Concentration of Measure for Machine Learning

An Introduction

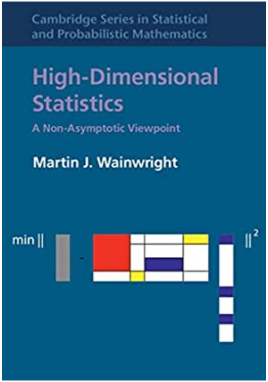
Aurélien Garivier

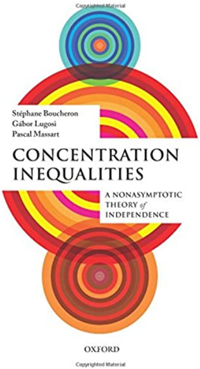
École Normale Supérieure de Lyon, UMPA & LIP

October 21st, 2021

Outline

References





Journal of Machine Learning Research 4 (2003) 895-911 Submitted 3/03; Published 10/03

Concentration Inequalities for the Missing Mass and for Histogram Rule Error

David McAllester MCALLESTER@TTI-C.ORG
Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago Il, 60637

Luis Ortiz LEORTIZ@LINC.CIS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

Editors: Ralf Herbrich and Thore Graepel

Abstract

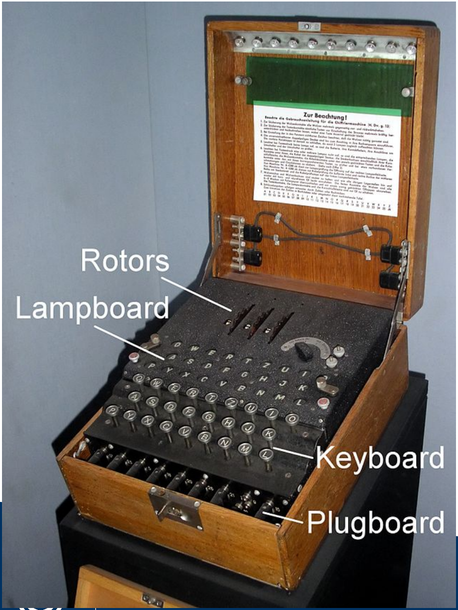
This paper gives distribution-free concentration inequalities for the missing mass and the error rate of histogram rules. Negative association methods can be used to reduce these concentration problems to concentration questions about independent sums. Although the sums are independent, they are highly heterogeneous. Such highly heterogeneous independent sums cannot be analyzed using standard concentration inequalities such as Hoeffding's inequality, the Angluin-Valiant bound, Bernstein's inequality, Bennett's inequality, or McDiarmid's theorem. The concentration inequality for histogram rule error is motivated by the desire to construct a new class of bounds on the generalization error of decision trees.

1. Introduction

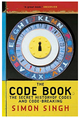
The Good-Turing missing mass estimator was developed in the 1940s to estimate the probability that the next item drawn from a fixed distribution will be an item not seen before. Since the publication of the Good-Turing missing mass estimator in 1953 (Good, 1953), this estimator has been used extensively in language modeling applications (Chen and Goodman, 1998, Church and Gale, 1991,

Outline

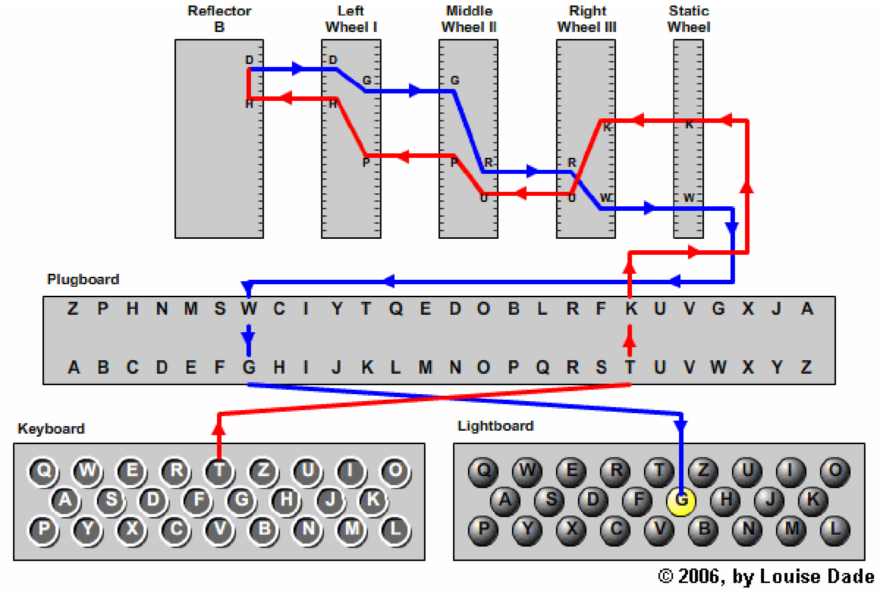
Enigma



- Electro-mechanical rotor cipher machines, 26 characters
- Invented at the end of WW1 by Arthur Scherbius
- Commercial use, then German Army during WW2
- First cracked by Marian Rejewski in the 1930s (Bomb), then improved to 3.10^{114} configurations
- Read Simon Singh, *The Code Book*

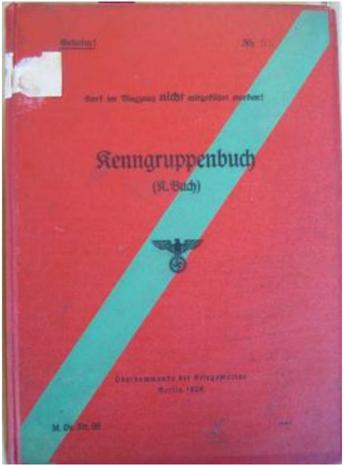


Enigma



Reflector B
Left Wheel I
Middle Wheel II
Right Wheel III
Static Wheel
D
H
D
G
H
P
G
P
R
U
K
R
U
W
K
W
Plugboard
Z P H N M S W C I Y T Q E D O B L R F K U V G X J A
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Keyboard
Q W E R T Z U I O
A S D F G H J K
P Y X C V B N M L
Lightboard
Q W E R T Z U I O
A S D F G H J K
P Y X C V B N M L
© 2006, by Louise Dade

Battle of the Atlantic



- Massively used by the German Kriegsmarine and Luftwaffe
- weakness: 3-letters setting to initiate communication, taken from the *Kenngruppenbuch*
- Government Code and Cypher School: Bletchley Park (on the train line between Cambridge and Oxford)
- Colossus (first programmable computers) in 1943

Estimating probabilities

- Discrete alphabet A.
- Unknown probability p on A
- Sample $X_1, \ldots, X_n$ of independent draws of p.
- Goal : use the sample to estimate $p(a)$ for all $a \in A$.

Natural idea:

$$\hat{p}(a) = \frac{N(a)}{n}, \quad \text{where } N(a) = \#\{i : X_i = a\}$$

[Src: Alon Orlitsky, https://ic.epfl.ch/files/content/sites/ic/files/Inka/Orlitsky%20Talk%202016.pdf]

Safari preparation

Observe animal sample

1 giraffe, 2 elephants, 3 zebras

Probability estimation?

Empirical frequency

Species	Probability
giraffes	1/6
elephants	2/6
zebras	3/6

[Src: Alon Orlitsky, https://ic.epfl.ch/files/content/sites/ic/files/Inka/Orlitsky%20Talk%202016.pdf]

ENS DE LYON

Bigram Model for NLP

[Src: https://nlp.stanford.edu/~wcmac]

Learning set:

- john read moby dick
- mary read a different book
- she read a book by cher

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)} \qquad p(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$p($	$john$	$read$	a	$book$	$)$
$=$	$p(john\|\cdot)$	$p(read\|john)$	$p(a\|read)$	$p(book\|a)$	$p(\cdot\|book)$
$=$	$\frac{c(\cdot\ john)}{\sum_w c(\cdot\ w)}$	$\frac{c(john\ read)}{\sum_w c(john\ w)}$	$\frac{c(reada)}{\sum_w c(read\ w)}$	$\frac{c(a\ book)}{\sum_w c(a\ w)}$	$\frac{c(book\ \cdot)}{\sum_w c(book\ w)}$
$=$	$\frac{1}{3}$	$\frac{1}{1}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{2}$
$\approx$	0.06				

Bigram Model for NLP

[Src: https://nlp.stanford.edu/~wcmac]

Learning set:

john read moby dick

mary read a different book

she read a book by cher

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)} \qquad p(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$p($	$cher$	$read$	a	$book$	$)$
$=$	$p(cher\|\cdot)$	$p(read\|cher)$	$p(a\|read)$	$p(book\|a)$	$P(\cdot\|book)$
$=$	$\frac{c(\cdot\ cher)}{\sum_w c(\cdot\ w)}$	$\frac{c(cher\ read)}{\sum_w c(cher\ w)}$	$\frac{c(reada)}{\sum_w c(read\ w)}$	$\frac{c(a\ book)}{\sum_w c(a\ w)}$	$\frac{c(book\ \cdot)}{\sum_w c(book\ w)}$
$=$	$\frac{0}{3}$	$\frac{0}{1}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{2}$
$=$	0				

$\Longrightarrow$ useless, the unseen **must** be treated correctly.

Bayesian Approach: Laplace Estimator

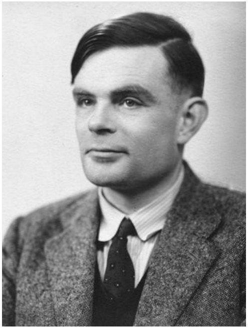
Pierre-Simon de Laplace (1749-1827), Thomas Bayes (1702-1761)
Will the sun rise tomorrow?

$$\hat{p}(a) = \frac{N(a) + 1}{n + |A|}$$

- good for small alphabets and many samples
- very bad when lots of items seen once (ex: DNA sequences)
- $|A|$ can be very large (or even infinite), but p concentrated on few items

$\Longrightarrow$ not a satisfying solution to the problem

Alan Turing



1912-1954

student of Godfrey Harold Hardy in Cambridge

PhD from Princeton with Alonzo Church

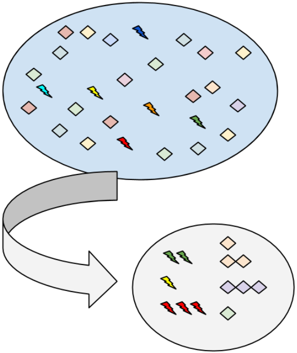
Irving J. Good

1916-2009

Graduated in Cambridge

Academic carrer in Bayesian statistics in Manchester and then in the University of Virginia (USA)

Missing mass estimation



$X_1, \ldots, X_n$ independent draws of $p \in \mathfrak{M}_1(A)$.

$$N_n(x) = \sum_{m=1}^{n} \mathbb{1}\{X_m = x\}$$

How to 'estimate' the total mass of the *unseen* items

$$M_n = \sum_{x \in A} p(x) \, \mathbb{1}\{N_n(x) = 0\} \, ?$$

Missing Mass

Let $A = \mathbb{N}$, let $p \in \mathcal{M}_1(\mathbb{N})$ and let $X_1, \ldots, X_n \overset{iid}{\sim} p$ and for every $x \in \mathbb{N}$, let $N_n(x) = \sum_{i=1}^n \mathbb{1}\{X_i = x\}$.
Pb: estimate the mass of the unseen

$$M_n = \mathbb{P}\left(X_{n+1} \notin \{X_1, \ldots, X_n\} \middle| X_1^n\right) = \sum_{x=0}^{\infty} p(x)\, \mathbb{1}\left\{N_n(x) = 0\right\}$$

