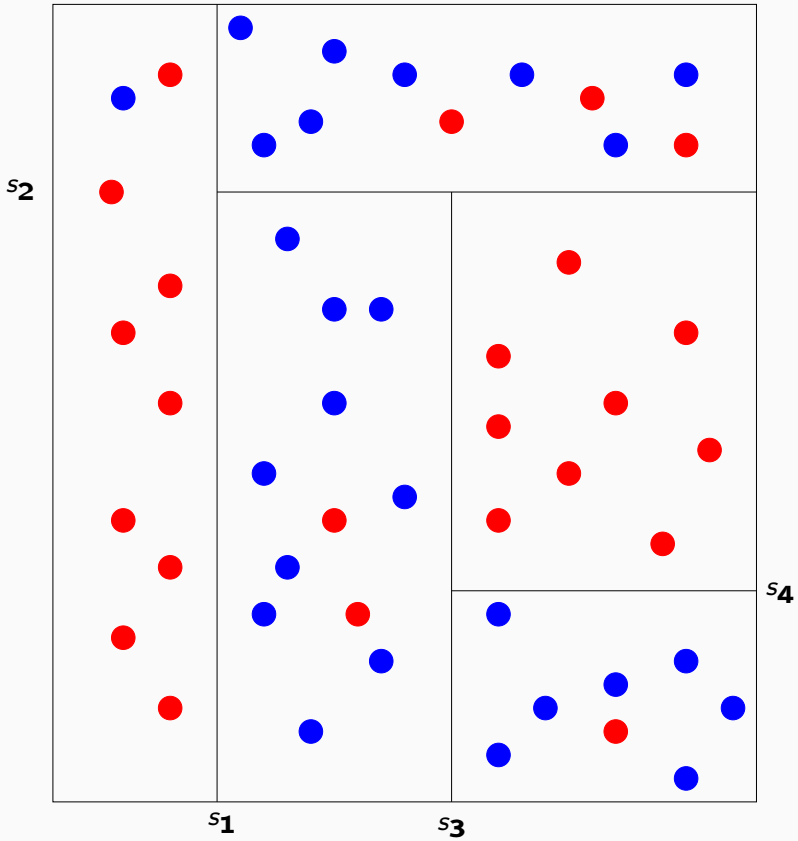
Example 2: tree classifiers

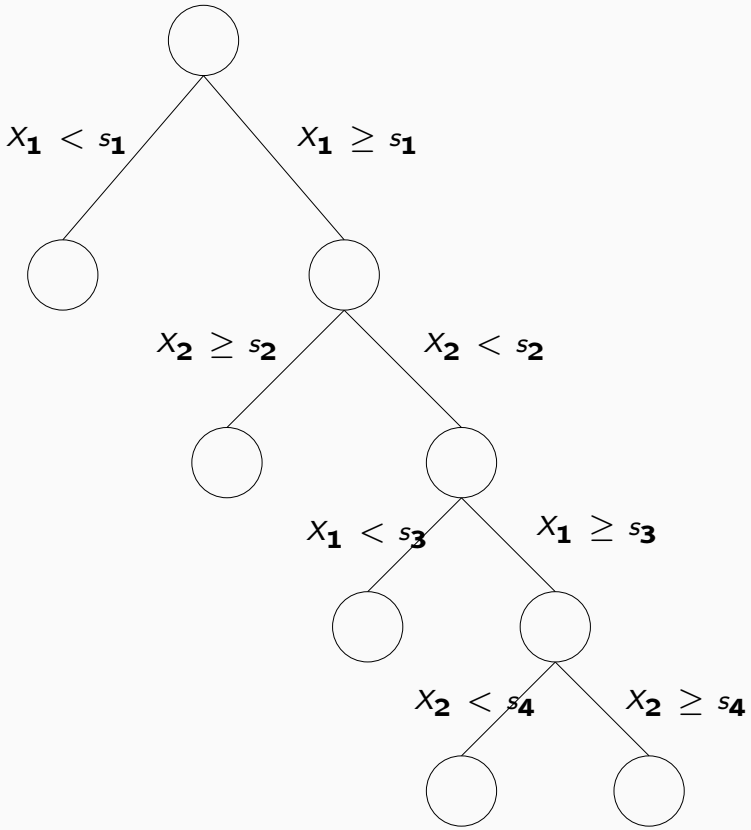


Example 2: tree classifiers

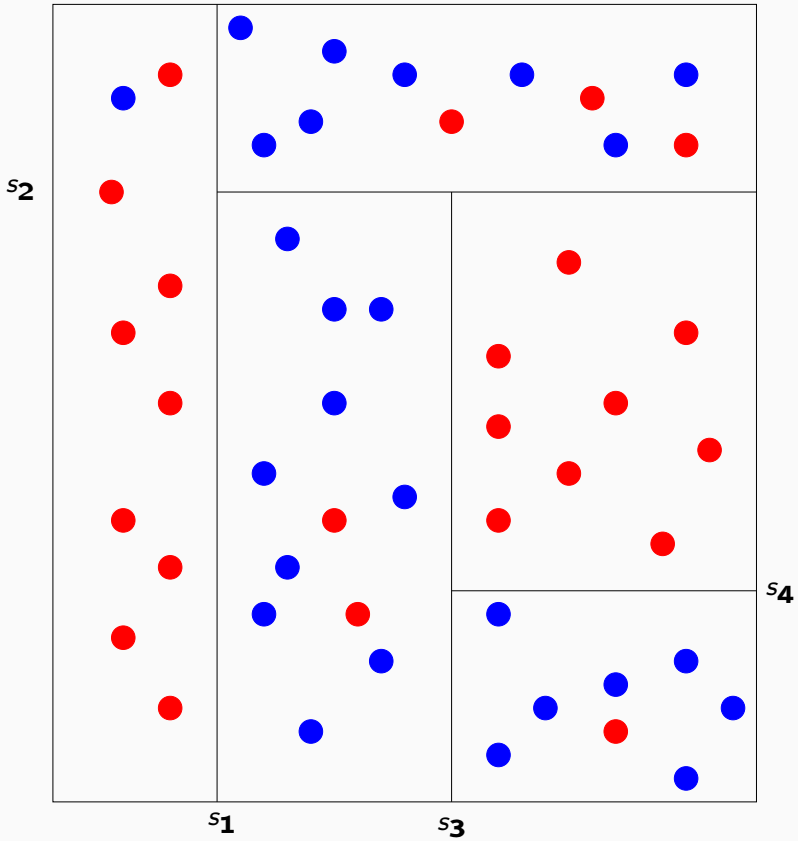


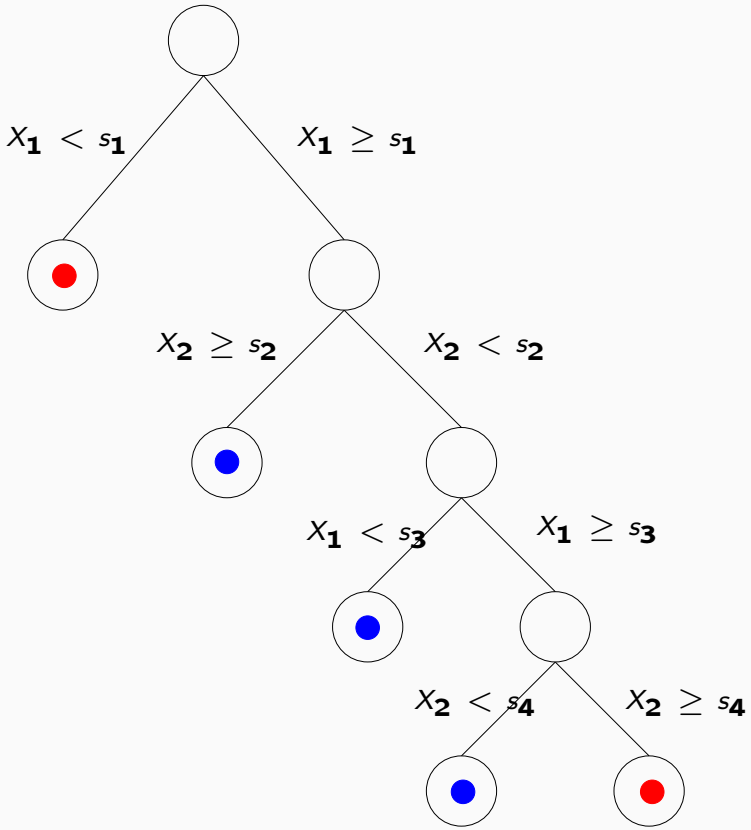
Example 2: tree classifiers





Example 2: tree classifiers





Binary trees

- Definition: recursive binary partitioning of $\mathbb{R}^d$, represented by a tree.
- A node has exactly either zero or two children.
- A node with zero children is called a leaf.
- If $u \leftrightarrow A$ and $u_L, u_R \leftrightarrow A_L, A_R$, then $A = A_L \cup A_R$ and $A_L \cap A_R = \emptyset$.
- The root $\leftrightarrow \mathbb{R}^d$ and the leaves $\leftrightarrow$ a partition of $\mathbb{R}^d$.
- We pass from A to A_L and A_R by answering a question on x:

  "Is $x^{(j)} \geq \alpha$?", for some coordinate j and some α.

- $\mathbb{R}^d$ is partitioned into hyperrectangles.
- Principle: x is passed into the root and then iteratively transmitted to the child nodes. This is repeated until a leaf is reached.

Tree classifiers

- Classifier: for $x \in A$,

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbb{1}_{[X_i \in A]} \mathbb{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbb{1}_{[X_i \in A]} \mathbb{1}_{[Y_i=0]} \\ 0 & \text{otherwise.} \end{cases}$$

- Two questions:
  - ▷ Do we cut?
  - ▷ In the affirmative, where do we cut?
- Many tree species (median, centered, CART, $C_{4.5}$, etc.).
- A data-modeling technique that proved itself to be effective.
- Trees are fast, simple, and able to explain complex data.
- Many modern predictive algorithms rely on tree principles.
- Impact in geometry, statistics, machine learning, and AI.

Median tree classifier

- Principle:
  - ▷ At each node: find the median according to one coordinate;
  - ▷ n points $\rightarrow$ two children with sizes $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$;
  - ▷ The median itself stays behind and is not sent down to the subtrees;
  - ▷ Repeat this for k levels of nodes, in a rotational manner.
- 2^k leaf regions, each having at least $n/2^k - 2$ and at most $n/2^k$ points.

Theorem

Assume that X has a density. If $k \rightarrow \infty$ and $\frac{n}{k2^k} \rightarrow \infty$, then the median tree classifier is consistent, i.e., $\mathbb{E}L(g_n) \rightarrow L^\star$. (Note: the conditions on k are fulfilled if $k \leq \log_2 n - 2\log_2 \log_2 n$, $k \rightarrow \infty$.)

A median tree is invariant under monotone transformations of the axes.

Label-dependent cuts

- Strategy: at each step, choose (j, s) with

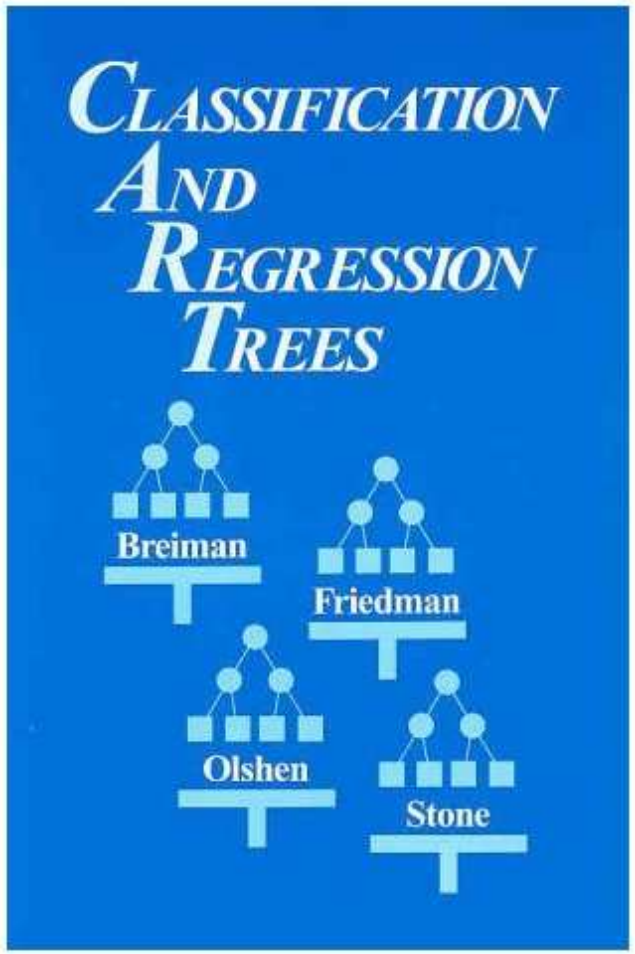
$$R_1(j,s) = \{x : x^{(j)} < s\} \quad \text{and} \quad R_2(j,s) = \{x : x^{(j)} \geq s\}.$$

- The pair (j, s) is selected by minimizing the impurity of the children.
- Impurity: low when the labels are homogeneous, high otherwise.
- Example: choose (j, s) that minimizes