Idea: use *hapaxes* = symbols $x \in \mathbb{N}$ that appear once in the sample

$$\hat{M}_n = \frac{1}{n} \sum_{x=0}^{\infty} \mathbb{1}\left\{N_n(x) = 1\right\}$$

= Good-Turing 'estimator'
= *leave-one-out* estimator of M_n: if $X_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$,

$$\hat{M}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{X_i \notin X_{-i}\right\}$$

’Bias’ of the Good-Turing estimator

Proposition [Good ’1953]

Whatever the law p,

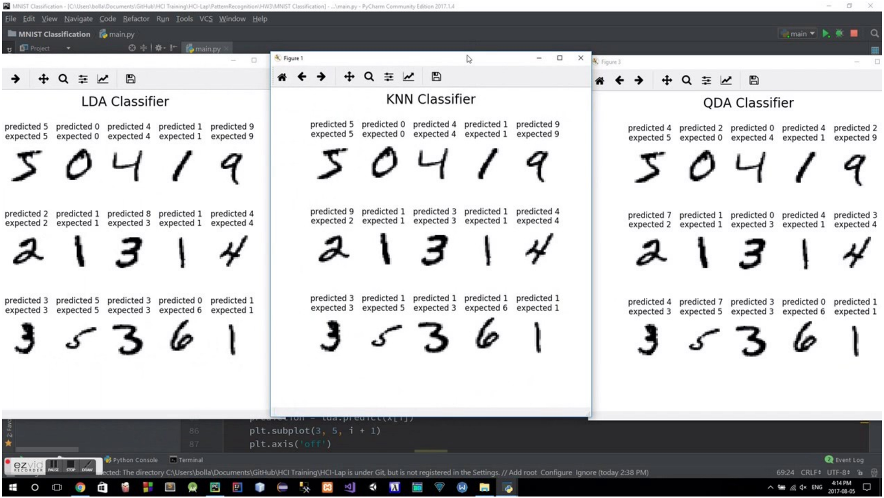
$$0 \leq \mathbb{E}\left[\hat{M}_n\right] - \mathbb{E}[M_n] \leq \frac{1}{n}$$

Proof:

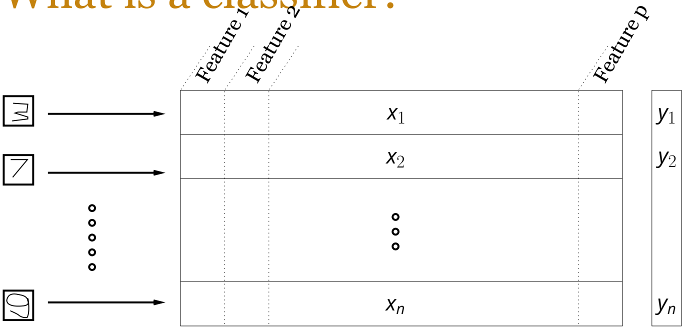
$$\begin{aligned}
\mathbb{E}\left[\hat{M}_n\right] - \mathbb{E}[M_n] &= \frac{1}{n}\mathbb{E}\left[\sum_{x\in\mathbb{N}} \mathbb{1}\{N_n(x) = 1\}\right] - \mathbb{E}\left[\sum_{x\in\mathbb{N}} p(X)\mathbb{1}\{N_n(x) = 0\}\right] \\
&= \frac{1}{n}\sum_{x\in\mathbb{N}} \mathbb{P}\big(N_n(x) = 1\big) - np(x)\ \mathbb{P}(N_n(x) = 0) \\
&= \frac{1}{n}\sum_{x\in\mathbb{N}} np(x)\big(1 - p(x)\big)^{n-1} - np(x)\big(1 - p(x)\big)^{n} \\
&= \frac{1}{n}\sum_{x\in\mathbb{N}} p(x) \times np(x)\big(1 - p(x)\big)^{n-1} \\
&= \frac{1}{n}\sum_{x\in\mathbb{N}} p(x)\ \mathbb{P}\big(N_n(x) = 1\big) \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{x\in\mathbb{N}} p(x)\mathbb{1}\big(N_n(x) = 1\big)\right] \in \left[0, \frac{1}{n}\right]
\end{aligned}$$

Outline

Example: MNIST dataset



What is a classifier?



$X \in \mathcal{M}_{n,p}(\mathbb{R})$ $\qquad$ $Y \in \mathcal{Y}^n$

Data: n-by-p matrix X

- n examples = points of observations
- p features = characteristics measured for each example

Classifier $\mathcal{A}_n$

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

 $\mapsto 6$

Statistical Learning Hypothesis

Assumption

- The examples $(X_i, Y_i)_{1 \leq i \leq n}$ are iid samples of an unknown joint distribution $\mathcal{D}$;
- The points to classify later are also independent draws of the *same* distribution $\mathcal{D}$.

Hence, for every *decision rule* $h : \mathcal{X} \to \mathcal{Y}$ we can define the *risk*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) \neq Y\big) = \mathcal{D}\Big(\big\{(x,y) : h(x) \neq y\big\}\Big) .$$

The goal of the learning algorithm is to *minimize the expected risk*:

$$R_n(\mathcal{A}_n) = \mathbb{E}_{\mathcal{D}^{\otimes n}}\Bigg[L_{\mathcal{D}}\Big(\underbrace{\mathcal{A}_n\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)}_{\hat{h}_n}\Big)\Bigg]$$

for *every* distribution $\mathcal{D}$, using only the examples.

Binary Classification

- Domain $\mathcal{X}$, label space $\mathcal{Y} = \{0, 1\}$
- Unknown distribution D on $\mathcal{X} \times \mathcal{Y}$
- Sample $S = (X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} D$
- $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$ hypothesis class
- loss function $\ell(y, y') = \mathbb{1}\{y \neq y'\}$
- generalization error (loss) $L_D(h) = \mathbb{E}_D\left[\ell\big(h(X), Y\big)\right] = \mathbb{E}_D\left[h(X) \neq Y\right]$
- training error $L_S(h) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}$
- *agnostic* learning $\neq$ realizable assumption (when there exists h^* such that $L_S(h^*) = 0$)
- learning algorithm: $S \mapsto \hat{h}_n$ such that $L_D(\hat{h}_n) - \inf_{h \in \mathcal{H}} L_D(h)$ small

Performance Limit: Bayes Classifier

Consider binary classification $\mathcal{Y} = \{0,1\}$, $\eta(x) := \mathcal{D}(Y = 1 | X = x)$.

Theorem

The Bayes classifier is defined by
$h^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\} = \mathbb{1}\{\eta(x) \geq 1 - \eta(x)\} = \mathbb{1}\{2\eta(x) - 1 \geq 0\}$.
For every classifier $h : \mathcal{X} \to \mathcal{Y} = \{0,1\}$,

$$L_{\mathcal{D}}(h) \geq L_{\mathcal{D}}(h^*) = \mathbb{E}\Big[\min\big(\eta(X), 1 - \eta(X)\big)\Big] .$$

The Bayes risk $L^*_{\mathcal{D}} = L_{\mathcal{D}}(h^*)$ *is called the* ***noise*** *of the problem.*
More precisely,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\Big[\big|2\eta(X) - 1\big| \, \mathbb{1}\{h(X) \neq h^*(X)\}\Big] .$$

Extends to $|\mathcal{Y}| > 2$.

Proof

$$
\begin{aligned}
L_D(h) - L_D(h^*) &= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(\\
&\quad \mathbb{1}\{Y=1\}\Big(\mathbb{1}\{h^*(X)=1\} - \mathbb{1}\{h^*(X)=0\}\Big) \\
&\quad + \mathbb{1}\{Y=0\}\Big(\mathbb{1}\{h^*(X)=0\} - \mathbb{1}\{h^*(X)=1\}\Big)\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{Y=1\}-1\Big)\Big(2\mathbb{1}\{h^*(X)=1\}-1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{Y=1\}-1\Big)\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\}-1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\}-1\Big)\mathbb{E}\Big[2\mathbb{1}\{Y=1\}-1 \mid X\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\}-1\Big)\Big(2\mathbb{E}[\mathbb{1}\{Y=1\} \mid X]-1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\,\mathrm{sign}\Big(\eta(X)-\frac{1}{2}\Big)(2\eta(X)-1)\Big]
\end{aligned}
$$

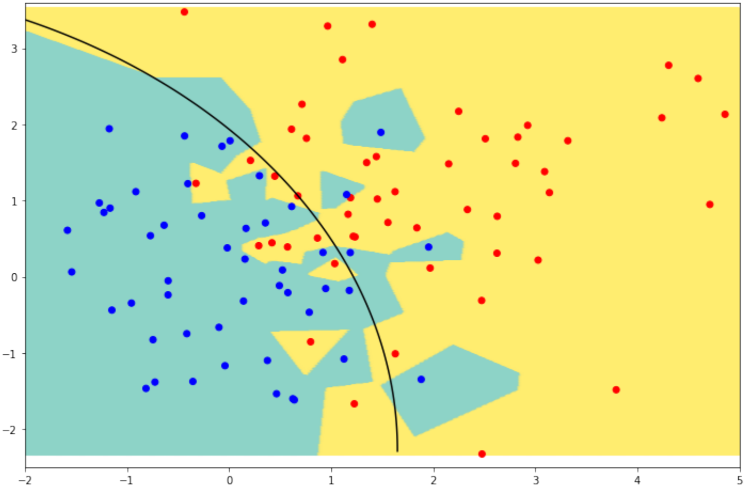
The Nearest-Neighbor Classifier

We assume that $\mathcal{X}$ is a metric space with distance d.
The nearest-neighbor classifier $\hat{h}_n^{NN} : \mathcal{X} \to \mathcal{Y}$ is defined as

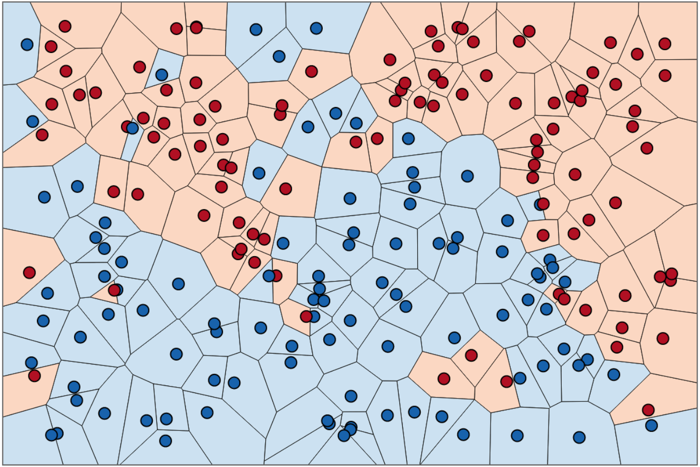
$$\hat{h}_n^{NN}(x) = Y_I \text{ where } I \in \arg\min_{1 \leq i \leq n} d(x - X_i) \; .$$

Typical distance: L^2 norm on $\mathbb{R}^d$: $\|x - x'\| = \sqrt{\sum_{j=1}^{d} (x_i - x_i')^2}$.
Buts many other possibilities: Hamming distance on $\{0, 1\}^d$, etc.

Numerically



Numerically



The most simple analysis of the most simple algorithm

A1. $\mathcal{Y} = \{0, 1\}$.

A2. $\mathcal{X} = [0, 1[^d$.

A3. η is c-Lipschitz continuous:

$$\forall x, x' \in \mathcal{X}, \left|\eta(x) - \eta(x')\right| \leq c\left\|x - x'\right\| .$$

Theorem

Under the previous assumptions, for all distributions $\mathcal{D}$ and all $m \geq 1$

$$\mathbb{E}\left[L_{\mathcal{D}}\left(\hat{h}_n^{NN}\right)\right] \leq 2L_{\mathcal{D}}^* + \frac{3c\sqrt{d}}{n^{1/(d+1)}} .$$

Proof Outline

- Conditioning: as $I(x) = \arg\min_{1\le i\le n} \|x - X_i\|$,

$$L_D(\hat{h}_n^{NN}) = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{Y \neq Y_{I(X)}\}\middle|X, X_1, \dots, X_n\right]\right] .$$

- $Y \sim \mathcal{B}(p),\ Y' \sim \mathcal{B}(q) \implies \mathbb{P}(Y \neq Y') \le 2\min(p, 1-p) + |p-q|$,

$$\mathbb{E}\left[\mathbb{1}\{Y \neq Y_{I(X)}\}|X, X_1, \dots, X_n\right] \le 2\min\left(\eta(X), 1-\eta(X)\right) + c\left\|X - X_{I(X)}\right\| .$$

- Partition $\mathcal{X}$ into $|\mathcal{C}| = T^d$ cells of diameter $\sqrt{d}/T$:

$$\mathcal{C} = \left\{\left[\frac{j_1 - 1}{T}, \frac{j_1}{T}\right[\times \cdots \times \left[\frac{j_d - 1}{T}, \frac{j_d}{T}\right[, \quad 1 \le j_1, \dots, j_d \le T\right\} .$$

- 2 cases: either the cell of X is occupied by a sample point, or not:

$$\left\|X - X_{I(X)}\right\| \le \sum_{c\in\mathcal{C}} \mathbb{1}\{X \in c\}\left(\frac{\sqrt{d}}{T}\mathbb{1}\bigcup_{i=1}^{n}\{X_i \in c\} + \sqrt{d}\mathbb{1}\bigcap_{i=1}^{n}\{X_i \notin c\}\right) .$$

$\implies \mathbb{E}\left[\left\|X - X_{I(X)}\right\|\right] \le \frac{\sqrt{d}}{T} + \frac{\sqrt{d}T^d}{e\,n}$ and choose $T = \left\lfloor n^{\frac{1}{d+1}}\right\rfloor$.

What does the analysis say?

- Is it loose? (sanity check: uniform $\mathcal{D}_X$)
- *Non-asympototic* (finite sample bound)
- The second term $\frac{3c\sqrt{d}}{n^{1/(d+1)}}$ is *distribution independent*
- Does not give the trajectorial decrease of risk
- In *expectation* only: concentrated?
- Exponential bound *d* (cannot be avoided...)

  $\Longrightarrow$ *curse of dimensionality*

- How to improve the classifier?

k-nearest neighbors

Let $\mathcal{X}$ be a (pre-compact) metric space with distance d.

k-NN classifier

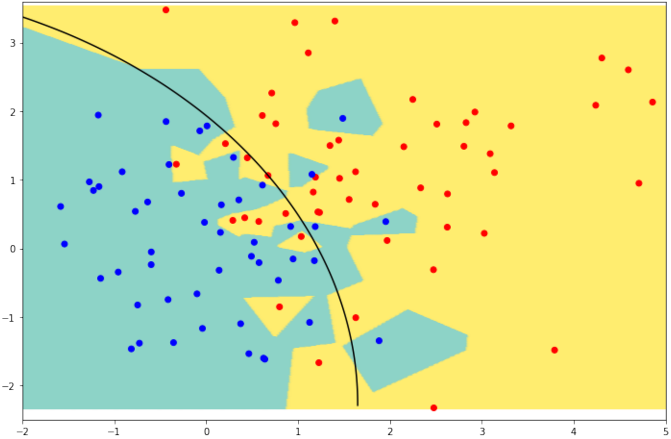
$h^{kNN} : x \mapsto \mathbb{1}\left\{\hat{\eta}(x) \geq 1/2\right\}$ = plugin for Bayes classifier with estimator

$$\hat{\eta}(x) = \frac{1}{k}\sum_{j=1}^{k} Y_{(j)}(X)$$

where

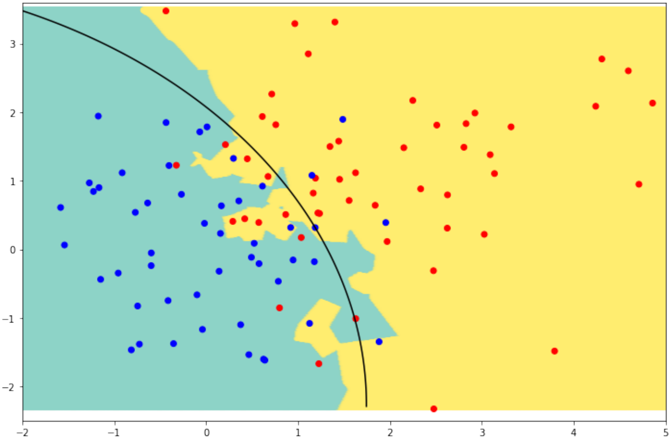
$$d\big(X_{(1)}(X), X\big) \leq d\big(X_{(2)}(X), X\big) \leq \cdots \leq d\big(X_{(n)}(X), X\big) \ .$$

More neighbors are better?



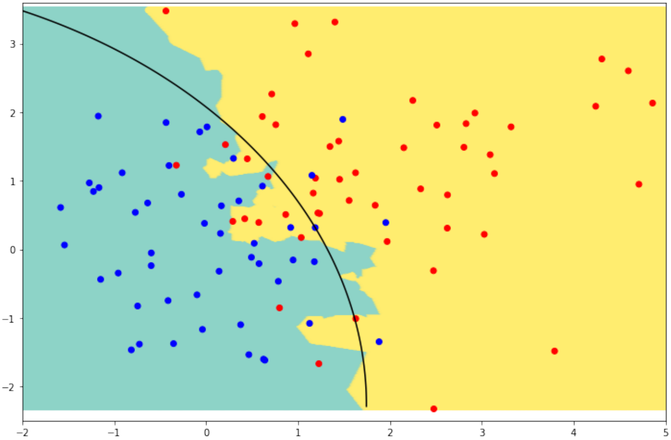
$k = 1$

More neighbors are better?



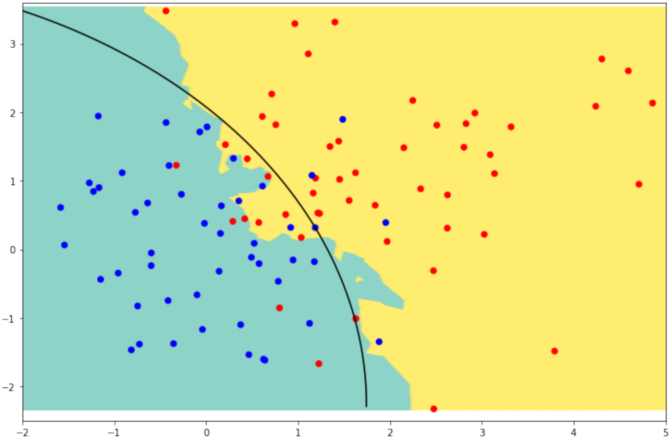
$k = 3$

More neighbors are better?



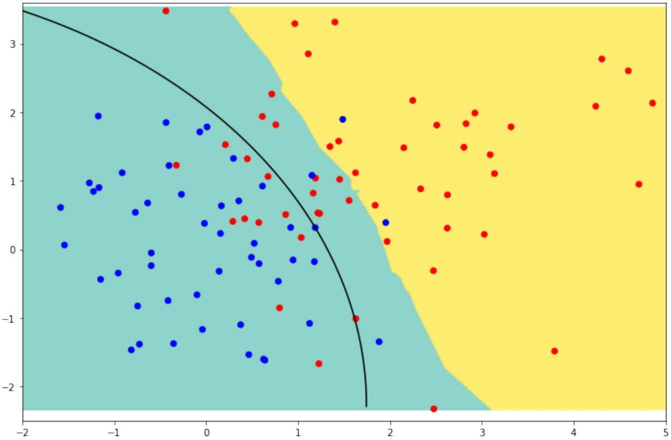
$k = 5$

More neighbors are better?



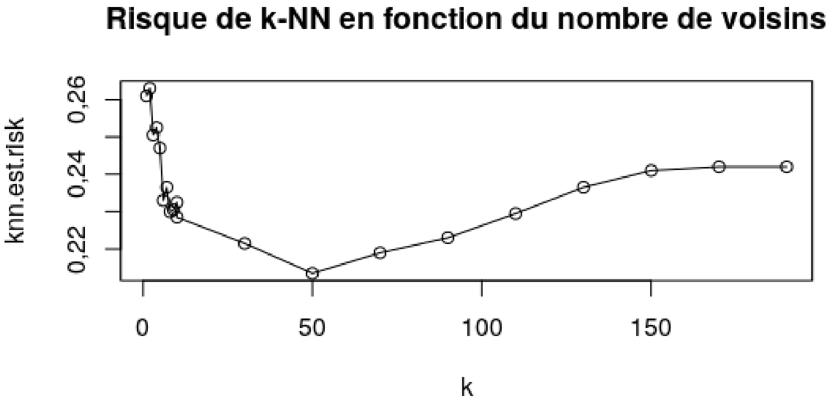
$k = 7$

More neighbors are better?



$k = 95$

Bias-Variance tradeoff



Outline

Agnostic PAC learnability

Definition

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_n$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D on $\mathcal{X} \times \mathcal{Y}$ when $S = \big((X_1, Y_1), \dots, (X_n, Y_n)\big) \overset{iid}{\sim} D$,

$$\mathbb{P}\left(L_D(\hat{h}_n) \geq \inf_{h' \in \mathcal{H}} L_D(h') + \epsilon\right) \leq \delta$$

for all $n \geq n_{\mathcal{H}}(\epsilon, \delta)$.

The smallest possible function $n_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

Learning via uniform convergence

Definition

A training set S is called ϵ-representative (wrt domain $\mathcal{X} \times \mathcal{Y}$, hypothese class $\mathcal{H}$, loss function ℓ and distribution D) if

$$\forall h \in \mathcal{H}, \left| L_S(h) - L_D(h) \right| \leq \epsilon \ .$$

Lemma

If S is $\epsilon/2$-representative, then any ERM $\hat{h}_n$ defined by $\hat{h}_n \in \arg\min_{h \in \mathcal{H}} L_S(h)$ satisfies:

$$L_D\left(\hat{h}_n\right) \leq \inf_{h \in \mathcal{H}} L_D(h) + \epsilon \ .$$

Proof: for every $h \in \mathcal{H}$,

$$L_D\left(\hat{h}_n\right) \leq L_S\left(\hat{h}_n\right) + \frac{\epsilon}{2} \leq L_S\left(h\right) + \frac{\epsilon}{2} \leq L_D\left(h\right) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \ .$$

Uniform Convergence Property

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (wrt $\mathcal{X} \times \mathcal{Y}$ and ℓ) if there exists a function $n_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every distribution D over $\mathcal{X} \times \mathcal{Y}$, a sample $S = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \overset{iid}{\sim} D$ of size $n \geq n_{\mathcal{H}}^{UC}(\epsilon, \delta)$ has probability at least $1 - \delta$ to be ϵ-representative.

Corollary

If $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostically PAC learnable with a sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq n_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$. Furthermore, the ERM is a successful PAC learner for $\mathcal{H}$.