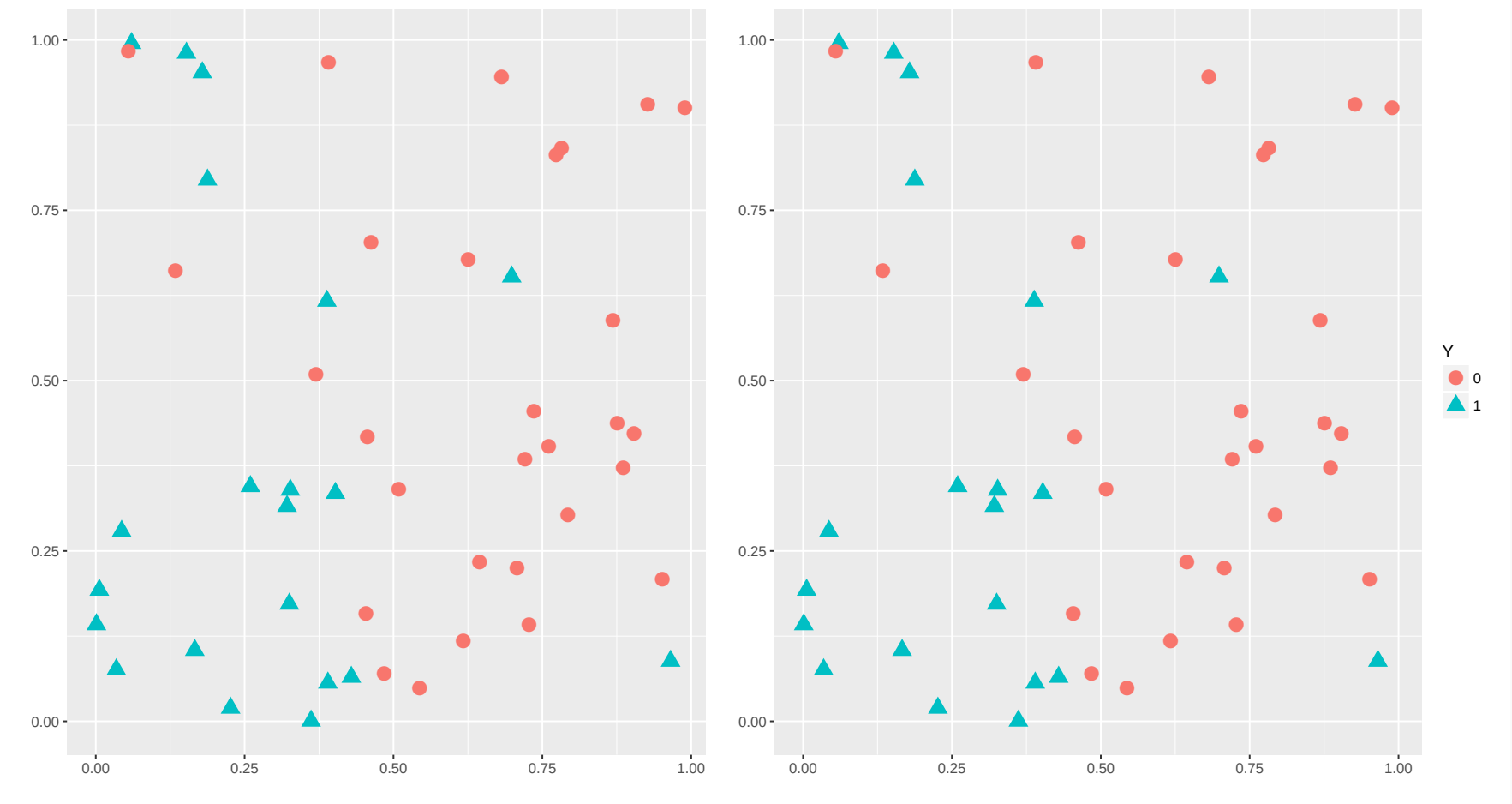
$$\sum_{X_i \in R_{\mathbf{1}}(j,s)} \mathbb{1}_{[Y_i \neq \hat{Y}_{\mathbf{1}}]} + \sum_{X_i \in R_{\mathbf{2}}(j,s)} \mathbb{1}_{[Y_i \neq \hat{Y}_{\mathbf{2}}]},$$

  where $\hat{Y}_k$ is the majority label in the node $R_k(j, s)$.
- There are other criteria: Gini impurity, entropy, etc.
- Example: CART algorithm.

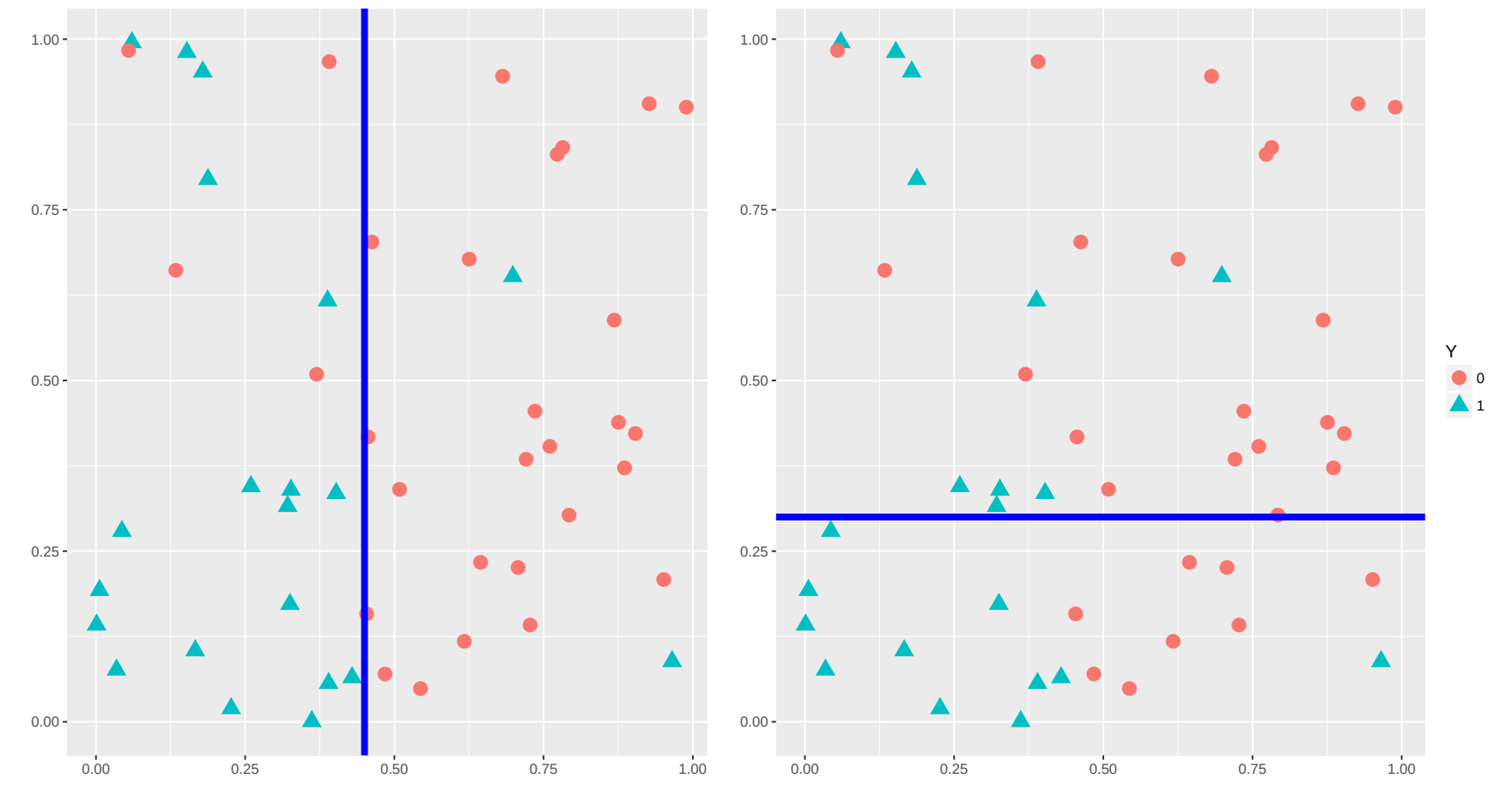
Leo Breiman


CLASSIFICATION
AND
REGRESSION
TREES
Breiman
Friedman
Olshen
Stone

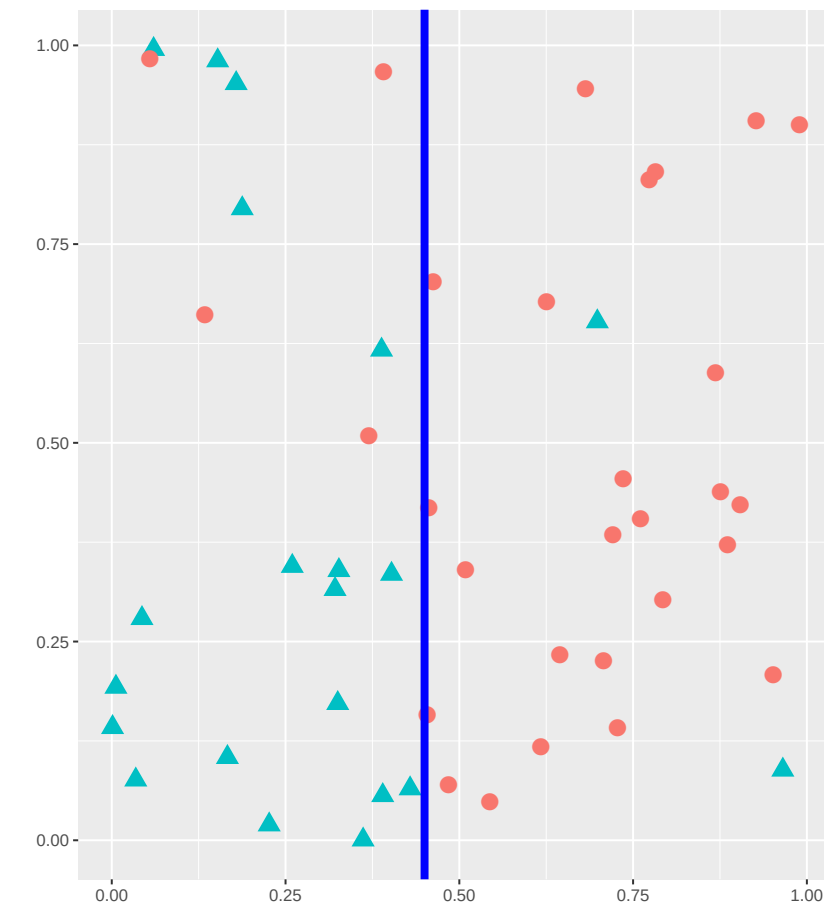
Example



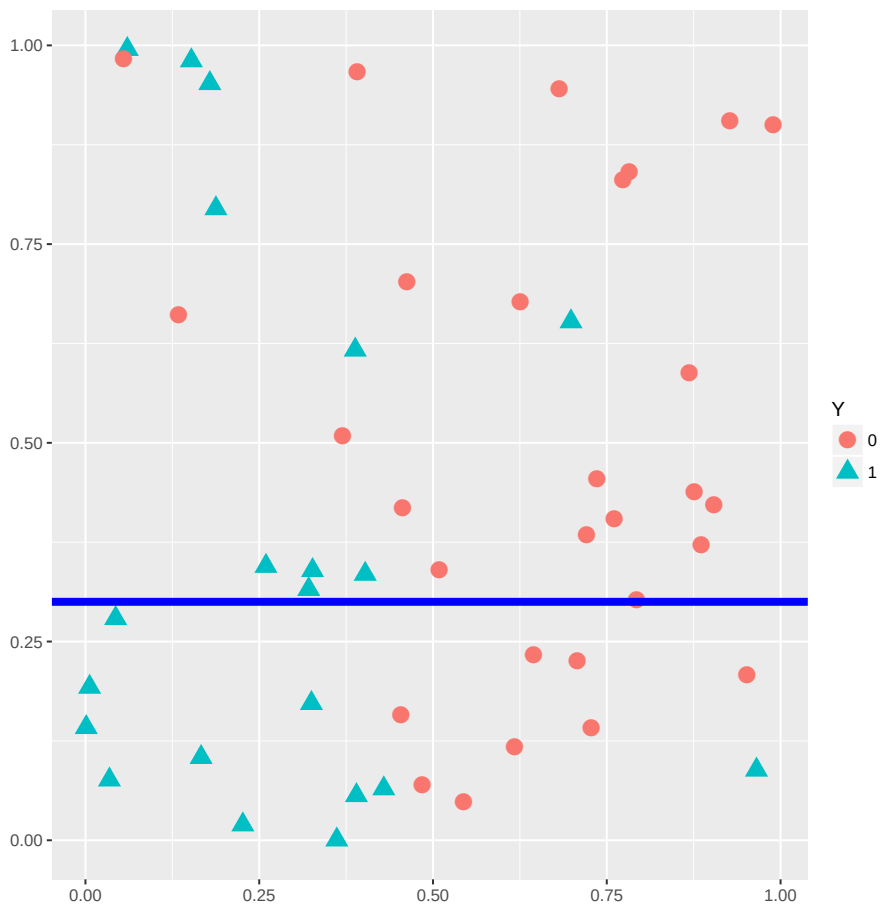
Example



Example

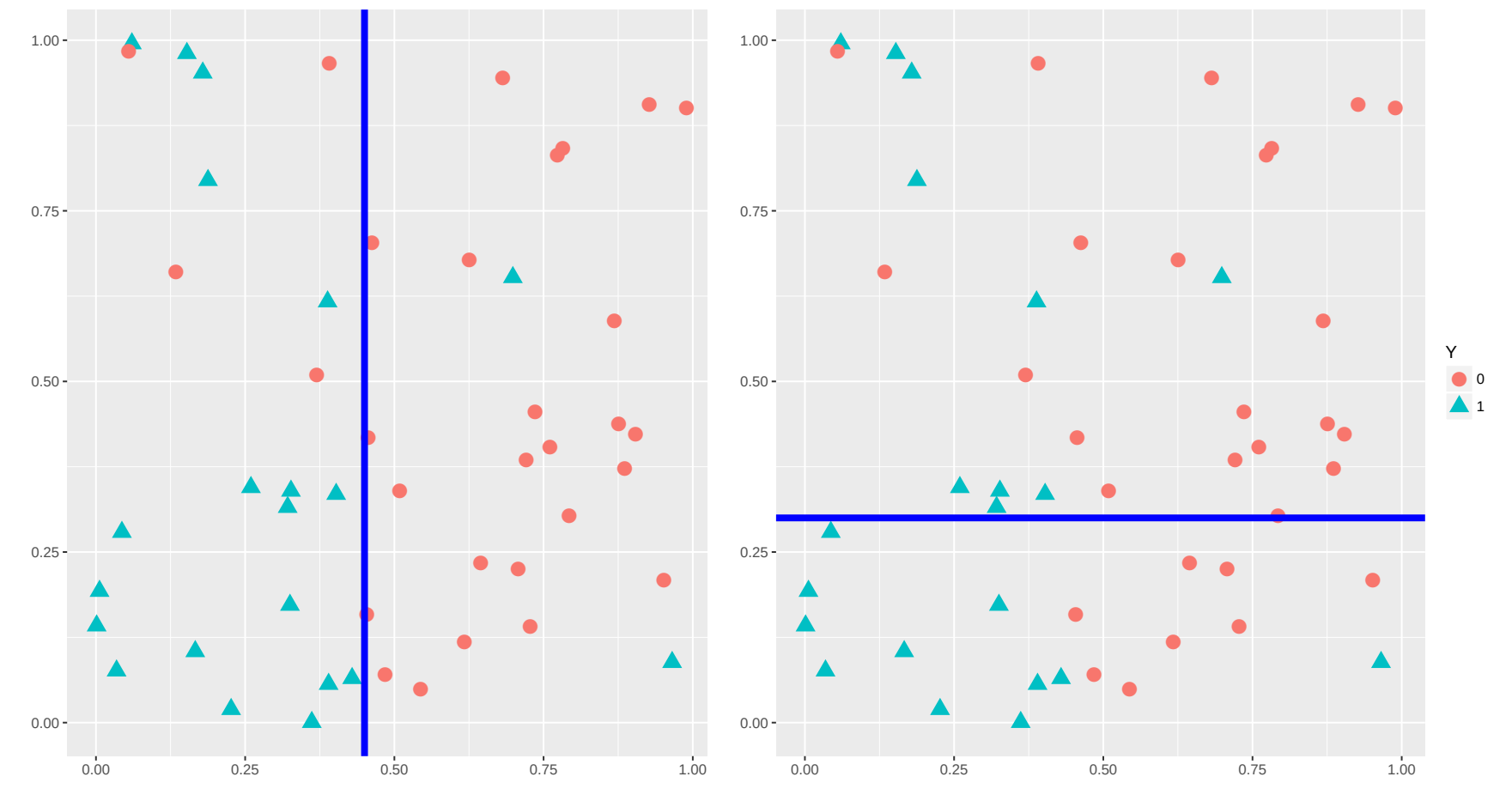


4+2 errors



8+10 errors

Example

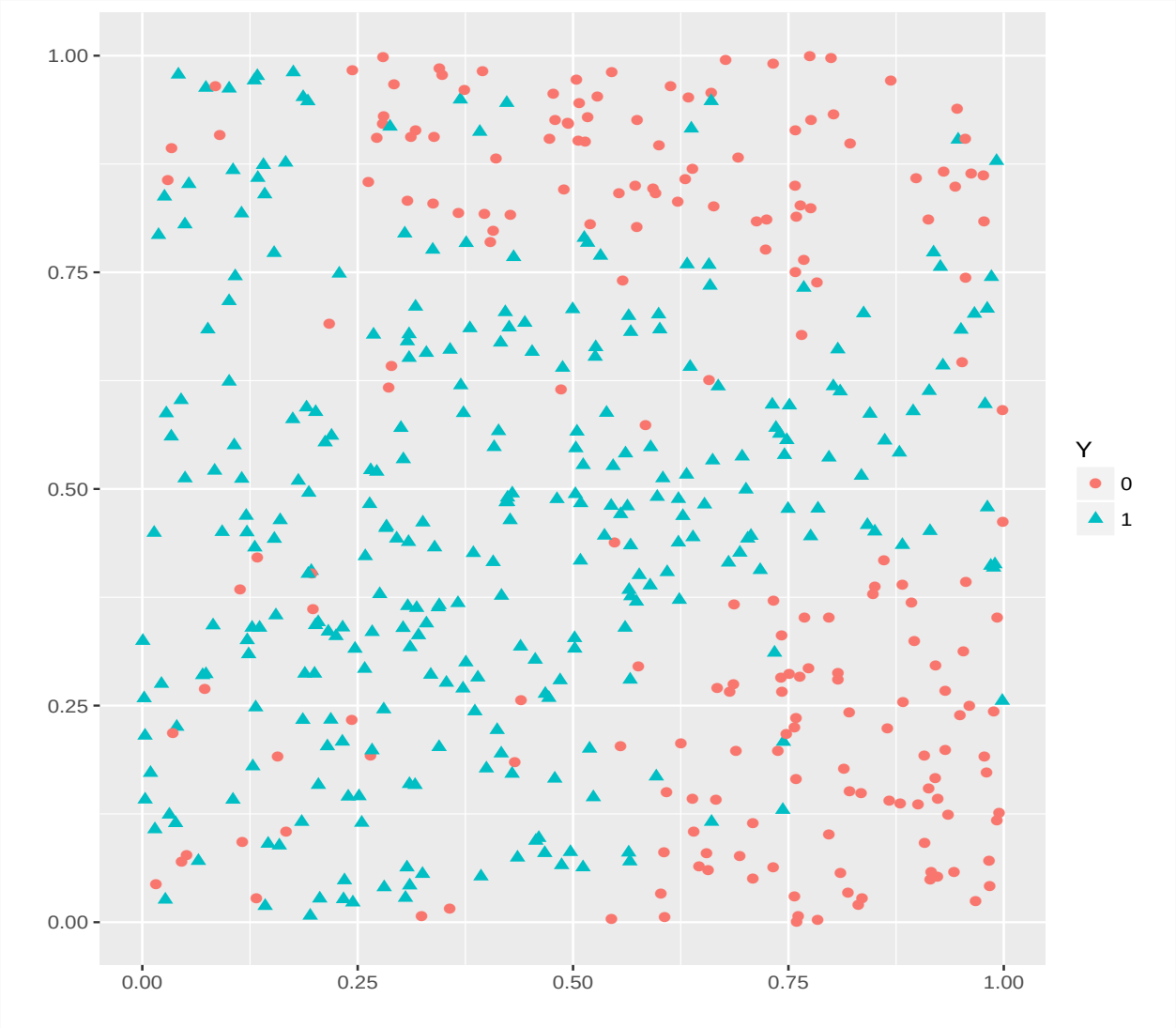


4+2 errors

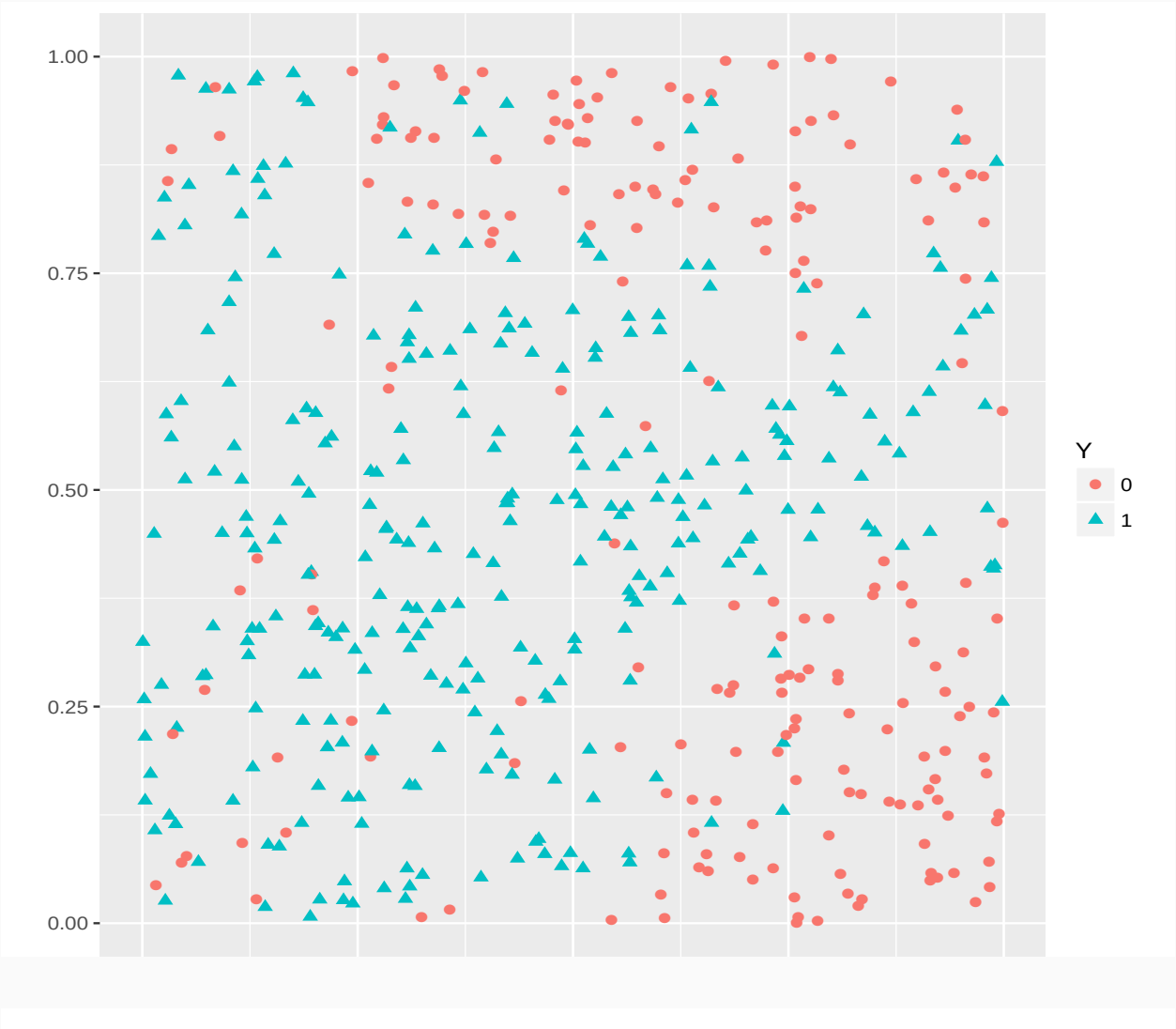
8+10 errors

Best cut: left.

When do we stop?

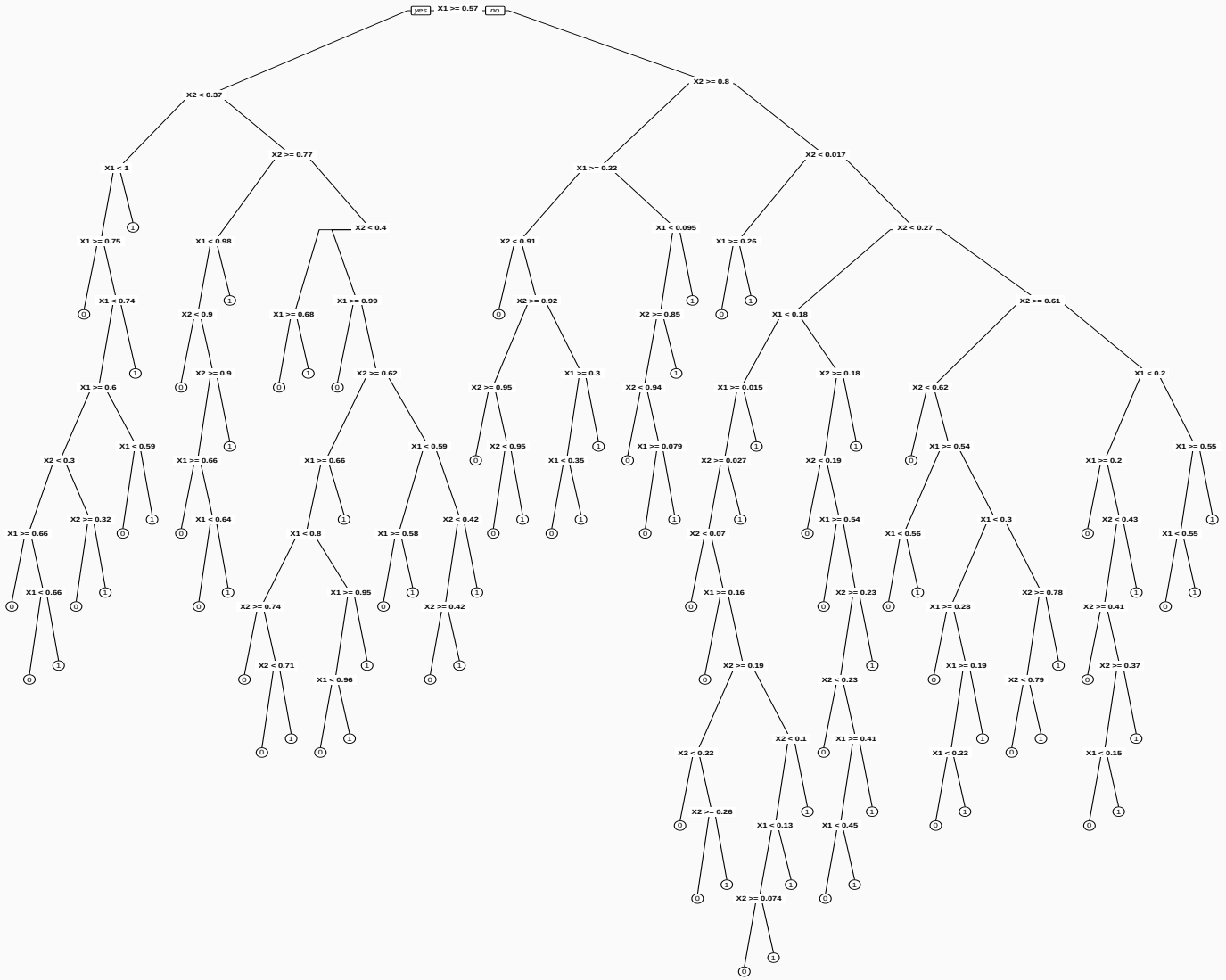


When do we stop?

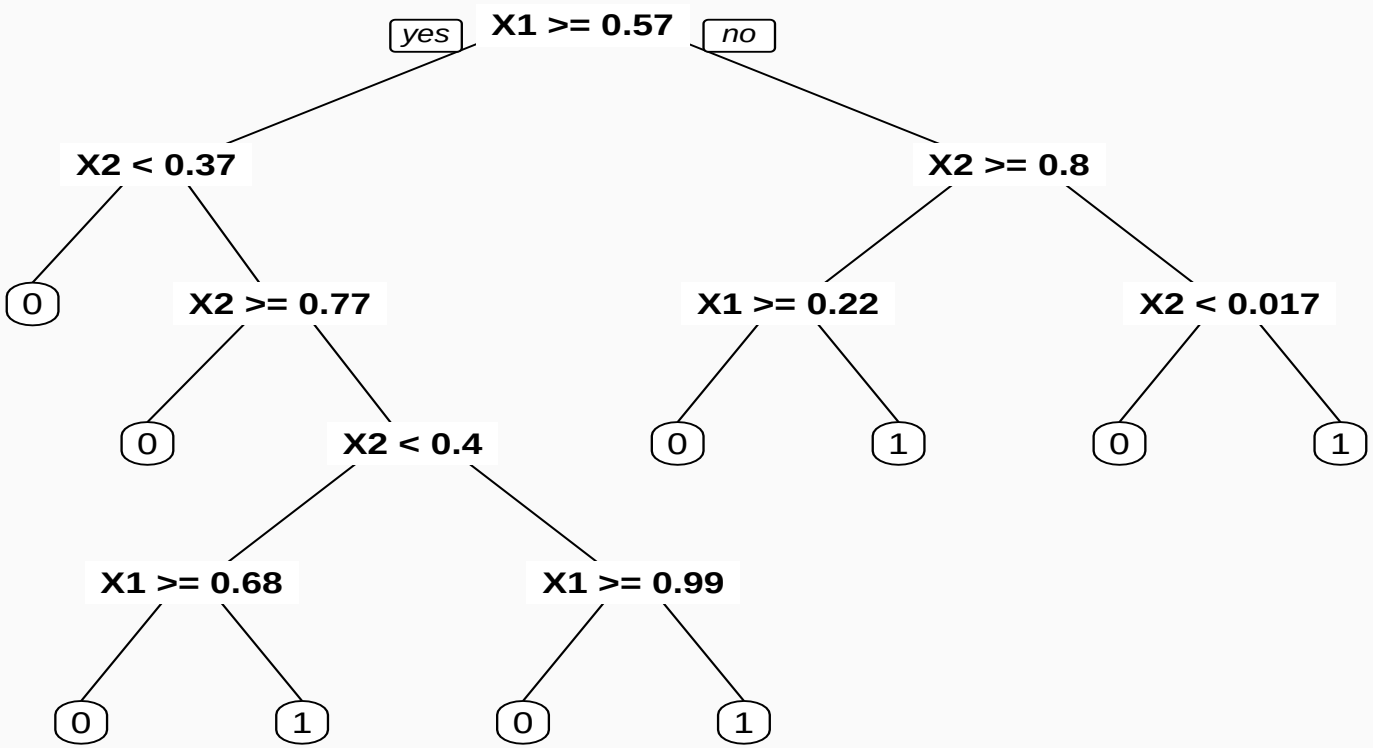


Guess: about 5 cells.