Outline

Dimensionality reduction

- Data: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathcal{M}_{n,p}(\mathbb{R})$, $p \gg 1$.
- Dimensionality reduction: replace x_i with $y_i = Wx_i$, where $W \in \mathcal{M}_{d,p}(\mathbb{R})$, $d \ll p$.
- Hopefully, we do not loose too much by replacing x_i by y_i.
  2 approaches:
  - Quasi-invertibility: there exists a recovering matrix $U \in \mathcal{M}_{p,d}(\mathbb{R})$ such that for all $i \in \{1, \dots, n\}$,
  $$\tilde{x}_i = Uy_i \approx x_i \ .$$
  - More modest goal: distance-preserving property
  $$\forall 1 \leq i, j \leq n, \quad \|y_i - y_j\| \approx \|x_i - x_j\|$$

Johnson-Lindenstrauss Lemma

Theorem

Let $x_1, \dots, x_n \in \mathbb{R}^p$, and let $\epsilon > 0$. Then, for every $d \geq \dfrac{4\log(n)}{\epsilon - \log(1+\epsilon)}$, there exists a matrix $A \in \mathcal{M}_{d,p}(\mathbb{R})$ such that

$$\forall 1 \leq i < j \leq n, \quad (1-\epsilon)\left\|x_i - x_j\right\|^2 \leq \left\|Ax_i - Ax_j\right\|^2 \leq (1+\epsilon)\left\|x_i - x_j\right\|^2 .$$

d is independent of p (!)

on the dependence on ϵ: $\dfrac{4\log(n)}{\epsilon - \log(1+\epsilon)} \leq \dfrac{8\log(n)}{\epsilon^2}\left(1 + \dfrac{\epsilon}{3}\right)^2 .$

Remark 2: how to find such a matrix A?

For every $d \geq \dfrac{4\log(n) + 2\log(1/\delta)}{\epsilon - \log(1+\epsilon)}$, the probability that a *random matrix* with entries $A_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$ satisfies the lemma is larger than $1 - \delta$.

Random Projections

Method: (constructive) probabilistic method: we choose

$$A_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right) \ .$$

Let $y \in \mathbb{R}^p$ and $Y = Ay$. Then $\forall 1 \leq k \leq d$,

$$Y_k = \sum_{\ell=1}^{p} A_{k,\ell} y_\ell \sim \mathcal{N}\left(0, \frac{\|y\|^2}{d}\right) \ .$$

Hence $\mathbb{E}\left[\|Y\|^2\right] = \|y\|^2$.

$\implies$ does it hold with large probability?

Outline

Classical Examples

- Gaussian
- Rademacher
- Bernoulli
- Poisson

Sub-Gaussian variables.

Chernoff's Bound

Theorem (Chernoff-Hoeffding Deviation Bound)

Let $\mu \in (0,1)$. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{B}(\mu)$, and let $x \in (\mu, 1]$.

(i) *Chernoffs' bound for Bernoulli variables:* $\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(-n\,\mathrm{kl}(x,\mu)\big)$ *, where* $\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. *Same for left deviations.*

(ii) *If* $\phi(x) = \mathrm{kl}(x,\mu)$, *then* $\phi''(x) = 1/[x(1-x)]$ *and*

$$
\begin{aligned}
\mathrm{kl}(x,\mu) &= \frac{(x-\mu)^2}{2}\int_0^1 \phi''\big(\mu + s(x-\mu)\big)\ 2(1-s)ds \\
&\geq \frac{(x-\mu)^2}{2\tilde{x}(1-\tilde{x})} \quad \text{with } \tilde{x} = \frac{2\mu + x}{3} \text{ by Jensen, since } \phi'' \text{ is convex and } \int_0^1 s\,2(1-s)ds = \frac{1}{3} \\
&\geq \frac{1}{2\max_{x\leq u\leq p} u(1-u)}(x-\mu)^2 \quad \geq 2(x-\mu)^2 \ .
\end{aligned}
$$

(iii) *Hoeffding's bound for Bernoulli variables:* $\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(-2n(x-\mu)^2\big)$.

(iv) *Inequalities* (3) *and* (**??**) *hold for arbitrary independent random variables with range* $[0,1]$ *and expectation* μ. *Reason:* $\exp(\lambda x) \leq (1-x)\exp(0) + x\exp(\lambda)$.

Examples

- If $\mu < 1/2$,

$$\mathbb{P}\left(\bar{X}_k > \frac{1}{2}\right) \leq \exp\left(-\frac{k}{2}(1-2\mu)^2\right) .$$

(Consequence of Chernoff or direct computation with $(1-u)^k \leq exp(-k\,u)$, or of Hoeffding).

- For all $\mu \in [0,1]$, Chernoff's bound with $\log(u) \geq (u-1)/u$ yields

$$\mathbb{P}\left(\bar{X}_m < \frac{\mu}{2}\right) \leq \exp\left(-\frac{1-\log(2)}{2}\,m\mu\right) \approx \exp\left(-0.153\,m\mu\right) \leq \exp\left(-\frac{m\mu}{7}\right) .$$

Hoeffding yields a very poor result, but (ii) gives:

$$\mathbb{P}\left(\bar{X}_m < \frac{\mu}{2}\right) \leq \exp\left(-\frac{3}{20}m\mu\right) = \exp\left(-0.15\,m\mu\right) \leq \exp\left(-\frac{m\mu}{8}\right) .$$

Outline

Proof of the Johnson-Lindenstrauss Lemma

Method: (constructive) probabilistic method: we choose $A_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$. Let $y \in \mathbb{R}^p$ and $Y = Ay$. Then $\forall 1 \leq k \leq d$, $Y_k = \sum_{\ell=1}^{p} A_{k,\ell} y_\ell \sim \mathcal{N}\left(0, \frac{\|y\|^2}{d}\right)$. Hence $\mathbb{E}\left[\|Y\|^2\right] = \|y\|^2$.

Besides, by the deviation bound for the χ^2 distribution given in the next slide,

$$\mathbb{P}\left(\|Y\|^2 \geq (1+\epsilon)\|y\|^2\right) = \mathbb{P}\left(\sum_{k=1}^{d}\left(\frac{\sqrt{d}Y_k}{\|y\|}\right)^2 \geq d(1+\epsilon)\right) \leq \exp\left(-d\,\phi^*(\epsilon)\right) \leq \frac{1}{n^2}$$

and similarly $\mathbb{P}\left(\|Y\|^2 \leq (1-\epsilon)\|y\|^2\right) \leq \exp\left(-d\,\phi^*(\epsilon)\right) \leq \frac{1}{n^2}$.

Applying this result to all $y_{i,j} = x_i - x_j$, $1 \leq i < j \leq n$, by the union bound:

$$\mathbb{P}\left(\bigcup_{1\leq i<j\leq n} \left\|A(x_i - x_j)\right\| \geq (1+\epsilon) \cup \left\|A(x_i - x_j)\right\| \leq (1-\epsilon)\right) \leq \frac{n(n-1)}{n^2} < 1 ,$$

and hence there exists at least a matrix A for which the lemma holds.

Deviations of the χ^2 distribution: rate function

Lemma

If $U \sim \mathcal{N}(0,1)$ and $X = U^2 - 1$, then

$$\phi^*(x) = \sup_\lambda \lambda x - \log \mathbb{E}\left[e^{\lambda X}\right] = \frac{x - \log(1+x)}{2} \geq \frac{x^2}{4\left(1+\frac{x}{3}\right)^2} \ .$$

Proof: For every $\lambda < 1/2$,

$$\mathbb{E}\left[e^{\lambda X}\right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda(u^2-1)} e^{-\frac{u^2}{2}} du = \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(1-2\lambda)u^2}{2}} du = e^{-\lambda} \frac{1}{\sqrt{1-2\lambda}} \ .$$

Hence $\phi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right] = -\frac{1}{2}\log(1-2\lambda) - \lambda$. The concave function $\lambda \mapsto \lambda x - \phi(\lambda)$ is maximized at λ^* s.t. $x = \phi'(\lambda^*) = \frac{1}{1-2\lambda^*} - 1$, that is at $\lambda^* = \frac{1}{2}\left(1 - \frac{1}{1+x}\right) = \frac{x}{2(1+x)}$. Hence $\phi^*(x) = \lambda^* x - \phi(\lambda^*) = \dfrac{x - \log(1+x)}{2}$. The last inequality is obtained by "Pollard's trick" applied to $g(x) = x - \log(1+x)$: since $g(0) = g'(0) = 0$ and since $g''(x) = 1/(1+x)^2$ is convex, by Jensen's inequality

$$\frac{x - \log(1+x)}{x^2/2} = \int_0^1 g''(sx) 2(1-s) ds \geq g''\left(x \int_0^1 s\, 2(1-s) ds\right) = g''\left(\frac{x}{3}\right) \ .$$

Deviations of the $\chi^2(d)$ distribution

By Chernoff's method, if $Z \sim \chi^2(d) \overset{\text{dist}}{=} U_1^2 + \cdots + U_d^2$ where $U_i \overset{iid}{\sim} \mathcal{N}(0,1)$:

$$\mathbb{P}\big(Z \geq d(1+\epsilon)\big) \leq \exp\big(-d\phi^*(\epsilon)\big) \leq \exp\left(-\frac{d\epsilon^2}{4\left(1+\frac{\epsilon}{3}\right)^2}\right).$$

Moreover, since $\phi^*(-\epsilon) = -\frac{\epsilon+\log(1-\epsilon)}{2} = \frac{1}{2}\sum_{k\geq 2}\frac{\epsilon^k}{k} \geq \frac{1}{2}\sum_{k\geq 2}(-1)^k\frac{\epsilon^k}{k} = \phi^*(\epsilon)$,
$\mathbb{P}\big(Z \leq d(1-\epsilon)\big) \leq \exp(-d\phi^*(\epsilon))$ and since $\phi^*(-\epsilon) = -\frac{\epsilon+\log(1-\epsilon)}{2} \geq \epsilon^2/4$,

$$\mathbb{P}\big(Z \leq d(1-\epsilon)\big) \leq \exp\left(-\frac{d\epsilon^2}{4}\right).$$

Note: the Laurent-Massart inequality states that for every $u > 0$, $\mathbb{P}\big(Z \geq d + 2\sqrt{du} + 2u\big) \leq \exp\big(-u\big)$. It can be deduced from the previous bound by noting that for every $x > 0$

$$\phi^*\left(2\sqrt{x}+2x\right) = x + \frac{1}{2}\left(2\sqrt{x} - \log\left(1 + 2\sqrt{x} + \frac{(2\sqrt{x})^2}{2}\right)\right)$$

$$\geq x + \frac{1}{2}\left(2\sqrt{x} - \log\left(\exp(2\sqrt{x})\right)\right) = x \text{, and}$$

$\mathbb{P}\big(Z \geq d + 2\sqrt{du} + 2u\big) = \mathbb{P}\big(\frac{1}{d}\sum_{i=1}^d (U_i^2 - 1) \geq 2\sqrt{\frac{u}{d}} + 2\frac{u}{d}\big) \leq \exp(-d\phi^*(2\sqrt{\frac{u}{d}} + 2\frac{u}{d})) \leq e^{-u}$. The proof of Laurent and Massart (which takes elements from Birgé and Massart 1998) is a bit different: they note that

$\phi(\lambda) = -\frac{1}{2}\log(1-2\lambda) - \lambda = \sum_{k=2}^{\infty}\frac{(2\lambda)^k}{2k} = \lambda^2\sum_{\ell=0}^{\infty}\frac{4(2\lambda)^\ell}{2(\ell+2)} \leq \lambda^2\sum_{\ell=0}^{\infty}(2\lambda)^\ell = \frac{\lambda^2}{1-2\lambda}$, and deduce that

$\phi^*(x) \geq \psi^*(x) = \sup_\lambda \lambda x - \frac{\lambda^2}{1-2\lambda} = \frac{x+1-\sqrt{2x+1}}{2}$, while $x > 0$ and $\psi^*(x) = u$ implies $x = 2\sqrt{u} + 2u$. Also note in passing that by Pollard's trick $\phi^*(x) \geq \psi^*(x) \geq \frac{x^2}{4\left(1+\frac{2x}{3}\right)^{3/2}}$.