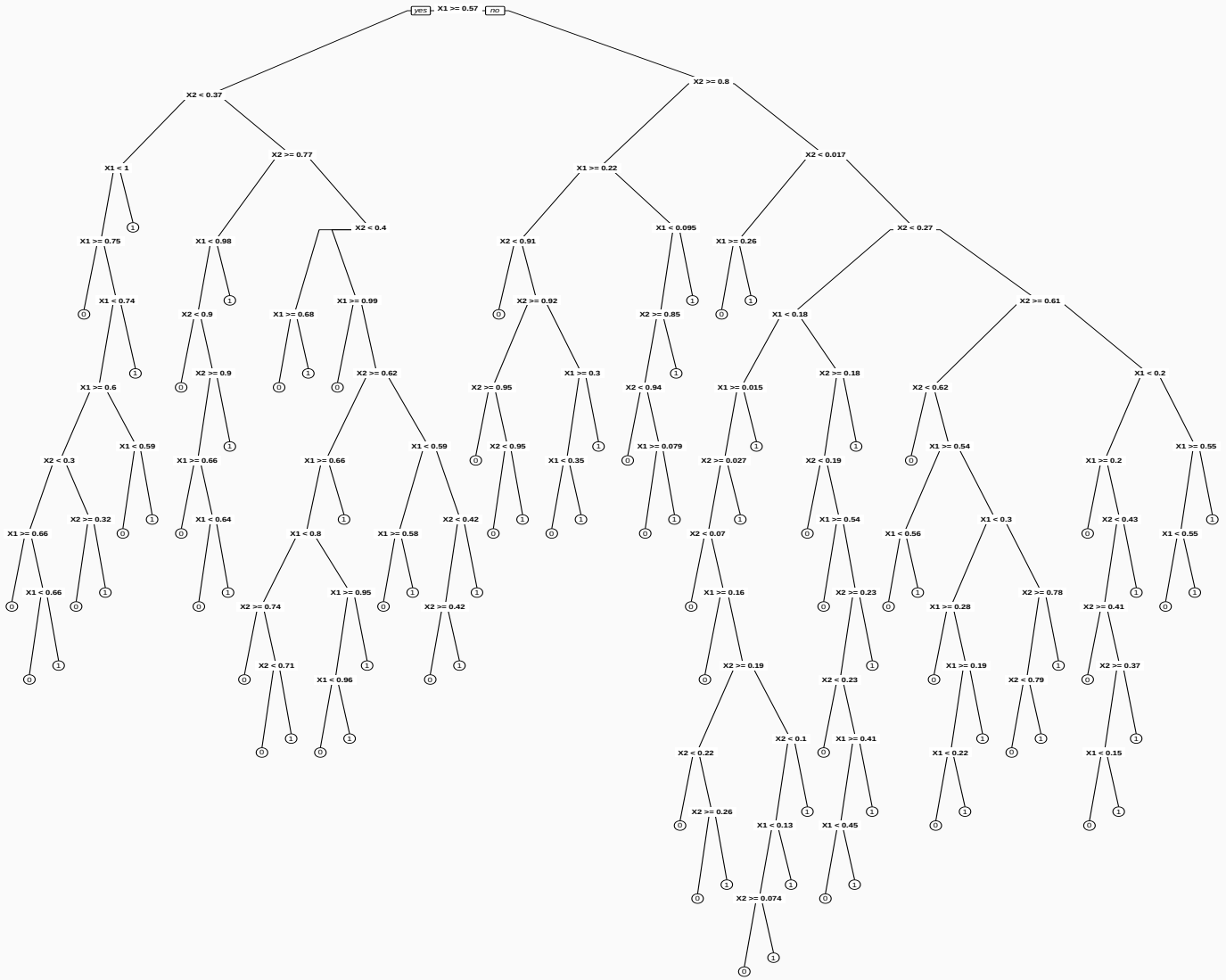
Maximal tree?



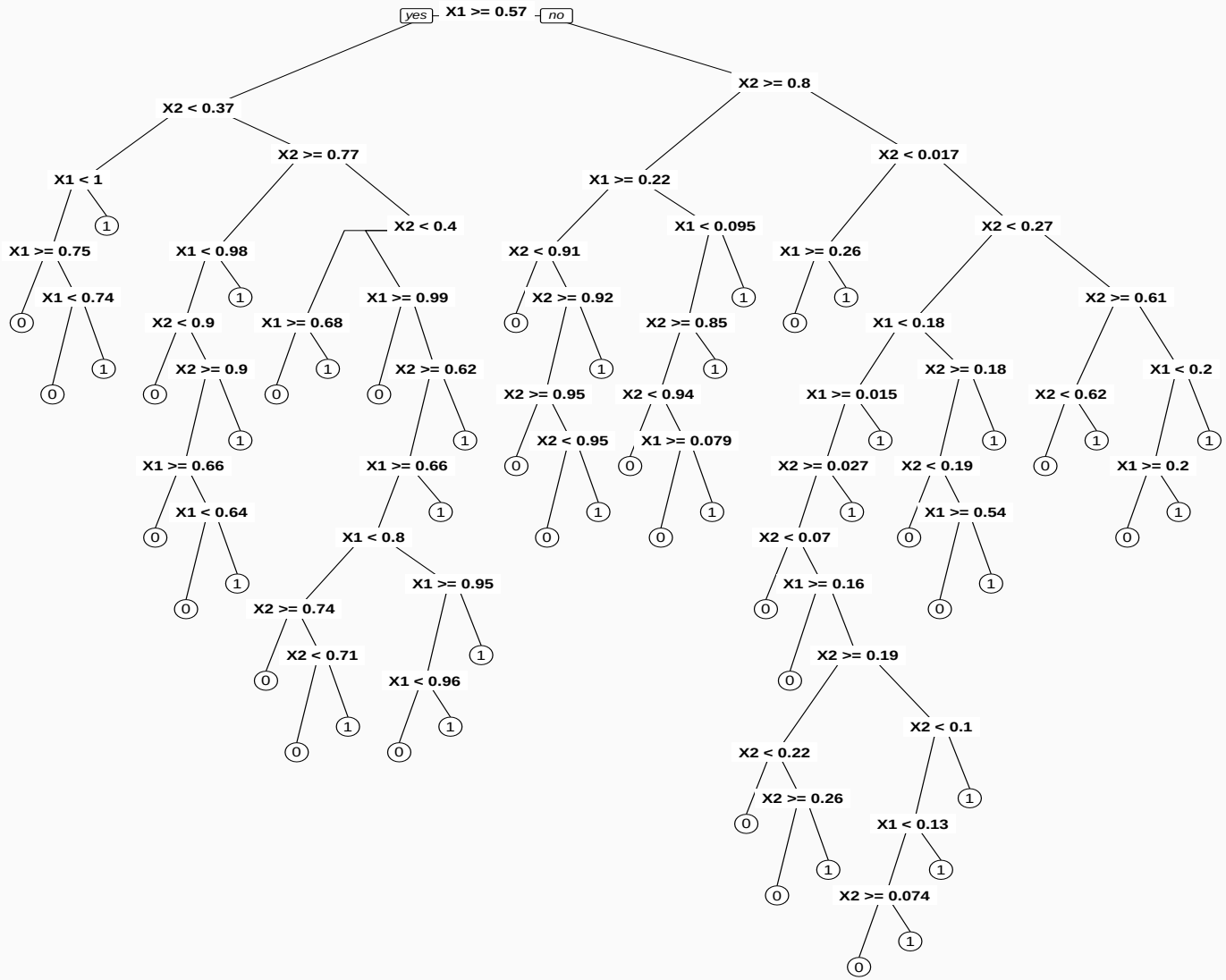
Smaller tree?