Outline

Bounded variables are sub-Gaussian

If $a \leq X \leq b$, then $\mathbb{V}\mathrm{ar}[X] \leq (b-a)^2/4$
By symmetrization, X is $(b-a)^2$ sub-Gaussian. In fact, one can prove better.

”Statistical Physics” View

Let X be a real-valued random variable with law P_X. For all $\lambda \in \mathbb{R}$, let $\phi_X(\lambda) = \ln \mathbb{E}\left[e^{\lambda X}\right]$. Then there is a largest open interval $[\lambda_{\min}, \lambda_{\max}]$ on which ϕ is defined. If it contains 0, let P_X^λ be defined by

$$\frac{dP_X^\lambda}{dP_X} = \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]} \ .$$

Then

$$\phi'(\lambda) = \mathbb{E}(P_X^\lambda) \quad \textit{and} \quad \phi''(\lambda) = \mathbb{V}\mathrm{ar}(P_X^\lambda)$$

Furthermore, let $(x_{\min}, x_{\max}) = [\lambda \mapsto \mathbb{E}(P_\lambda)](\lambda_{\min}, \lambda_{\max})$, and let $\lambda(x)$ be it reciprocal mapping. Then for every $x > \mu := \mathbb{E}[X]$, $\mathbb{P}(Z > x) \leq \exp(-I(x, \mu))$ and for every $x < \mathbb{E}[X]$, $\mathbb{P}(X < x) \leq \exp(-I(x, \mu))$ where

$$I(x, \mu) = \sup_{\lambda_{\min} < \lambda < \lambda_{\max}} \lambda x - \phi_X(\lambda) \ .$$

Gibbs-Variance lemma

For any real-valued X with expectation $\mathbb{E}[X] = \mu$, any $x \in (x_{\min}, x_{\max})$ and $\lambda \in (\lambda_{\min}, \lambda_{\max})$,

$$\phi_X(\lambda) = \lambda\mu + \int_0^\lambda \int_0^\lambda \sigma^2(t)\, dt\, du \ ,$$

and

$$\begin{aligned} I(x,\mu) &= \lambda(x)\beta(x) - \phi_X(\lambda(x)) \\ &= \mathrm{KL}\left(P_{\beta(x)}, P_X\right) = \inf_{\mathbb{E}[Q]\geq x} \mathrm{KL}(Q, P_X) \\ &= \int_\mu^x \int_\mu^u \frac{1}{\sigma^2(\lambda(t))}\, dt\, du \ . \end{aligned}$$

UNIVERSITÉ DE LYON

ENS DE LYON

Chernoff's rate function and KL divergence

Let $P = P_{M_n}$ and for $\lambda \in \mathbb{R}$ let P_λ be defined by $\frac{dP_\lambda}{dP}(x) = \frac{e^{\lambda x}}{Z(\lambda)}$, ie for all measurable, non-negative function f: $\mathbb{E}_\lambda\left[f(X)\right] = \int_{\mathbb{R}} f(x) \frac{e^{\lambda x}}{Z(\lambda)} dP(x)$

Prop:

$$\mathrm{KL}(P_\lambda, P) = \lambda \mathbb{E}_\lambda[X] - \Lambda(\lambda) = \inf\left\{ \mathrm{KL}(Q, P) : \mathbb{E}_Q[X] \geq \mathbb{E}_\lambda[X] \right\}$$

Proof: For every $Q \ll P$ with $\mathbb{E}_Q[X] \geq x$,

$$\begin{aligned}
\mathrm{KL}(Q, P) &= \int_{\mathbb{R}} \log\left(\frac{dQ}{dP}(x)\right) dQ(x) \\
&= \int_{\mathbb{R}} \log\left(\frac{dQ}{dP_\lambda}(x) \frac{dP_\lambda}{dP}(x)\right) dQ(x) \\
&= \mathrm{KL}(Q, P_\lambda) + \int_{\mathbb{R}} \log\left(\frac{e^{\lambda x}}{Z(\lambda)}\right) dQ(x) \\
&= \mathrm{KL}(Q, P_\lambda) + \lambda \mathbb{E}_Q[X] - \log\left(Z(\lambda)\right) \\
&\geq \quad 0 \quad + \lambda \mathbb{E}_\lambda[X] - \quad \Lambda(\lambda) \quad = \mathrm{KL}(P_\lambda, P)
\end{aligned}$$

Cor: since $\lambda(x)$ is such that $\mathbb{E}(P_{\lambda(x)}) = x$, $\quad I(x) = \mathrm{KL}(P_{\lambda(x)}, P)$

Chernoff's rate function and KL divergence

Let $P = P_{M_n}$ and for $\lambda \in \mathbb{R}$ let P_λ be defined by $\frac{dP_\lambda}{dP}(x) = \frac{e^{\lambda x}}{Z(\lambda)}$, ie for all measurable, non-negative function f: $\mathbb{E}_\lambda\left[f(X)\right] = \int_{\mathbb{R}} f(x) \frac{e^{\lambda x}}{Z(\lambda)} dP(x)$

Prop:
$\mathrm{KL}(P_\lambda, P) = \lambda \mathbb{E}_\lambda[X] - \Lambda(\lambda) = \inf\left\{ \mathrm{KL}(Q, P) : \mathbb{E}_Q[X] \geq \mathbb{E}_\lambda[X] \right\}$

Cor: since $\lambda(x)$ is such that $\mathbb{E}(P_{\lambda(x)}) = x$, $\quad I(x) = \mathrm{KL}(P_{\lambda(x)}, P)$

Since $\Lambda'(\lambda) = \dfrac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} = \mathbb{E}_\lambda[X]$ and

$\Lambda''(\lambda) = \dfrac{\mathbb{E}\left[X^2 e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} - \left(\dfrac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}\right)^2 = \mathbb{V}\mathrm{ar}_\lambda[X] > 0$, the C^∞ mapping

$\lambda \mapsto \lambda x - \Lambda(\lambda)$ is maximal where at $\lambda(x)$ where $x = \Lambda'\big(\lambda(x)\big) = \mathbb{E}_{\lambda(x)}[X]$ and then

$$
\begin{aligned}
I(x) &= \lambda(x)x - \Lambda\big(\lambda(x)\big)\\
&= \lambda(x)x - \Big(\lambda(x)\mathbb{E}_{\lambda(x)}[X] - \mathrm{KL}\left(P_{\lambda(x)}, P\right)\Big)\\
&= \mathrm{KL}\left(P_{\lambda(x)}, P\right)
\end{aligned}
$$

Hoeffding's inequality

A $[a, b]$-bounded variable is $(b - a)^2/4$-sub-Gaussian.

Application: Finite classes are agnostically PAC-learnable

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2} \right\rceil .$$

Moreover, $\mathcal{H}$ is agnostically PAC learnable using an ERM algorithm with sample complexity

$$n_{\mathcal{H}}(\epsilon, \delta) \leq 2n_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2\log \frac{2|\mathcal{H}|}{\delta}}{\epsilon^2} \right\rceil .$$

Proof: Hoeffding's inequality and the union bound.

Sub-Gaussian inequalities

Bennett's and Bernstein's inequalities

Let $(X_i)_{1\leq i\leq n}$ be independent random variables upper-bounded by 1, let $\bar{\mu} = (\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n])/n$, let σ^2 be such that $\mathbb{E}[X_i^2] \leq \sigma^2$ for all i and let $\phi(u) = (1+u)\log(1+u) - u$. Then, for all $x > 0$,

$$\mathbb{P}\big(\bar{X} \geq \bar{\mu} + x\big) \leq \exp\left(-n\,\sigma^2 \phi\left(\frac{x}{\sigma^2}\right)\right) \leq \exp\left(-\frac{n\,x^2/2}{\sigma^2 + x/3}\right) .$$

Bernstein from Bennett: $\phi(x) \geq \dfrac{x^2}{2\left(1+\frac{x}{3}\right)}$ since $\psi(x) = 2\left(1+\frac{x}{3}\right)\phi(x) - x^2 \geq 0.$

Extension: if $X_i \leq b$ with $b > 0$,

$$\mathbb{P}\big(\bar{X}_n \geq \bar{\mu} + x\big) \leq \exp\left(-\frac{n\sigma^2}{b^2}\phi\left(\frac{bx}{\sigma^2}\right)\right) \leq \exp\left(-\frac{n\,x^2/2}{\sigma^2 + bx/3}\right) .$$

Example: for X with range in $[0, 1]$,

$$\cdots \frac{\mu}{2}\Big) \leq \exp\left(-m\left(\frac{3}{2}\log\frac{3}{2} - \frac{1}{2}\right)\mu\right) \leq \exp\left(-\frac{3m\mu}{28}\right) .$$

Parenthesis: "Pollard's trick"

From [Pollard, MiniEmpirical ex.14, http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf]

For any sufficiently smooth real-valued function g defined at least in a neighborhood of 0 let

$$G(x) = \frac{g(x) - g(0) - xg'(0)}{x^2/2} \text{ if } x \neq 0, \text{ and } G(0) = g''(0) \ .$$

By Taylor's integral formula $g(x) - g(0) - xg'(0) = \int_0^x g''(u)(x-u)du = x^2 \int_0^1 g''(sx)(1-s)ds$.

Thus, $G(x) = \int g''(sx)d\nu(s)$, where $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$.

Hence, if g is convex then $g'' \geq 0$ and $G \geq 0$. Moreover, if g'' is increasing then the functions $x \mapsto g''(sx)$ for $s \in [0,1]$ are all increasing and G is also increasing as an average of increasing functions. For $g(u) = \exp(u)$, this yields that $(\exp(u) - u - 1)/u^2$ is increasing, as required for the proof of Bernstein's inequality.

Similarly, if g'' is convex then G is also convex as an average of convex functions $\left(x \mapsto g''(sx)\right)_s$. Moreover, by Jensen's inequality applied to convex function $\psi(s) = g''(xs)$ with the probability measure $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$

$$G(x) = \int_0^1 g''(xs)\ 2(1-s)ds \geq g''\left(x\int_0^1 s \times 2(1-s)ds\right) = g''\left(\frac{x}{3}\right) \ .$$

For $g(u) = (1+u)\log(1+u) - u$, $g''(u) = 1/(1+u)$ and this yields:

$$\frac{g(u)}{u^2/2} \geq g''\left(\frac{u}{3}\right) = \frac{1}{1+u/3} \ .$$

Exercise: for $X_i \overset{iid}{\sim} \mathcal{B}(\mu)$, $\mathbb{P}(\bar{X}_m \geq 2\mu) \leq \exp(-m \times ?)$

Chernoff + Taylor: since $\log(u) \geq (u-1)/u$,

$$\mathrm{kl}(2\mu, \mu) = 2\mu \log(2) + (1-2\mu) \log \frac{1-2\mu}{1-2\mu} \geq 2\mu \log(2) - \mu = \mu(2\log(2) - 1) \approx 0.386\,\mu\ .$$

Chernoff with convexity:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{4/3\mu} = \frac{3}{8}\,\mu = 0.375\mu\ .$$

Improved Hoeffding:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{\max_{\mu \leq u \leq 2\mu} u(1-u)} \geq \frac{\mu^2/2}{2\mu} = \frac{1}{4}\,\mu = 0.25\mu\ .$$

Bennett:

$$2\mu \log \frac{2\mu}{\mu} - (2\mu - \mu) = \mu(2\log(2) - 1) \approx 0.386\,\mu\ .$$

Bernstein:

$$\frac{(2\mu-\mu)^2/2}{\mu(1-\mu) + (2\mu-\mu)/3} \geq \frac{\mu^2/2}{\mu + \mu/3}\,\frac{3}{8}\,\mu = 0.375\mu\ .$$

Hoeffding: $2(2\mu-\mu)^2 = 2\mu^2$, very poor (as expected) when μ is small.

UNIVERSITÉ DE LYON

ENS DE LYON

Bennett's inequality

Theorem

Let $b \geq 0$ and let X be a centered variable such that $\mathbb{E}[X^2] \leq \sigma^2$. If $\mathbb{P}(X \leq b) = 1$, then for all $\lambda > 0$:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\sigma^2}{b^2}\left(e^{\lambda b} - \lambda b - 1\right)\right) .$$

Hence, if $X = X_1 + \cdots + X_n$ where the (X_i) are independent, $X_i \leq b$, $\mathbb{E}[X_i] = 0$ and $\mathbb{V}\mathrm{ar}[X_i] \leq \sigma_i^2$, then for every $x > 0$,

$$\mathbb{P}(X > x) \leq \exp\left(-\frac{\sigma^2}{b^2} H\left(\frac{bx}{\sigma^2}\right)\right)$$

with $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Bernstein's inequality

Theorem

If for all $k \geq 3, \mathbb{E}[X^k] \leq 1/2 k! \sigma^2 b^{k-2}$, then for all $\lambda \in (0, 1/b)$:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2(1-\lambda b)}\right) .$$

Hence, if $X = X_1 + \cdots + X_n$ where the (X_i) are independent and $\forall k \geq 3, \mathbb{E}[X_i^k] \leq 1/2 k! \sigma_i^2 b^{k-2}$, then for every $x > 0$,

$$\mathbb{P}(X > x) \leq \exp\left(-\frac{x^2}{2\left(\sigma^2 + xb\right)}\right)$$

with $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Proof: choose $\lambda = x/(\sigma^2 + tb)$

Remark: Bennett's condition is stronger since it implies $\mathbb{E}[X^k] \leq \mathbb{E}[X^2 b^{k-2}] \leq \sigma^2 b^{k-2}$.

Outline

Hoeffding-Azuma

Th: Let $X_0, \dots, X_n$ be a martingale such that $\forall 1 \leq k \leq n$, $|X_k - X_{k-1}| \leq c_k$. Then for all $x > 0$,

$$\mathbb{P}\big(|X_n - X_0| > x\big) \leq 2 \exp\left(-\frac{x^2}{2\sum_{k=1}^n c_k^2}\right)$$

Mc-Diarmid's ineqality

McDiarmid's inequality: If $X_1, \dots X_n$ are independent random variables on $\mathcal{X}$ and $f : \mathcal{X}^n \to \mathbb{R}$ is such that $\forall 1 \leq i \leq n, \forall x_1, \dots, x_n, x_i'$,

$$\left| f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n) \right| \leq c_i \ ,$$

then

$$\mathbb{P}\left(\left| f(X_1, \dots, X_n) - \mathbb{E}\left[f(X_1, \dots, X_n) \right] \right| \geq x \right) \leq \exp\left(\frac{-2x^2}{\sum_{i=1}^n c_i^2} \right) \ .$$

Sanity check: $f(x) = \sum x_i$

Application to the concentration of the Good-Turing estimator.

Outline

A first concentration result with Chebishev: negative correlation permits to bound the variance of M_n by $1/(en)$.

Teaser: Missing mass - negative correlation

References For Negative Association

Negative Association - Definition, Properties, and Applications, *by David Wajc*
`https://www.cs.cmu.edu/~dwajc/notes/Negative%20Association.pdf`

Balls and Bins:A Study in Negative Dependence, *by Balls and Bins:A Study in Negative Dependence*, `https://www.brics.dk/RS/96/25/BRICS-RS-96-25.pdf`

Definition

Intuitively: $X_1, \ldots, X_n$ are negatively associated when, if a subset I a variables is "high", a disjoint subset J has to be "low".

Definition

A set of real-valued random variables $X_1, X_2, \ldots, X_n$ is said to be negatively associated (NA) if for any two disjoint index sets $I, J \subset [n]$ and two functions f, g both monotone increasing or both monotone decreasing, it holds

$$\mathbb{E}\left[f(X_i : i \in I)\, g\left(X_j : j \in J\right)\right] \leq \mathbb{E}\left[f(X_i : i \in I)\right] \mathbb{E}\left[g\left(X_j : j \in J\right)\right]$$

NB: f is *monotone increasing* if $\forall i \in I, x_i \leq x'_i$ implies $f(x) \leq f(x')$.

First properties

Let $X_1, X_2, \ldots, X_n$ be NA.

- For all $i \neq j$, $\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ i.e. $\mathbf{Cov}(X_i, X_j) \leq 0$.
- For any disjoints subsets $I, J \subset [n]$ and all $x_1, \ldots, x_n$,

$$\mathbb{P}\big(X_i \geq x_i : i \in I \cup J\big) \leq \mathbb{P}\big(X_i \geq x_i : i \in I\big)\,\mathbb{P}\big(X_j \geq x_j : j \in J\big) \quad \text{and}$$
$$\mathbb{P}\big(X_i \leq x_i : i \in I \cup J\big) \leq \mathbb{P}\big(X_i \leq x_i : i \in I\big)\,\mathbb{P}\big(X_j \leq x_j : j \in J\big)$$

- For all monotone increasing functions $f_1, \ldots, f_k$ depending on disjoint subsets of the $(X_i)_i$,

$$\mathbb{E}\Big[\prod_j f_j(X)\Big] \leq \prod_j \mathbb{E}[f_j(X)]$$

- For all $x_1, \ldots, x_n$,

$$\mathbb{P}\left(\bigcap_i \{X_i \geq x_i\right) \leq \prod_i \mathbb{P}(X_i \geq x_i) \quad \text{and} \quad \mathbb{P}\left(\bigcap_i \{X_i \leq x_i\}\right) \leq \prod_i \mathbb{P}(X_i \leq x_i)$$

Consequence: NA concentrates better than independent

For Chernoff's method (which relies on exponential moments), NA variables can simply be treated as independent!
In particular:

Chernoff-Hoeffding bound

Let $X_1, \ldots, X_n$ be NA random variables with $X_i \in [a_i, b_i]$ a.s. Then $S = X_1 + \cdots + X_n$ satifies Hoeffding's tail bound: for all $t \geq 0$,

$$\mathbb{P}\Big[\big|S - E[S]\big| \geq t\Big] \leq 2\exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

Examples of NA variables

- Independent variables...
- **0-1 principle** If $X_1, \ldots X_n$ are Bernoulli variables and $\sum_i X_i \leq 1$ a.s., then they are NA.

  Let f and g are monotically increasing and depend on disjoint subsets of indices. $\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)] \iff \mathbb{E}[\tilde{f}(X)\tilde{g}(X)] \leq \mathbb{E}[\tilde{f}(X)]\mathbb{E}[\tilde{g}(X)]$, where $\tilde{f}(X) = f(X) - f(\vec{0})$ and $\tilde{g}(X) = g(X) - g(\vec{0})$. But $\tilde{f}(X)\tilde{g}(X) = 0$ always, while $\tilde{f}(X) \geq 0$ and $\tilde{g}(X) \geq 0$.
- **Permutation distributions** If $x_1 \leq \cdots \leq x_n$ and if $X_1, \ldots, X_n$ are random variables such that $\left\{X_1, \ldots, X_n\right\} = \left\{x_1, \ldots, x_n\right\}$ a.s., with all assignments equally likely, then they are NA.
- **Sampling without replacement** If $X_1, \ldots, X_n$ are sample without replacement from $\{x_1, \ldots, x_N\}$ (with $N \geq n$), then they are NA.

Closure properties

Union

If the $\{X_i : i \in I\}$ are NA, if $\{Y_j : j \in J\}$ are NA, and if the $\{X_i\}$ are independent from the $\{Y_j\}$, then the $\{X_i, Y_j : i \in I, j \in J\}$ are NA.

Concordant monotone

If the $\{X_i : i \in I\}$ are NA, if $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ are all monotonically increasing and depend on different subsets of $[n]$, then $\left\{f_j(X) : 1 \leq j \leq k\right\}$ are NA.
The same holds if $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ are all monotonically decreasing.

Bins and balls

The standard bins and balls process consists of m balls and n bins.

- each ball b is independently placed in bin i with probability $p_{b,i}$: $X_b \overset{indep}{\sim} \mathcal{M}ulti(p_{b,.})$.
- *occupancy number* $B_i = \sum_{b=1}^{m} \mathbb{1}\{X_b = i\}$ number of balls in bin i.

In particular $\displaystyle\sum_{i=1}^{n} B_i = m$.

Prop: The B_i are NA.

Let $X_{b,i} = \mathbb{1}\{\text{ball b fell into bin i}\}$. By the $0-1$ principle, for all $1 \leq b \leq m$ the $\{X_{b,i} : 1 \leq i \leq n\}$ are NA. By independence and closure under union, so are the $\{X_{b,i} : 1 \leq b \leq m, 1 \leq i \leq n\}$. By closure under concordant monotone functions, the $B_i = \displaystyle\sum_{b=1}^{m} X_{b,i}$ are NA.

Consequence: Concentration of the number $N = \sum_i \mathbb{1}\{B_i = 0\}$ of empty bins, since the $(\mathbb{1}\{B_i = 0\})_i$ are NA.
If $p_{b,i} = 1/n$, then the number N of empty bins satisfies $N = n\, e^{-m/n} \pm O\left(\sqrt{n\, e^{-m/n}}\right)$.

Applications

- missing mass
- histogram rules for binary classification

Outline

Kullback-Leibler divergence

Definition

Let P and Q be two probability distributions on a measurable set Ω. The Kullback-Leibler divergence from Q to P is defined as follows:

- if P is not absolutely continuous with respect to Q, then $\mathrm{KL}(P, Q) = +\infty$;
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

$$\mathrm{KL}(P, Q) = \int_\Omega \log \frac{dP}{dQ}\, dP = \int_\Omega \frac{dP}{dQ} \log \frac{dP}{dQ}\, dQ\ .$$

Property: $0 \leq \mathrm{KL}(P, Q) \leq +\infty$, $\mathrm{KL}(P, Q) = 0$ iff $P = Q$.

If $P \ll Q$ and $f = \frac{dP}{dQ}$, $\int_\Omega f \log(f)\, dQ = \int_\Omega \left[f \log(f)\right]_+ dQ - \int_\Omega \left[f \log(f)\right]_- dQ$, the later is finite since $\left[f \log(f)\right]_- \leq 1/e$.