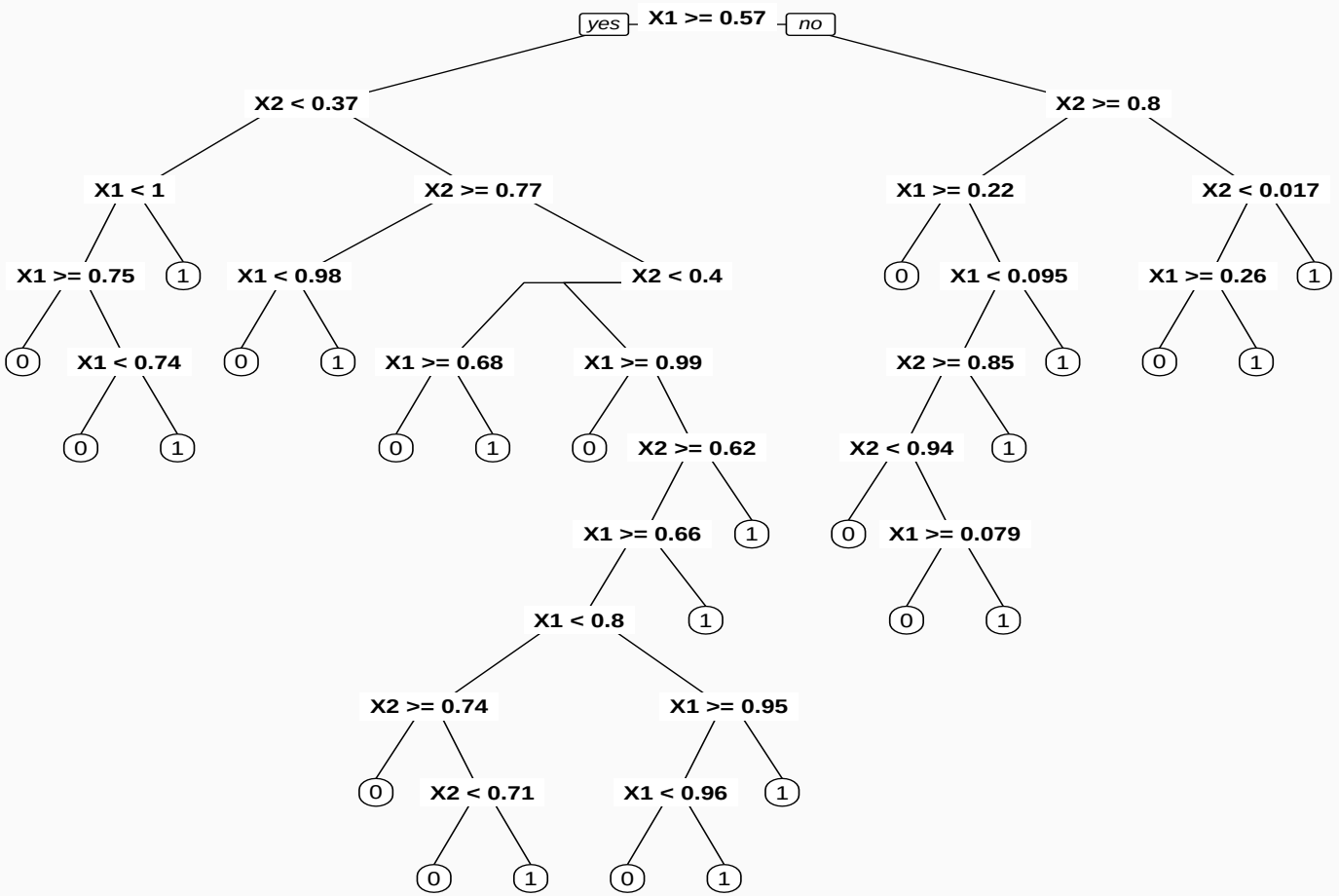
Nested trees



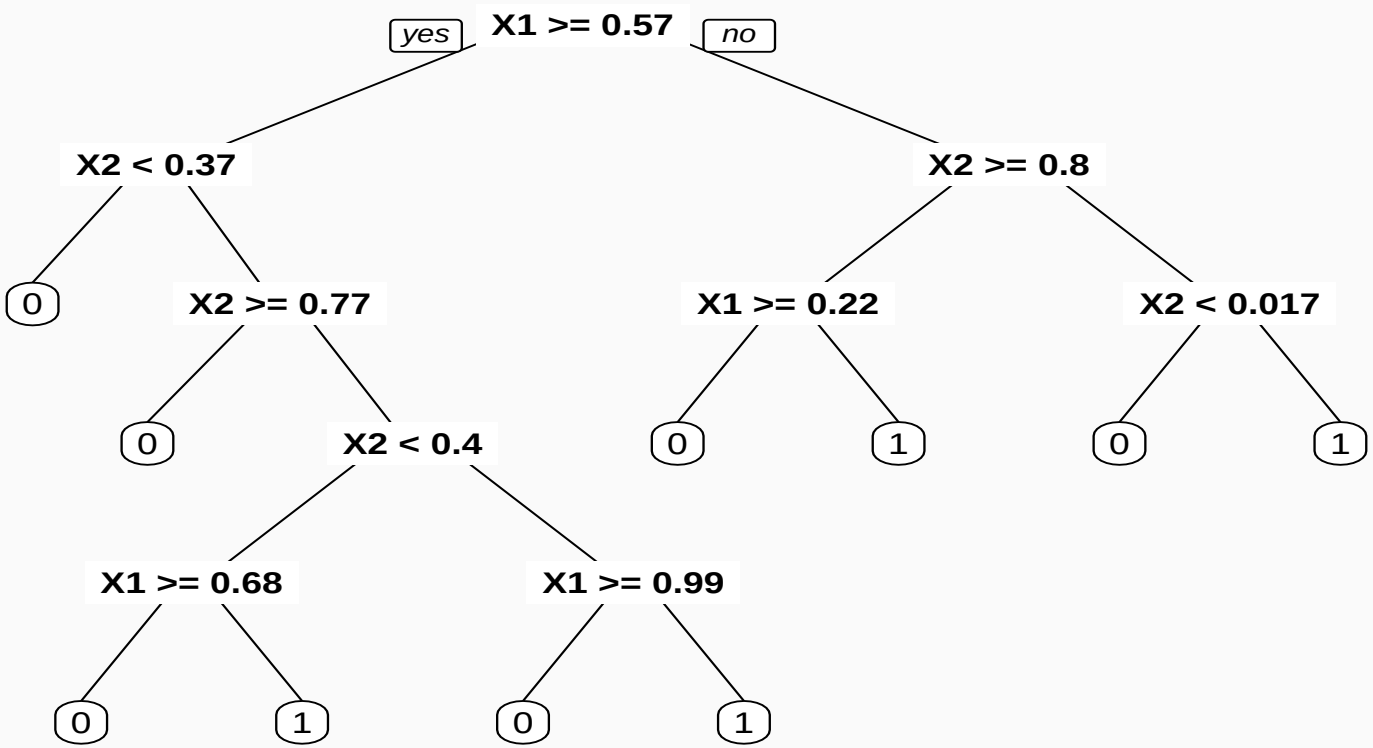
Nested trees



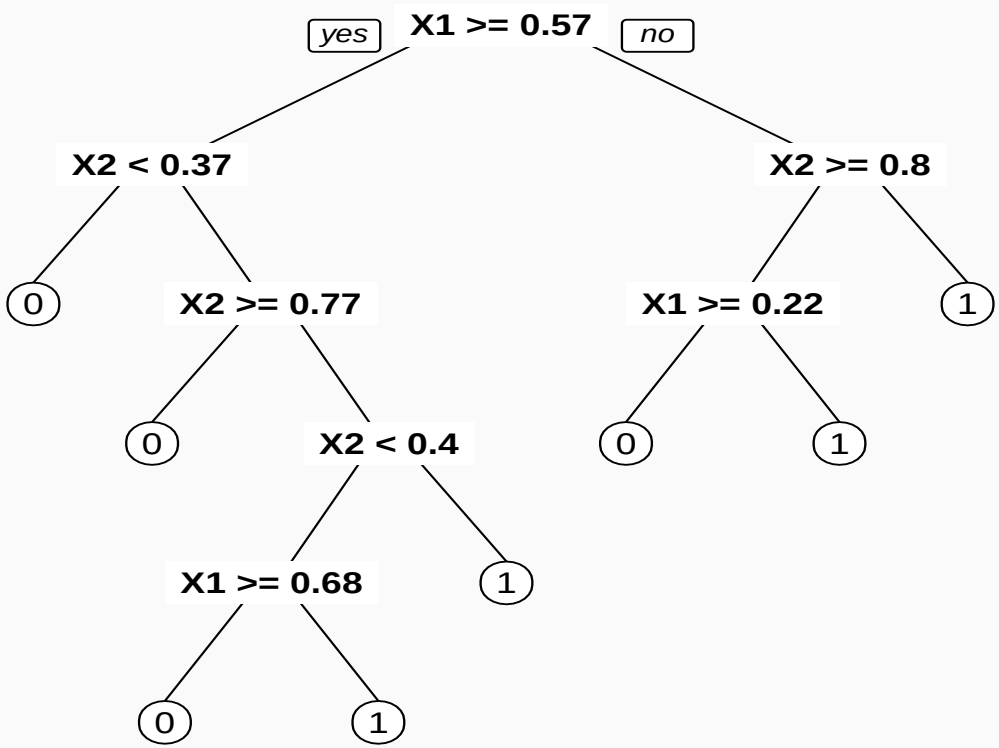
Nested trees



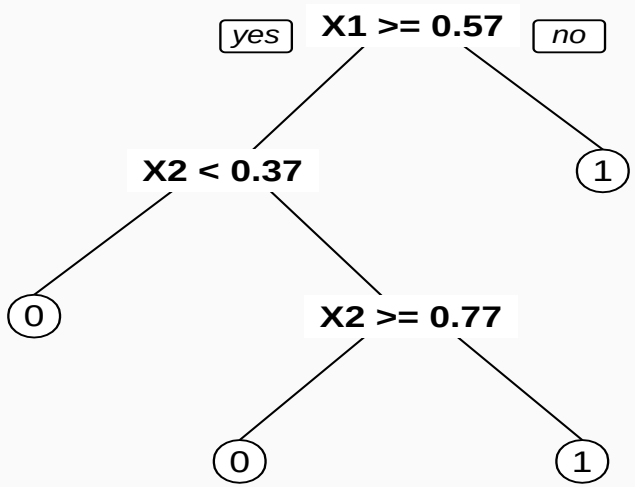
Nested trees



Nested trees

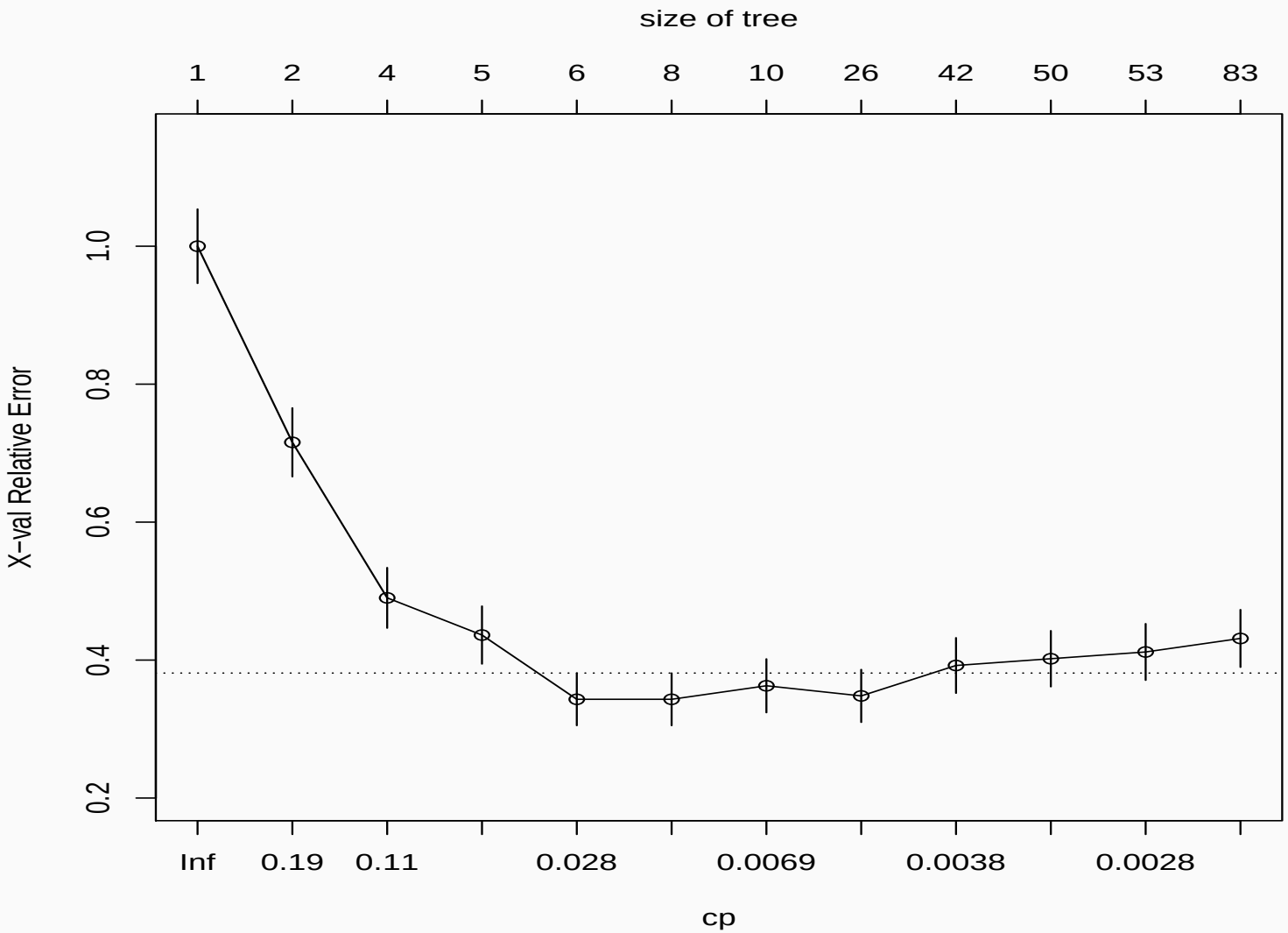


Nested trees

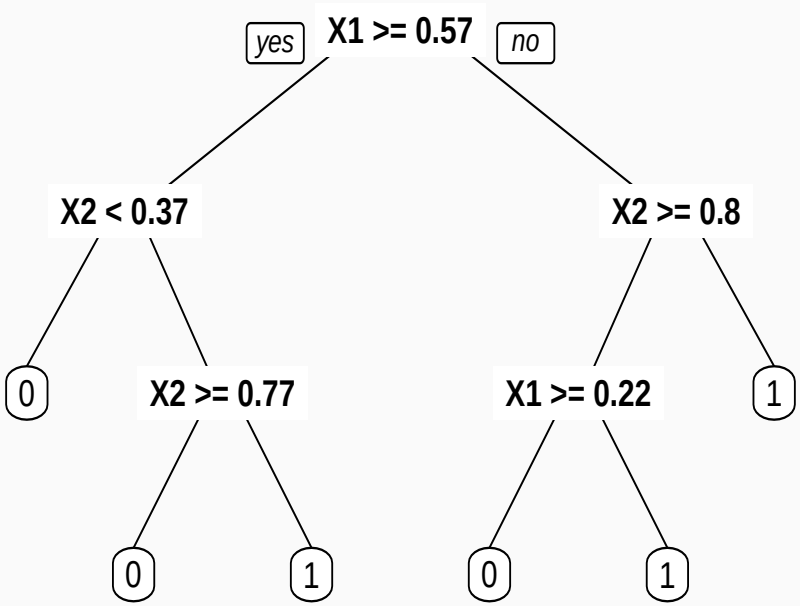


Nested trees

... and pruning

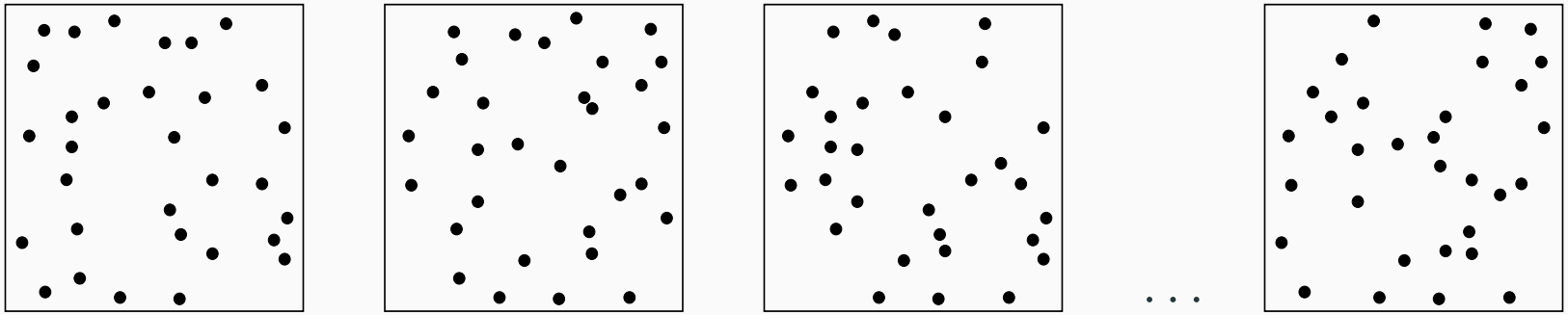

size of tree
1
2
4
5
6
8
10
26
42
50
53
83
X−val Relative Error
0.2
0.4
0.6
0.8
1.0
Inf
0.19
0.11
0.028
0.0069
0.0038
0.0028
cp

Final tree



... Breiman's random forests

1. Data resampling (bootstrap or subsampling).
2. At each cell, randomly preselect a small number of features.
3. Construct the trees by a CART-type procedure.
4. Stopping rule: leave one point in the leaves.
5. Final decision: majority vote.



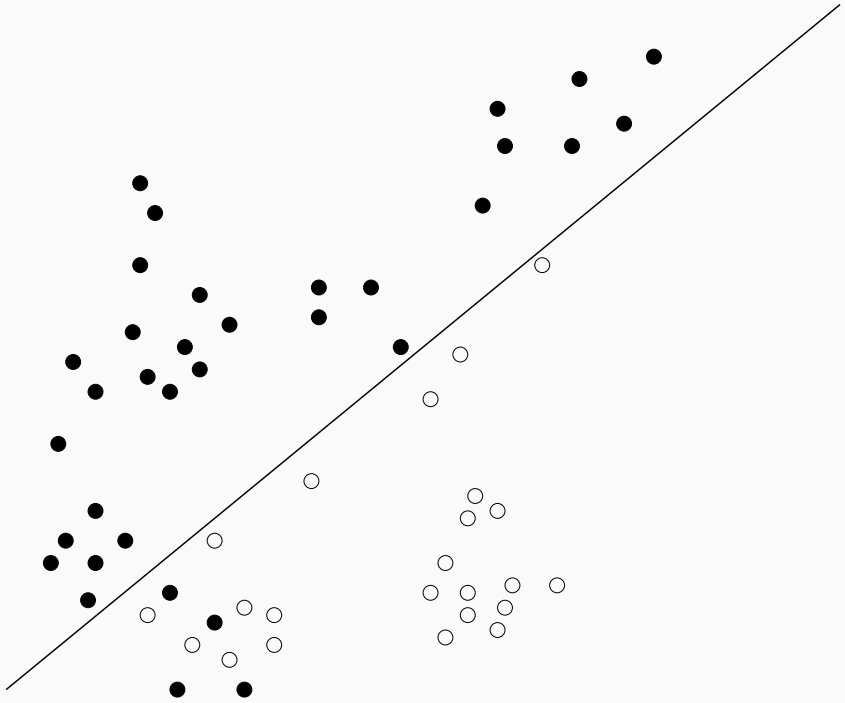
Neural networks

Rosenblatt's perceptron (1958)

- Linear discriminant or perceptron:

$$g(x) = \begin{cases} 1 & \text{if } \psi(x) > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

where $\psi(x) = \sum_{j=1}^{d} c_j x^{(j)} + c_0 = c^\top x + c_0$.



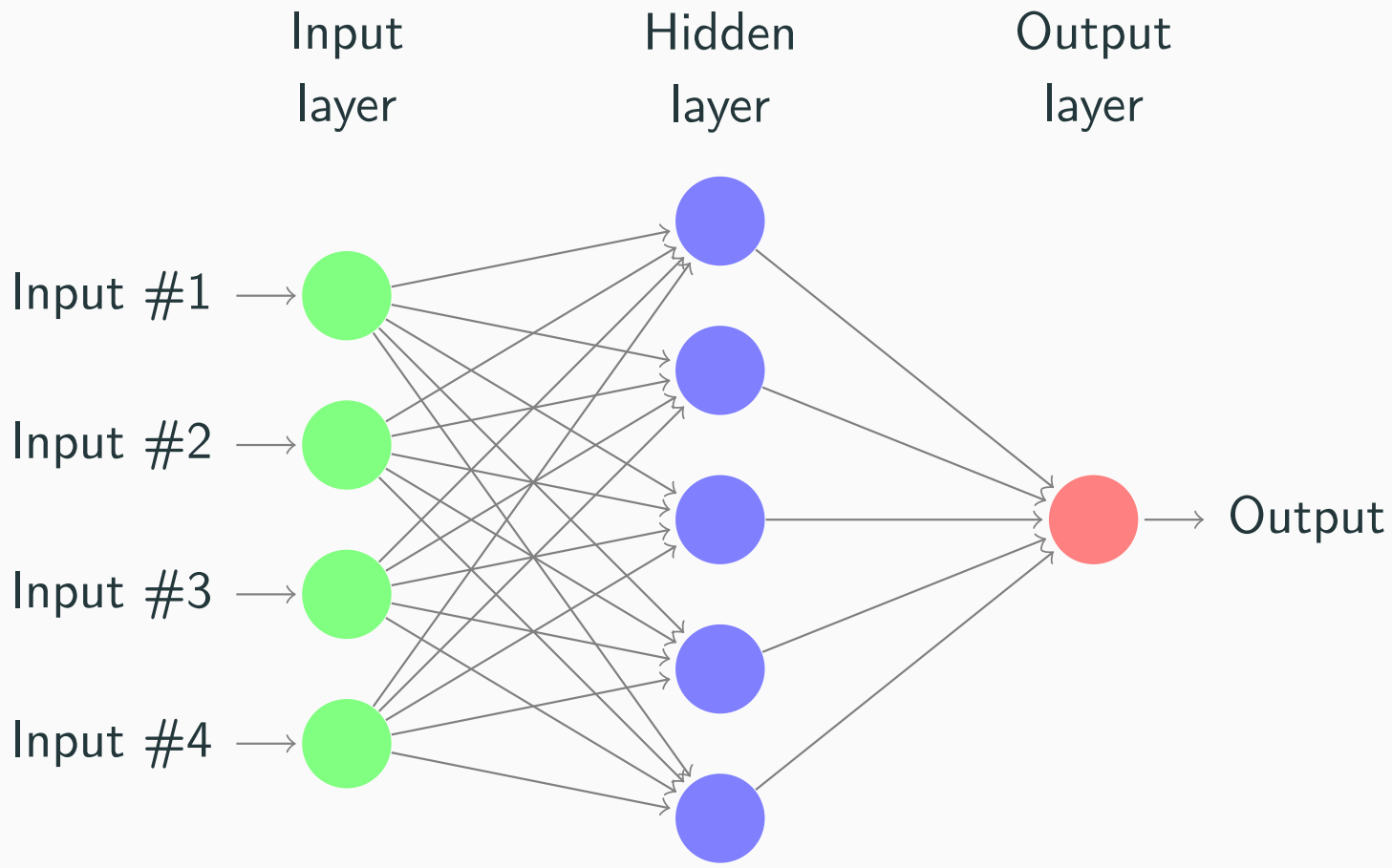
- This is a neural network without hidden layers.

Neural networks

Neural network with one hidden layer:

$$\psi(x) = \sum_{i=1}^{k} c_i \sigma(\psi_i(x)) + c_0,$$

where $\psi_i(x) = \sum_{j=1}^{d} a_{ij} x^{(j)} + a_{i0}$ and σ is the activation function.



Activation function

- Definition: nondecreasing function with

$$\lim_{x \to -\infty} \sigma(x) = -1 \quad \text{and} \quad \lim_{x \to \infty} \sigma(x) = 1.$$

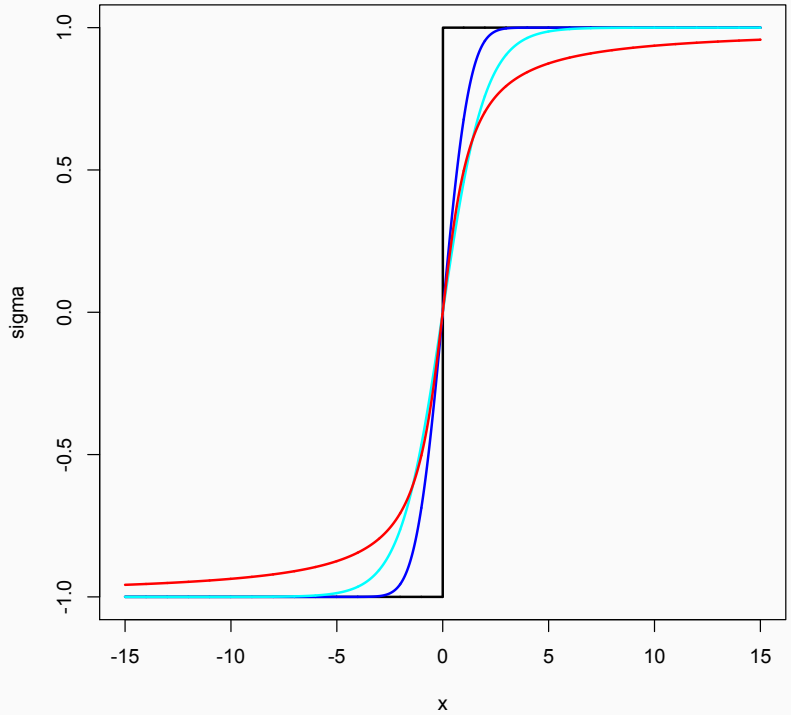
- Examples:

▷ Threshold: $\sigma(x) = 2\mathbb{1}_{[x \geq 0]} - 1$

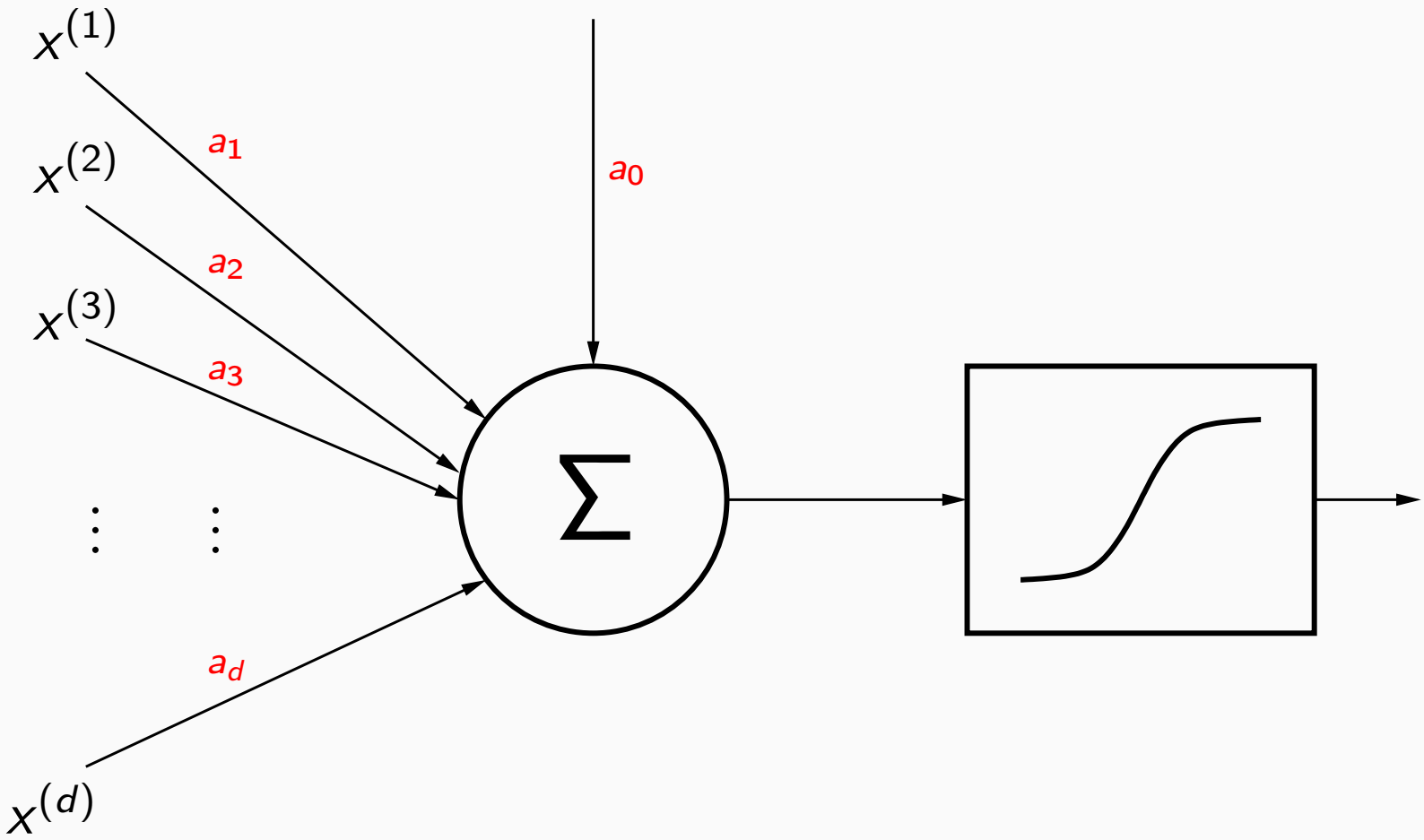
▷ Gaussian: $\sigma(x) = 2\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2} \mathrm{d}t - 1$

▷ Logistic: $\sigma(x) = \frac{1-e^{-x}}{1+e^{-x}}$

▷ Arctan: $\sigma(x) = \frac{2}{\pi} \arctan(x)$



A neuron



More layers, more neurons

- A one-hidden-layer neural network has k hidden neurons.
- Output of the i-th neuron: $u_i = \sigma(\psi_i(x))$. So,

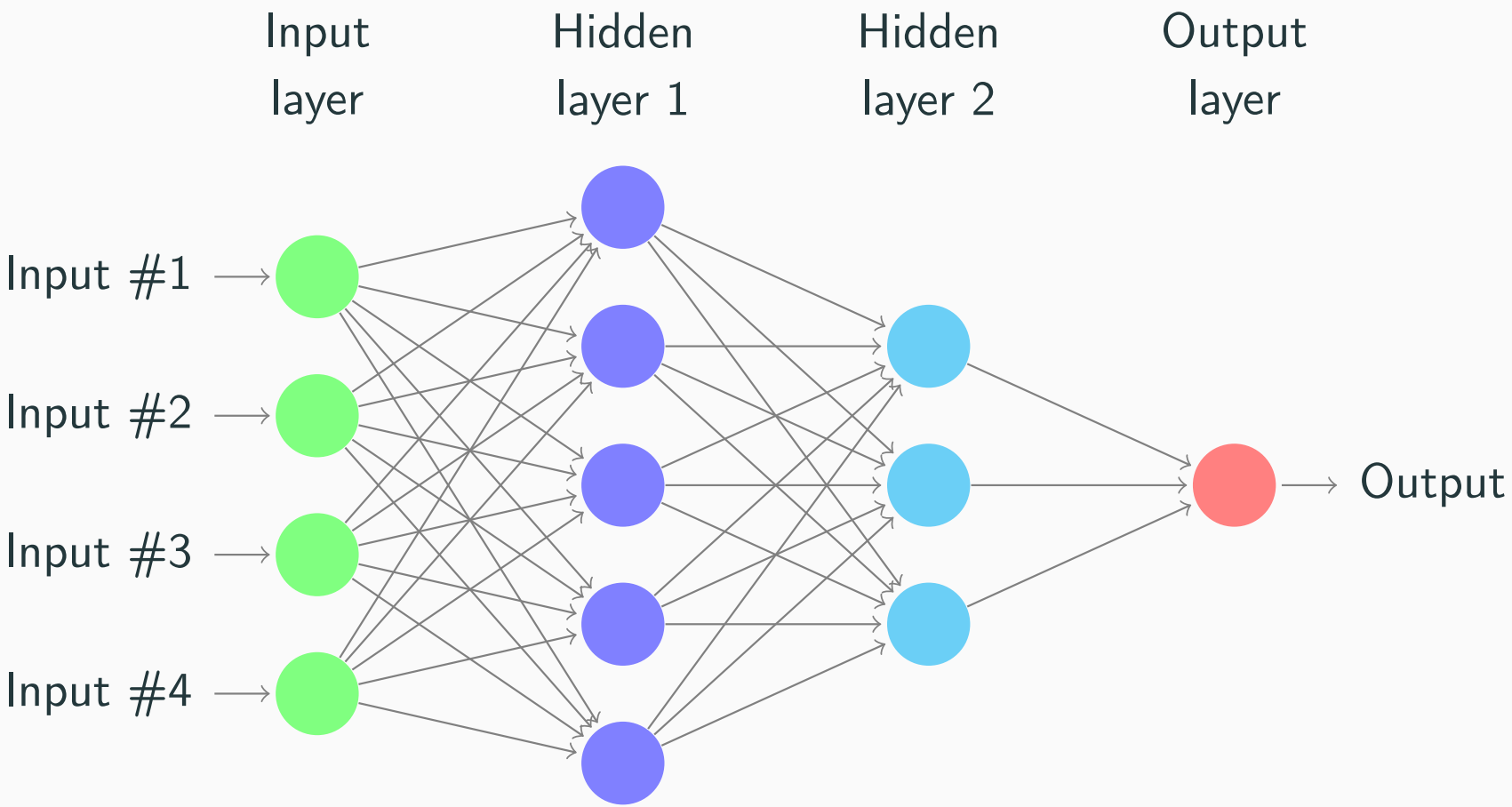
$$\psi(x) = \sum_{i=1}^{k} c_i u_i + c_0.$$

- Two-hidden-layer neural network: $\psi(x) = \sum_{i=1}^{\ell} c_i z_i + c_0$, where

$$z_i = \sigma\Big(\sum_{j=1}^{k} b_{ij} u_j + b_{i0}\Big) \quad \text{and} \quad u_j = \sigma\Big(\sum_{i=1}^{d} a_{ji} x^{(i)} + a_{j0}\Big).$$

- First hidden layer: k neurons. Second hidden layer: ℓ neurons.
- And so on... Deep learning: more than 100 layers!

Two-hidden-layer neural network



Theoretical roots

- A theorem by Kolmogorov (1957) and Lorentz (1976): for every continuous function f on $[0,1]^d$,

$$f(x) = \sum_{i=1}^{2d+1} F_i\Big(\sum_{j=1}^{d} G_{ij}(x^{(j)})\Big),$$

  where the G_{ij} and the F_i are continuous functions depending on f.
- Example: with $d = 2$ and $f(x) = x^{(1)}x^{(2)}$,

$$f(x) = \frac{1}{4}\Big((x^{(1)} + x^{(2)})^2 - (x^{(1)} - x^{(2)})^2\Big).$$

- Next step: approximate the G_{ij} and F_i by functions $\sigma(a^\top x + a_0)$.
- One-hidden-layer: universally consistent classifiers provided $k_n \to \infty$.
- Unrealistic, since the hardware is fixed beforehand.