Examples:

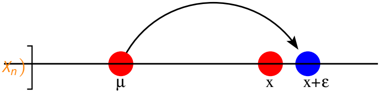
$\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = \mathrm{kl}(p, q)$, $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}\ .$

UNIVERSITÉ DE LYON

ENS DE LYON

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,

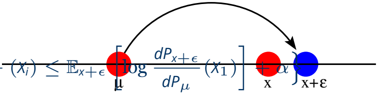


$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times \frac{d\mathbb{P}_\mu}{d\mathbb{P}_{x+\epsilon}}(X_1, \ldots, X_n)\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad \times e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\quad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon, \mu) + \alpha\right\}}\left(1 - o_n(1)\right).
\end{aligned}
$$

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,

$$\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad \times e^{-\sum_{i=1}^{n}\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right) .
\end{aligned}$$



Asymptotic Optimality (Large Deviation Lower Bound)

$$\liminf_n \frac{1}{n}\log\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \geq -\mathrm{kl}(x,\mu) .$$

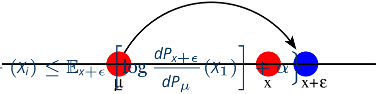
UNIVERSITÉ DE LYON

ENS DE LYON

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,

$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad\qquad \times e^{-\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\quad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right)\ .
\end{aligned}
$$



Asymptotic Optimality (Large Deviation Principle)

$$
\frac{1}{n}\log\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \underset{n\to\infty}{\longrightarrow} -\mathrm{kl}(x,\mu)\ .
$$

Properties of KL divergence

Tensorization of entropy:

If $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, then

$$\mathrm{KL}(P, Q) = \mathrm{KL}(P_1, Q_1) + \mathrm{KL}(P_2, Q_2) \ .$$

Contraction of entropy data-processing inequality:

Let $(\Omega, \mathcal{A})$ be a measurable space, and let P and Q be two probability measures on $(\Omega, \mathcal{A})$. Let $X : \Omega \to (\mathcal{X}, \mathcal{B})$ be a random variable, and let P^X (resp. Q^X) be the push-forward measures, ie the laws of X wrt P (resp. Q). Then

$$\mathrm{KL}\left(P^X, Q^X\right) \leq \mathrm{KL}(P, Q) \ .$$

Pinsker's inequality:

Let $P, Q \in \mathfrak{M}_1(\Omega, \mathcal{A})$. Then $\|P - Q\|_{TV} \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \leq \sqrt{\dfrac{\mathrm{KL}(P, Q)}{2}}$.

Proof: contraction

Contraction: if $\mathrm{KL}(P, Q) = +\infty$, the result is obvious. Otherwise, $P \ll Q$ and there exists $\frac{dP}{dQ} : \Omega \to \mathbb{R}$ such that for all measurable $f : \Omega \to \mathbb{R}$, $\int_\Omega f\, dP = \int_\Omega f \frac{dP}{dQ}\, dQ$.

- We first prove that $P^X \ll Q^X$ and, if $\gamma(x) := \mathbb{E}_Q\left[\frac{dP}{dQ}\middle| X = x\right]$ is the Q-a.s. unique function such that $\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|X\right] = \gamma(X)$, then $\gamma = \frac{dP^X}{dQ^X}$. Indeed, for all $B \in \mathcal{B}$,

$$
\begin{aligned}
P^X(B) = P(X \in B) &= \int_{X\in B} \frac{dP}{dQ} dQ = \mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X \in B\}\right] \\
&= \mathbb{E}_Q\left[\mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X \in B\}\middle|X\right]\right] = \mathbb{E}_Q\left[\mathbb{1}\{X \in B\}\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|X\right]\right] \\
&= \mathbb{E}_Q\left[\mathbb{1}\{X \in B\}\gamma(X)\right] = \int_{X\in B} \gamma(X) dQ = \int_B \gamma dQ^X
\end{aligned}
$$

  and hence $P^X \ll Q^X$ and $\frac{dP^X}{dQ^X} = \gamma$.
- Now,

$$
\begin{aligned}
\mathrm{KL}\left(P^X, Q^X\right) &= \int_{\mathcal{X}} \gamma \log \gamma \; dQ^X = \int_\Omega \gamma(X) \log \gamma(X)\; dQ \\
&= \mathbb{E}_Q\left[\phi\left(E_Q\left[\frac{dP}{dQ}\middle|X\right]\right)\right] \quad \text{where } \phi := x \mapsto x\log(x) \text{ is convex} \\
&\leq \mathbb{E}_Q\left[\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\middle|X\right]\right] \quad \text{by (conditional) Jensen's inequality} \\
&= \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right] = \mathrm{KL}(P, Q)\ .
\end{aligned}
$$

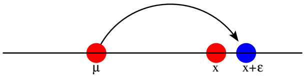
Proof: Pinsker

Let $A \in \mathcal{A}$, $p = P(A)$ and $q = Q(A)$. By contraction,

$$\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^{\mathbb{1}_A}, Q^{\mathbb{1}_A}) = \mathrm{KL}\left(\mathcal{B}\big(P(A)\big), \mathcal{B}\big(Q(A)\big)\right) = \mathrm{kl}\,\big(P(A), Q(A)\big) \geq 2\big(P(A) - Q(A)\big)^2 \ .$$

Lower Bound: the Entropic Way

Let $\Omega = \{0,1\}^n$, $X_i(\omega) = \omega_i$
Probability laws on Ω: $\mathbb{P}_p = \mathcal{B}(p)^{\otimes n}$.
For all $\epsilon > 0$,



$$n\ \mathrm{kl}(x+\epsilon, \mu) = \mathrm{KL}\left(\mathbb{P}_{x+\epsilon}, \mathbb{P}_\mu\right) \qquad \mathrm{KL}(P \otimes P', Q \otimes Q') = \mathrm{KL}(P, Q) + \mathrm{KL}(P', Q')$$

$$\geq \mathrm{KL}\left(\mathbb{P}_{x+\epsilon}^{\mathbb{1}\{\bar{X}_n \geq x\}}, \mathbb{P}_\mu^{\mathbb{1}\{\bar{X}_n \geq x\}}\right) \qquad \mathrm{KL}(P, Q) \geq \mathrm{KL}(P^X, Q^X)$$

contraction of entropy
= data-processing inequality

$$= \mathrm{kl}\left(\mathbb{P}_{x+\epsilon}\left(\bar{X}_n \geq x\right), \mathbb{P}_\mu\left(\bar{X}_n \geq x\right)\right)$$

$$\geq \mathbb{P}_{x+\epsilon}\left(\bar{X}_n \geq x\right) \log \frac{1}{\mathbb{P}_\mu\left(\bar{X}_n \geq x\right)} - \log(2) \qquad \mathrm{kl}(p, q) \geq p \log \frac{1}{q} - \log 2$$

A non-asymptotic lower bound

$$\forall \epsilon > 0, \qquad \mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \geq e^{-\frac{n\ \mathrm{kl}(x+\epsilon, \mu) + \log(2)}{1 - e^{-2n\epsilon^2}}}.$$

Outline

The No-Free-Lunch theorem

A learning algorithm A for binary classification maps a sample $S \sim \mathcal{D}^{\otimes n}$ to a decision rule $\hat{h}_n$.

Theorem

Let A be any learning algorithm for binary classification over a domain $\mathcal{X}$. If the training set size is $n \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:

- there exists a function $f : \mathcal{X} \to \{0, 1\}$ with $L_D(f) = 0$;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes n}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8} .$$

Note that the ERM over $\mathcal{H} = \{f\}$, or over any set $\mathcal{H}$ such that $n \geq 8 \log(7|\mathcal{H}|/6)$, is a successful learner in that setting.

Proof

Take $C \subset \mathcal{X}$ of cardinality $2n$, and $\{0,1\}^C = \{f_1, \ldots, f_T\}$ where $T = 2^{2n}$. For each $1 \leq i \leq T$, we denote by D_i the probability distribution on $C \times \{0,1\}$ defined by $D_i(\{x,y\}) = \begin{cases} \frac{1}{2n} \text{ if } y = f_i(x) \ , \\ 0 \text{ otherwise.} \end{cases}$

We will show that $\max_{1 \leq i \leq T} \mathbb{E}[L_{D_i}(A(S))] \geq 1/4$, which entails the result thanks to the small lemma: if $P(0 \leq Z \leq 1) = 1$ and $\mathbb{E}[Z] \geq 1/4$, then $\mathbb{P}(Z \geq 1/8) \geq 1/7$. Indeed, $1/4 \leq \mathbb{E}[Z] \leq \mathbb{P}(Z < 1/8)/8 + \mathbb{P}(Z \geq 1/8) = 1/8 - 7\,\mathbb{P}(Z \geq 1/8)/8$.
All the X-samples $S_1^X, \ldots, S_k^X$, for $k = (2n)^n$, are equaly likely. For $1 \leq j \leq k$, if $S_j^X = (x_1, \ldots, x_n)$ we denote by $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_n, f_i(x_n))$, and $\hat{f}_j^i = A(S_j^i)$.

$$\max_{1 \leq i \leq T} \mathbb{E}\left[L_{D_i}(A(S))\right] = \max_{1 \leq i \leq T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(\hat{f}_j^i\right) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(\hat{f}_j^i\right)$$

$$= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) \geq \min_{1 \leq j \leq k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) .$$

Fix $1 \leq j \leq k$, denote $S_j^X = (x_1, \ldots, x_n)$ and define $\{v_1, \ldots, v_p\} = C \setminus \{x_1, \ldots, x_n\}$, where $p \geq n$. Then

$$L_{D_i}\left(\hat{f}_j^i\right) = \frac{1}{2n} \sum_{x \in C} \mathbb{1}\{\hat{f}_j^i(x) \neq f_i(x)\} \geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\}$$

and hence $$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} \geq \frac{1}{2} \min_{1 \leq r \leq p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} .$$

Fix $1 \leq r \leq p$. Then the functions $\{f_i : 1 \leq i \leq T\}$ can be grouped into $T/2$ pairs of functions $(\tilde{f}_i^0, \tilde{f}_i^1)$, $1 \leq i \leq T/2$ which agree on all $x \in C$ except on v_r, and for all $1 \leq i \leq T/2$ it holds that $\mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^1(v_r)\} = 1$. Hence,

$\sum_{i=1}^{T} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} = \sum_{i=1}^{T/2} \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^1(v_r)\} = T/2$, which concludes the proof.

Consequence: infinite VC-dimension $\implies$ no learnability

Recall that a hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_n$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D on $\mathcal{X} \times \mathcal{Y}$ when $S = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \overset{iid}{\sim} D$,

$$\mathbb{P}\left(L_D\big(\hat{h}_n\big) \geq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon\right) \leq \delta$$

for all $n \geq n_{\mathcal{H}}(\epsilon, \delta)$.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then $\mathcal{H}$ is not PAC-learnable.

Proof: for every training size n, there exists a set $C \subset \mathcal{X}$ of size $2n$ that is shattered by $\mathcal{H}$. By the NFL theorem, for every learning algorithm A there exists a probability distribution D over $\mathcal{X} \times \{0,1\}$ and $h : \mathcal{X} \to \{0,1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.

Consequence: Curse of Dimensionality

Theorem

Let $c > 1$ be a Lipschitz constant. Let A be any learning algorithm for binary classification over a domain $\mathcal{X} = [0,1]^d$. If the training set size is $n \leq (c+1)^d/2$, then there exists a distribution $\mathcal{D}$ over $[0,1]^d \times \{0,1\}$ such that:

- $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ is c-Lipschitz;
- the Bayes error of the distribution is 0;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes n}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8}\ .$$

Shattering

Definition

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of functions $C \to \{0,1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \Big\{ (x_1, \ldots, x_m) \to \big(h(x_1), \ldots, h(x_m)\big) : h \in \mathcal{H} \Big\} .$$

Shattering

A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Example:

- $\mathcal{H} = \{ \mathbb{1}_{]-\infty,a]} : a \in \mathbb{R} \}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \{ h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1 \text{ and } a_2 \leq b_2 \}$ where

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \, ; \\ 0 & \text{otherwise} . \end{cases}$$

UNIVERSITÉ DE LYON

ENS DE LYON

VC dimension

Definition

The *Vapnik Chervonenkis dimension* $\mathrm{VCdim}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Example:

- $\mathcal{H} = \left\{\mathbb{1}_{]-\infty,a]} : a \in \mathbb{R}\right\}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \left\{\mathbb{R}^2 \ni x \mapsto \mathbb{1}_{[a_1,b_1]}(x_1)\mathbb{1}_{[a_2,b_2]}(x_2) : a_1 \leq b_1 \text{ and } a_2 \leq b_2\right\}$

Fundamental theorem of PAC learning

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Then the following propositions are equivalent:

1. $\mathcal{H}$ has the uniform convergence property,
2. any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$,
3. $\mathcal{H}$ is agnostic PAC learnable,
4. $\mathcal{H}$ has finite VC-dimension.