Approximation by neural networks

Theorem

If $\mathscr{C}_k$ is the class of all neural network classifiers with the threshold activation function and k neurons in two hidden layers, then

$$\lim_{k\to\infty} \inf_{g\in\mathscr{C}_k} L(g) - L^\star = 0$$

for all distributions of (X, Y).

Theorem

If $\mathscr{C}_k$ is the class of all neural network classifiers with one hidden layer of k nodes and an arbitrary activation function, then

$$\lim_{k\to\infty} \inf_{g\in\mathscr{C}_k} L(g) - L^\star = 0$$

for all distributions of (X, Y).

Approximation results

Lemma

Let $(\mathscr{F}_k)_k$ be a sequence of classes of functions $f : \mathbb{R}^d \to \mathbb{R}$, and let $(\mathscr{C}_k)_k$ be the companion sequence of classes of classifiers. Assume that for every $a, b \in \mathbb{R}^d$ and every continuous function h on $[a, b]^d$,

$$\lim_{k\to\infty} \inf_{f\in\mathscr{F}_k} \sup_{x\in[a,b]^d} |h(x) - f(x)| = 0.$$

Then, for any distribution of (X, Y),

$$\lim_{k\to\infty} \inf_{g\in\mathscr{C}_k} L(g) - L^\star = 0.$$

Approximation results, cont.

Theorem

For every continuous function $h : [a, b]^d \to \mathbb{R}$ *and for every* $\varepsilon > 0$, *there exists a neural network with one hidden layer, of the form*

$$\psi(x) = \sum_{i=1}^{k} c_i \sigma(\psi_i(x)) + c_0,$$

such that

$$\sup_{x \in [a,b]^d} |h(x) - \psi(x)| \leq \varepsilon.$$

Consistency results

- Assumptions:
  - $\mathscr{C}_k$ = one-hidden-layer neural network classifiers with k nodes;
  - σ = threshold activation function.
- Shatter coefficient:

$$\mathsf{S}_{\mathscr{A}}(n) \leq (ne)^{kd+2k+1}.$$

- VC dimension:

$$2\lfloor k/2 \rfloor d \leq V_{\mathscr{C}_k} \leq (2kd + 4k + 2)\log_2(e(kd + 2k + 1)).$$

- Consequence: if $k \to \infty$ and $k \log n/n \to 0$, then $\mathbb{E}L(g_n^\star) \to L^\star$ for all distributions of (X, Y).
- Optimization: gradient descent (backpropagation) $\to$ smooth σ needed.
- But careful: there exists an activation σ that is monotone increasing, continuous, concave on $(0, \infty)$ and convex on $(-\infty, 0)$, such that $V_{\mathscr{C}_k} = \infty$ for each $k \geq 8$...

Quantization and clustering

References

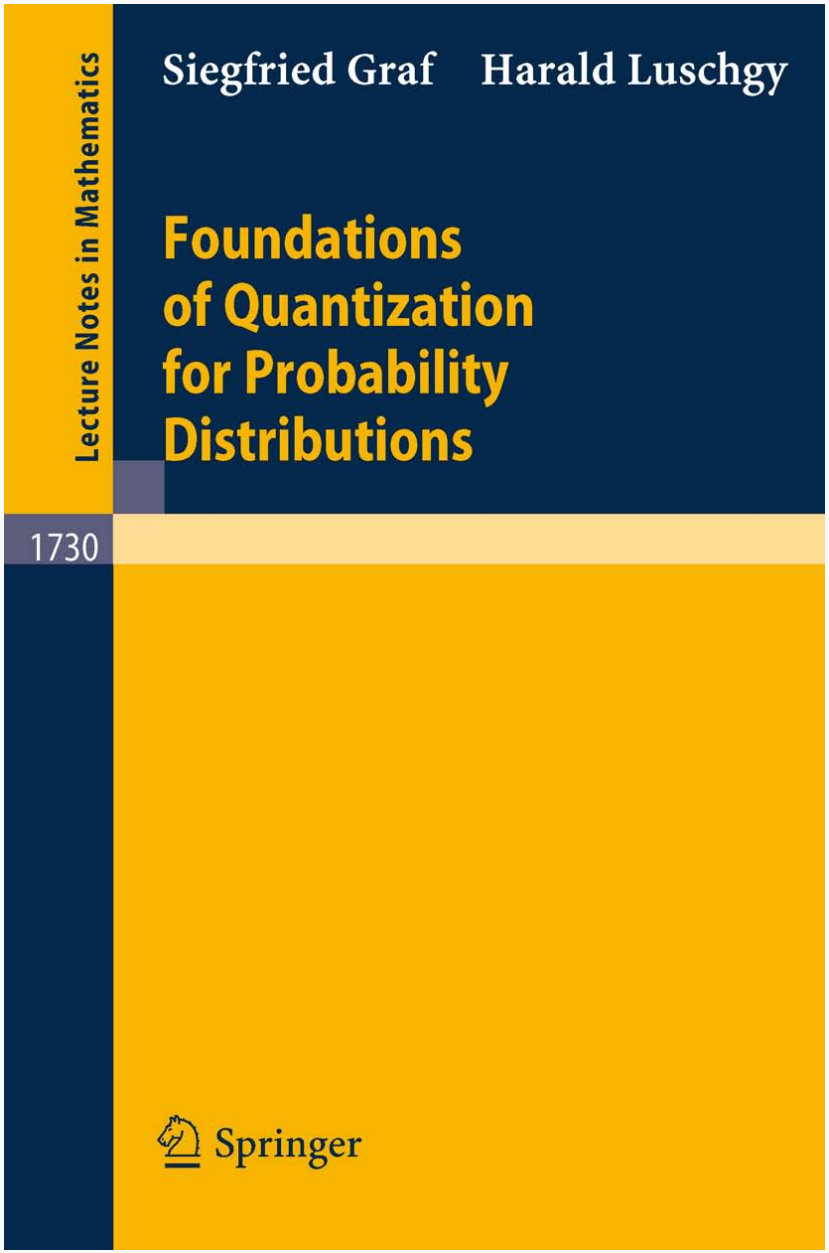
Learning-Theoretic Methods
in Vector Quantization

Lecture Notes for the
Advanced School on the Principles of Nonparametric Learning
Udine, Italy, July 9-13, 2001.

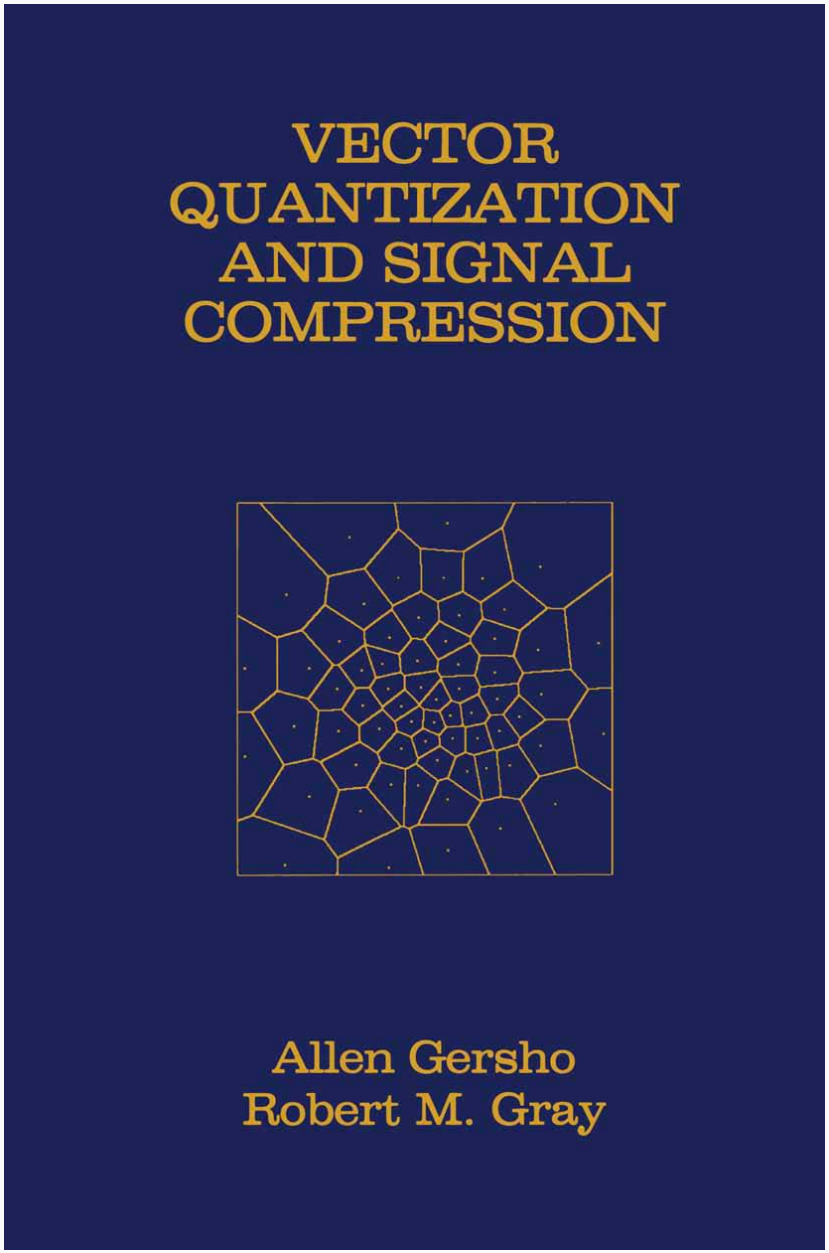
To appear in:
Principles of Nonparametric Learning
L. Györfi, editor, CISM Lecture Notes, Wien, New York: Springer 2001.

Tamás Linder
Department of Mathematics & Statistics
and
Department of Electrical & Computer Engineering
Queen's University
Kingston, Ontario, Canada, K7L 3N6
email: linder@mast.queensu.ca

References


Lecture Notes in Mathematics
Siegfried Graf Harald Luschgy
Foundations of Quantization for Probability Distributions
1730
Springer

References


VECTOR
QUANTIZATION
AND SIGNAL
COMPRESSION
Allen Gersho
Robert M. Gray

Quantization

- Quantization: probabilistic principle to compress information.
- Context: a random variable X taking values in $(\mathbb{R}^d, \|\cdot\|)$.
- $\|\cdot\|$ = Euclidean norm.
- Assumption: $\mathbb{E}\|X\|^2 < \infty \Leftrightarrow \int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x) < \infty$.

Definition (Quantizer)

Let $k \geq 1$ be an integer. A quantizer q of order k is a Borel measurable function $q : \mathbb{R}^d \to \mathscr{C} \subseteq \mathbb{R}^d$, with $|\mathscr{C}| \leq k$.

- A quantizer q of order k is characterized by:
  - ▷ A codebook $\mathscr{C} = \{c_1, \ldots, c_k\}$;
  - ▷ A partition $\mathscr{P} = \{A_1, \ldots, A_k\}$ of $\mathbb{R}^d$, with $q(x) = c_j \Leftrightarrow x \in A_j$.
- Notation: $q = (\mathscr{C}, \mathscr{P})$.

Quality measure of compression

Definition (Distortion)

The distortion (for X or μ) of a quantizer $q = (\mathscr{C}, \mathscr{P})$ of order k is

$$D(\mu, q) = \mathbb{E}\|X - q(X)\|^2 = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu(\mathrm{d}x).$$

The minimal distortion at the order k is $D_k^\star(\mu) = \inf_q D(\mu, q)$, where the infimum is taken over all quantizers of order k.

- ▷ The smaller the distortion, the better the compression.
- ▷ The compression quality improves with k.

Lemma

One has $D_k^\star(\mu) \downarrow 0$ as $k \to \infty$.

Nearest neighbor (NN) quantizers

- Context: quantizers of order k.
- Voronoi partition: for $\mathscr{C} = \{c_1, \ldots, c_k\}$, the Voronoi partition is

$$A_1 = \left\{x \in \mathbb{R}^d : \|x - c_1\| \leq \|x - c_\ell\|, \forall \ell = 1, \ldots, k\right\}, \text{ and}$$

$$A_j = \left\{x \in \mathbb{R}^d : \|x - c_j\| \leq \|x - c_\ell\|, \forall \ell = 1, \ldots, k\right\} \setminus \bigcup_{t=1}^{j-1} A_t,$$

for $2 \leq j \leq k$.

Definition (NN quantizer)

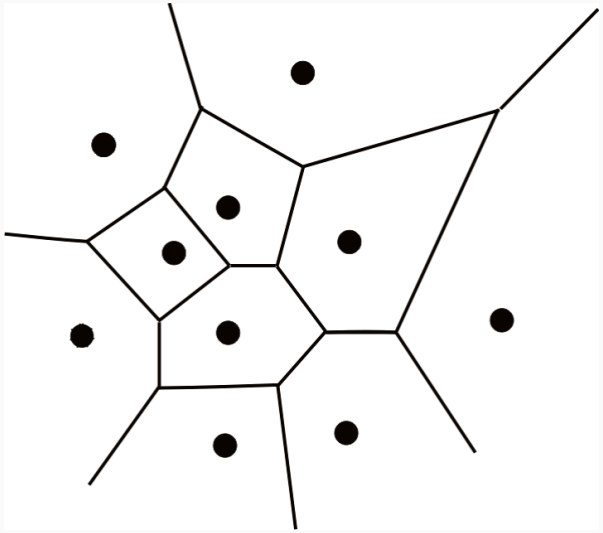
A quantizer of order k is a NN quantizer if its partition is the Voronoi partition associated with its codebook. Thus, a NN quantizer takes the form $q = (\mathscr{C}, \mathscr{P}_{\mathrm{V}}(\mathscr{C}))$, where $|\mathscr{C}| \leq k$.

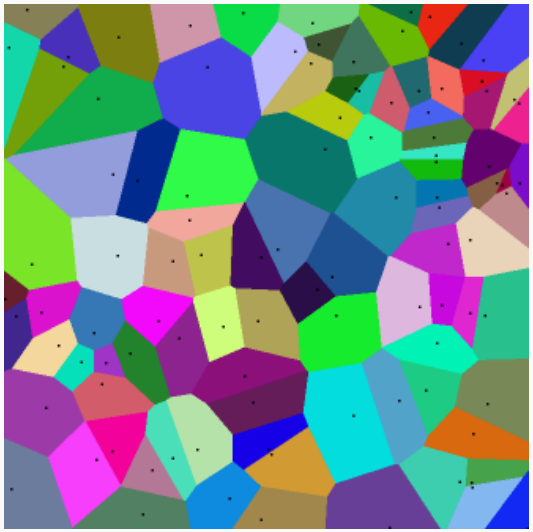
- ▷ A NN quantizer is entirely characterized by its codebook, via the rule

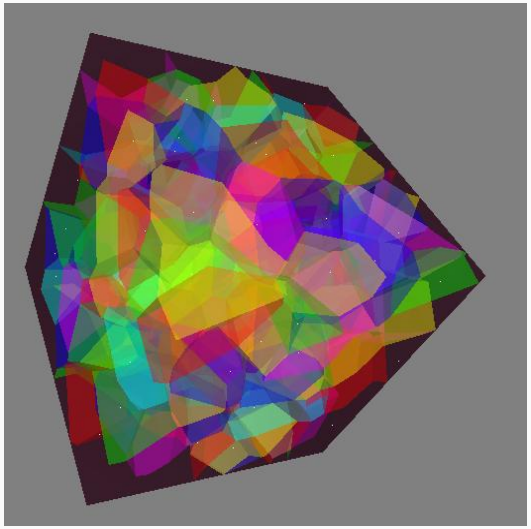
$$\|x - q(x)\| = \min_{c_j \in \mathscr{C}} \|x - c_j\|.$$

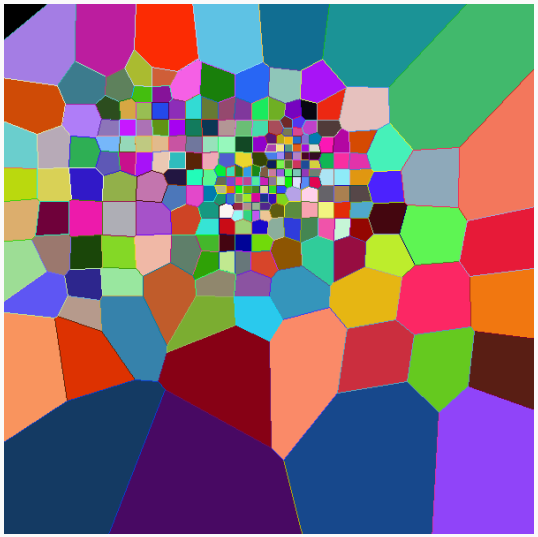
- ▷ Vocabulary: the c_j are the centers or the centroids.

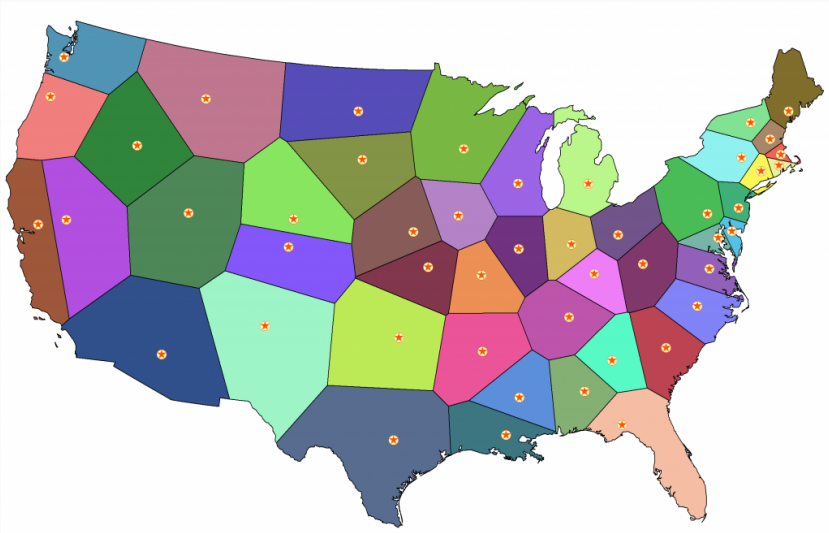
Voronoi diagrams



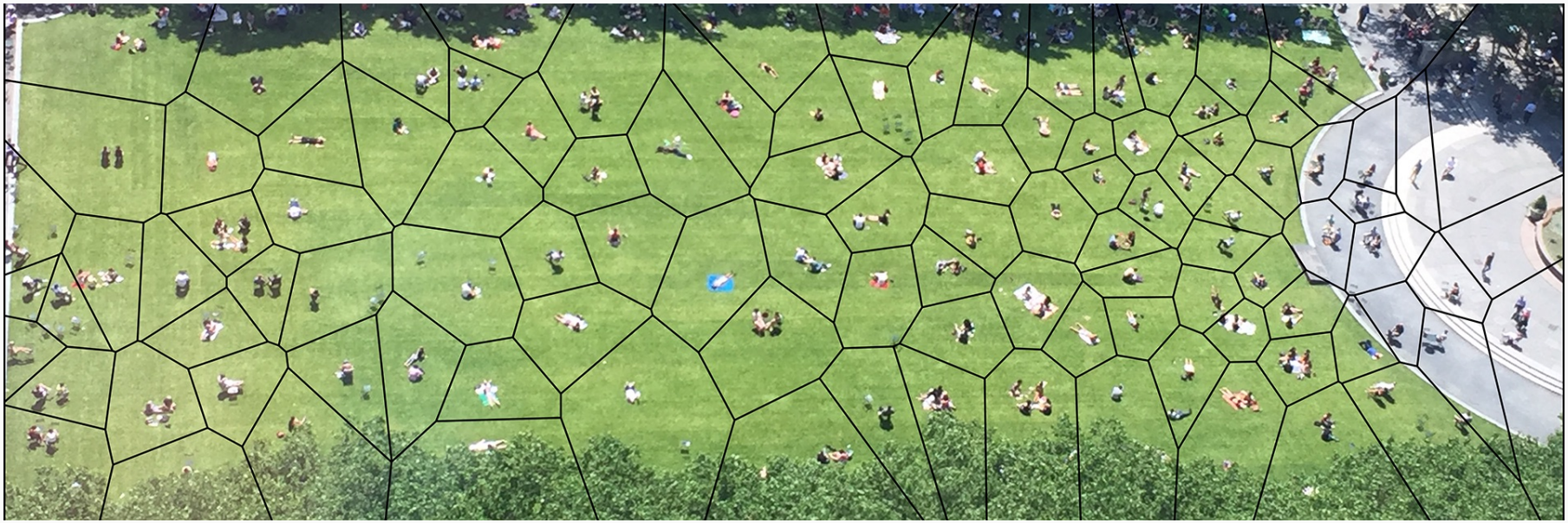








Bryant park



Properties of NN quantizers

Proposition

Let q_{NN} be a NN quantizer with codebook $\mathscr{C} = \{c_1, \ldots, c_k\}$. Then

$$D(\mu, q_{\mathrm{NN}}) = \mathbb{E} \min_{1 \le j \le k} \|X - c_j\|^2 = \int_{\mathbb{R}^d} \min_{1 \le j \le k} \|x - c_j\|^2 \mu(\mathrm{d}x).$$

In addition, for any quantizer $q = (\mathscr{C}, \mathscr{P})$, $D(\mu, q_{\mathrm{NN}}) \le D(\mu, q)$.

- If quantizers with minimal distortion exist, they are NN quantizers.
- Notation: $q_{\mathsf{NN}} = (\mathsf{c}, \mathscr{P}_{\mathrm{V}}(\mathsf{c}))$, with $\mathsf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ and distortion

$$W(\mu, \mathsf{c}) \stackrel{\mathrm{def}}{=} D(\mu, q_{\mathsf{NN}}).$$

Theorem

There exists a quantizer with minimal distortion.

Empirical quantization

- In practice, the distribution of X is unknown.
- Sample: $X_1, \ldots, X_n$ i.i.d., distributed as (and independent of) X.
- Objective: construct a good $q_n(\cdot) = q_n(\cdot\,; X_1, \ldots, X_n)$.
- The distortion of q_n is naturally defined by

$$D(\mu, q_n) = \mathbb{E}\big(\|X - q_n(X)\|^2 \,|\, X_1, \ldots, X_n\big) = \int_{\mathbb{R}^d} \|x - q_n(x)\|^2 \mu(\mathrm{d}x).$$

- Empirical measure: $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.
- Empirical distortion:

$$D(\mu_n, q) = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu_n(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^n \|X_i - q(X_i)\|^2.$$

- For $q_{\mathrm{NN}} = (\mathrm{c}, \mathscr{P}_{\mathrm{V}}(\mathrm{c}))$, with $\mathrm{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$,

$$D(\mu_n, q_{\mathrm{NN}}) = W(\mu_n, \mathrm{c}) = \frac{1}{n}\sum_{i=1}^n \min_{1\le j\le k} \|X_i - c_j\|^2.$$

Consistent quantization

Definition (Consistency)

A quantizer is consistent if

$$\mathbb{E}D(\mu, q_n) \to D^\star(\mu) \quad \text{as } n \to \infty.$$

It has rate of convergence $(v_n)_n$ *if*

$$\mathbb{E}D(\mu, q_n) - D^\star(\mu) = \mathrm{O}(1/v_n), \quad \text{with } v_n \to \infty.$$

- Exercise: consistency $\Leftrightarrow D(\mu, q_n) \xrightarrow{L^1} D^\star(\mu)$.
- Natural choice: $q_n^\star$ that minimizes the empirical distortion over all NN quantizers.
- Definition: $\mathrm{c}_n^\star = (c_{n,1}^\star, \ldots, c_{n,k}^\star)$ such that

$$W(\mu_n, \mathrm{c}_n^\star) = \inf_{\mathrm{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathrm{c}).$$

So,

$$q_n^\star = \left(\mathrm{c}_n^\star, \mathscr{P}_{\mathrm{V}}(\mathrm{c}_n^\star)\right).$$

Quantization and clustering

- $q_n^\star$ allows a clustering of $X_1, \ldots, X_n$ into k groups.
- Principle: X_i is affected to group j if $q_n^\star(X_i) = j$.
- Cluster $\sharp j$ = the X_i such that $\|X_i - c_{n,j}^\star\| \leq \|X_i - c_{n,\ell}^\star\|$, $\forall \ell = 1, \ldots, k$.

Clustering = unsupervised learning.

Towards an algorithm

- Computation of $q_n^\star$ is often a NP hard problem $\rightarrow$ k-means algorithm.
- Basic idea: for $\mathscr{C} = \{c_1, \ldots, c_k\}$ and $\mathscr{P} = \{A_1, \ldots, A_k\}$, let $q = (\mathscr{C}, \mathscr{P})$ and $q_n = (\mathscr{C}_n, \mathscr{P})$, with $\mathscr{C}_n = \{c_{n,1}, \ldots, c_{n,k}\}$ such that

$$c_{n,j} = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \sum_{i=1}^n \|X_i - y\|^2 \mathbb{1}_{[X_i \in A_j]} = \frac{\sum_{i=1}^n X_i \mathbb{1}_{[X_i \in A_j]}}{\sum_{i=1}^n \mathbb{1}_{[X_i \in A_j]}}, \ 1 \leq j \leq k.$$

Then

$$\begin{aligned} D(\mu_n, q) &= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^n \|X_i - c_j\|^2 \mathbb{1}_{[X_i \in A_j]} \\ &\geq \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^n \|X_i - c_{n,j}\|^2 \mathbb{1}_{[X_i \in A_j]} \\ &= D(\mu_n, q_n). \end{aligned}$$

k-means algorithm

1. Initialization: $\mathscr{C}^{(1)} = \{c_1^{(1)}, \ldots, c_k^{(1)}\}$ and $\mathscr{P}_V^{(1)} = \{A_1^{(1)}, \ldots, A_k^{(1)}\}$.
2. Lloyd's iteration: compute $\mathscr{C}^{(\ell+1)} = \{c_1^{(\ell+1)}, \ldots, c_k^{(\ell+1)}\}$ from $\mathscr{C}^{(\ell)} = \{c_1^{(\ell)}, \ldots, c_k^{(\ell)}\}$ via the iteration

$$c_j^{(\ell+1)} = \frac{\sum_{i=1}^n X_i \mathbb{1}_{[X_i \in A_j^{(\ell)}]}}{\sum_{i=1}^n \mathbb{1}_{[X_i \in A_j^{(\ell)}]}}, \quad 1 \leq j \leq k,$$

   where $\{A_1^{(\ell)}, \ldots, A_k^{(\ell)}\}$ is the Voronoi partition associated with $\mathscr{C}^{(\ell)}$.
3. The algorithm stops after a finite number of iterations.
4. Warning: the output codebook is not $\mathrm{c}_n^\star$.

Consistency of $q_n^\star$

- Reminder: $\mathrm{c}_n^\star = (c_{n,1}^\star, \ldots, c_{n,k}^\star)$ such that

$$W(\mu_n, \mathrm{c}_n^\star) = \inf_{\mathrm{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathrm{c}).$$

  So,

$$q_n^\star = (\mathrm{c}_n^\star, \mathscr{P}_{\mathrm{V}}(\mathrm{c}_n^\star)).$$

- Next step: prove the consistency of $q_n^\star$.

Definition (Wasserstein distance)

Let ν_1 and ν_2 be probability measures on $\mathbb{R}^d$ with finite second moment. The Wasserstein distance ρ_W between ν_1 and ν_2 is

$$\rho_W(\nu_1, \nu_2) = \inf_{X \overset{\mathscr{D}}{=} \nu_1, Y \overset{\mathscr{D}}{=} \nu_2} \sqrt{\mathbb{E}\|X - Y\|^2}.$$

Wasserstein distance

- Property 1: There exists (X_0, Y_0) such that $X_0 \overset{\mathscr{D}}{=} \nu_1$, $Y_0 \overset{\mathscr{D}}{=} \nu_2$, and

$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

- Property 2: One has $\rho_W(\nu_n, \nu) \to 0$ if

$$\nu_n \Rightarrow \nu \quad \text{and} \quad \int_{\mathbb{R}^d} \|x\|^2 \nu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^2 \nu(\mathrm{d}x).$$

Proposition

Let ν_1 and ν_2 be probability measures on $\mathbb{R}^d$ with finite second moment. If q is a NN quantizer, then

$$\left| D(\nu_1, q)^{1/2} - D(\nu_2, q)^{1/2} \right| \leq \rho_W(\nu_1, \nu_2).$$

Consistency of $q_n^\star$

Theorem

One has $D(\mu, q_n^\star) \to D^\star(\mu)$ almost surely, and $\mathbb{E}D(\mu, q_n^\star) \to D^\star(\mu)$.

- ▷ Conclusion: the quantizer $q_n^\star$ is consistent.

Theorem

If $\|X\| \leq R$ with probability one, then

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

- ▷ $\|X\| \leq R$ is called the peak power constraint.
- ▷ Take-home message: the rate of convergence does not depend on d!