Fundamental theorem of PAC learning (quantitative version)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Assume that $d := \mathrm{VCdim}(\mathcal{H}) < \infty$. Then there exist constants C_1, C_2 such that:

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

Sauer's lemma

Definition

Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as the maximal number of different functions that can be obtained by restricting $\mathcal{H}$ to a set of size m: $\tau_{\mathcal{H}}(m) = \max_{C \subset X : |C| = m} \left|\mathcal{H}_C\right|$.

Note: if $\mathrm{VCdim}(\mathcal{H}) = d$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$.

Sauer's lemma

Let $\mathcal{H}$ be a hypothesis class with $d = \mathrm{VCdim}(\mathcal{H}) < \infty$. Then, for all $m \geq d$,

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d .$$

Think of example: $\mathcal{H} = \left\{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\right\}$ with $d = \mathrm{VCdim}(\mathcal{H}) = 1$.

Proof of Sauer's lemma 1/2

In fact we prove the stronger claim:

$$|\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i} .$$

where the last inequality holds since no set of size larger than d is shattered by $\mathcal{H}$. The proof is by induction.
m=1: The empty set is always considered to be shattered by $\mathcal{H}$. Hence, either $|\mathcal{H}_C| = 1$ and $d = 0$, inequality $1 \leq 1$, or $d \geq 1$ and the inequality is $2 \leq 2$.
Induction: Let $C = \{x_1, \ldots, x_m\}$, and let $C' = \{x_2, \ldots, x_m\}$. We note functions like vectors, and we define

$$Y_0 = \left\{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ or } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\right\}, \quad \text{and}$$

$$Y_1 = \left\{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ and } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\right\} .$$

Then $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Moreover, $Y_0 = \mathcal{H}_{C'}$ and hence by the induction hypothesis:

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B' \subset C' : \mathcal{H} \text{ shatters } B'\}| = |\{B \subset C : x_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}|$$

Next, define

$$\mathcal{H}' = \left\{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } \forall 1 \leq i \leq m, h'(x_i) = \begin{cases} 1 - h(x_1) \text{ if } i = 1 \\ h(x_i) \text{ otherwise} \end{cases}\right\}$$

Note that $\mathcal{H}'$ shatters $B' \subset C'$ iff $\mathcal{H}'$ shatters $B' \cup \{x_1\}$, and that $Y_1 = \mathcal{H}'_{C'}$. Hence, by the induction hypothesis,

$$\begin{aligned} |Y_1| &= |\mathcal{H}'_{C'}| \leq |\{B' \subset C' : \mathcal{H}' \text{ shatters } B'\}| = |\{B' \subset C' : \mathcal{H}' \text{ shatters } B' \cup \{x_1\}\}| \\ &= |\{B \subset C : x_1 \in B \text{ and } \mathcal{H}' \text{ shatters } B\}| \leq |\{B \subset C : x_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| . \end{aligned}$$

Overall,

$$|\mathcal{H}_C| = |Y_0| + |Y_1| \leq |\{B \subset C : x_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| + |\{B \subset C : x_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| = |\{B \subset C : \mathcal{H} \text{ shatters } B\}| .$$

Proof of Sauer's lemma 2/2

For the last inequality, one may observe that if $m \geq 2d$, defining $N \sim \mathcal{B}(m, 1/2)$, Chernoff's inequality and inequality $\log(u) \geq (u-1)/u$ yield

$$-\log \mathbb{P}(N \leq d) \geq m\,\mathrm{kl}\left(\frac{d}{m}, \frac{1}{2}\right) \geq d \log \frac{2d}{m} + (m-d)\log \frac{2(m-d)}{m}$$

$$\geq m\log(2) + d\log\frac{d}{m} + (m-d)\frac{-d/m}{(m-d)/m} = m\log(2) + d\log\frac{d}{em}\ ,$$

and hence

$$\sum_{i=0}^{d} \binom{m}{i} = 2^m \mathbb{P}(N \leq d) \leq \exp\left(-d\log\frac{d}{em}\right) = \left(\frac{em}{d}\right)^d\ .$$

Besides, for the case $d \leq m \leq 2d$, the inequality is obvious since $(em/d)^d \geq 2^m$: indeed, function $f : x \mapsto -x\log(x/e)$ is increasing on $[0,1]$, and hence for all $d \leq m \leq 2d$:

$$\frac{d}{m}\log\frac{em}{d} = f(d/m) \geq f(1/2) = \frac{1}{2}\log(2e) \geq \log(2)\ ,$$

which implies $\left(\frac{em}{d}\right)^d = \exp\left(d\log\frac{em}{d}\right) \geq \exp(m\log(2)) = 2^m$.
Alternately, you may simply observe that for all $m \geq d$,

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d}\binom{m}{i} \leq \sum_{i=0}^{d}\left(\frac{d}{m}\right)^i\binom{m}{i} \leq \sum_{i=0}^{m}\left(\frac{d}{m}\right)^i\binom{m}{i} = \left(1+\frac{d}{m}\right)^m \leq e^d\ .$$

Finite VC dimension implies Uniform Convergence

Theorem

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every distribution D dans for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of the sample $S \sim D^{\otimes n}$ we have

$$\sup_{h \in \mathcal{H}} \left| L_D(h) - L_S(h) \right| \leq \frac{1 + \sqrt{\log\left(\tau_{\mathcal{H}}(2n)\right)}}{\delta \sqrt{n/2}} .$$

Note: this result is sufficient to prove that finite VC-dim $\implies$ learnable, but the dependency in δ is not correct at all: roughly speaking, the factor $1/\delta$ can be replaced by $\log(1/\delta)$.

Proof: symmetrization and Rademacher complexity (1/2)

We consider the 0-1 loss $\ell(h,(x,y)) = \mathbb{1}\{h(x) \neq y\}$, or any $[0,1]-$valued loss ℓ. We denote $Z_i = (X_i, Y_i)$, and observe that $L_D(h) = \mathbb{E}_{Z_i}[\ell(h, Z_i)] = \mathbb{E}_{S'}[L_{S'}(h)]$ if $S' = Z'_1, \dots, Z'_n$ denotes another iid sample of D. Hence,

$$
\begin{aligned}
\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|L_D(h) - L_S(h)\right|\right] &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|\mathbb{E}_{S'}[L_{S'}(h)] - L_S(h)\right|\right] = \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|\mathbb{E}_{S'}\left[L_{S'}(h) - L_S(h)\right]\right|\right] \\
&\leq \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'}\left[\left|L_{S'}(h) - L_S(h)\right|\right]\right] \leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}} \left|L_{S'}(h) - L_S(h)\right|\right]\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \ell(h, Z'_i) - \ell(h, Z_i)\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i\left(\ell(h, Z'_i) - \ell(h, Z_i)\right)\right|\right] \quad \text{for all } \sigma \in \{\pm 1\}^n \\
&= \mathbb{E}_\Sigma \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\left(\ell(h, Z'_i) - \ell(h, Z_i)\right)\right|\right] \quad \text{if } \Sigma \sim \mathcal{U}\left(\{\pm 1\}^n\right) \\
&= \mathbb{E}_{S,S'} \mathbb{E}_\Sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\left(\ell(h, Z'_i) - \ell(h, Z_i)\right)\right|\right] .
\end{aligned}
$$

Proof: symmetrization and Rademacher complexity (1/2)

We consider the 0-1 loss $\ell(h,(x,y)) = \mathbb{1}\{h(x) \neq y\}$, or any $[0,1]-$valued loss ℓ. We denote $Z_i = (X_i, Y_i)$, and observe that $L_D(h) = \mathbb{E}_{Z_i}[\ell(h, Z_i)] = \mathbb{E}_{S'}[L_{S'}(h)]$ if $S' = Z'_1, \ldots, Z'_n$ denotes another iid sample of D. Hence,

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|L_D(h) - L_S(h)\right|\right] = \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|\mathbb{E}_{S'}[L_{S'}(h)] - L_S(h)\right|\right] = \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left|\mathbb{E}_{S'}\left[L_{S'}(h) - L_S(h)\right]\right|\right]$$

$$= \mathbb{E}_{S,S'}\mathbb{E}_{\Sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\big(\ell(h,Z'_i) - \ell(h,Z_i)\big)\right|\right] .$$

Now, for every S, S', let $C = C_{S,S'} = \left\{x : \exists i \in \{1,\ldots,n\} : x = X_i \text{ or } X'_i\right\}$. Then $\forall \sigma \in \{-1,1\}^n$,

$$\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\ell(h,Z'_i) - \ell(h,Z_i)\big)\right| = \max_{h\in\mathcal{H}_C} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\ell(h,Z'_i) - \ell(h,Z_i)\big)\right| .$$

ENS DE LYON

Proof: symmetrization and Rademacher complexity (2/2)

Moreover, for every $h \in \mathcal{H}_C$ let $Z_h = \frac{1}{n} \sum_{i=1}^n \Sigma_i \big(\ell(h, Z_i') - \ell(h, Z_i)\big)$. Then $\mathbb{E}_\Sigma[Z_h] = 0$, each summand belongs to $[-1, 1]$ and by Hoeffding's inequality, for every $\epsilon > 0$:

$$\mathbb{P}_\Sigma\left[|Z_h| \geq \epsilon\right] \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right) .$$

Hence, by the union bound,

$$\mathbb{P}_\Sigma\left[\max_{h \in \mathcal{H}_C} |Z_h| \geq \epsilon\right] \leq 2\big|\mathcal{H}_C\big| \exp\left(-\frac{n\epsilon^2}{2}\right) .$$

The following lemma permits to deduce that

$$\mathbb{E}_\Sigma\left[\max_{h \in \mathcal{H}_C} |Z_h|\right] \leq \frac{1 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{n/2}} \leq \frac{1 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{n/2}} .$$

since $|C| \leq 2n$. Hence,

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \big|L_D(h) - L_S(h)\big|\right] \leq \mathbb{E}_{S,S'} \mathbb{E}_\Sigma\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left|\sum_{i=1}^n \Sigma_i \big(\ell(h, Z_i') - \ell(h, Z_i)\big)\right|\right] \leq \frac{1 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{n/2}} ,$$

and we conclude by using Markov's inequality (poor idea! Better: McDiarmid's inequality).

Technical Lemma

Lemma

Let $a > 0$, $b > 1$, and let Z be a real-valued random variable such that for all $t \geq 0$, $\mathbb{P}(Z \geq t) \leq 2b \exp\left(-\frac{t^2}{a^2}\right)$. Then

$$\mathbb{E}[Z] \leq a\left(\sqrt{\log(b)} + \frac{1}{\sqrt{\log(b)}}\right) .$$

Proof:

$$\begin{aligned}
\mathbb{E}[Z] &\leq \int_0^\infty \mathbb{P}(Z \geq t)dt \leq a\sqrt{\log(b)} + \int_{a\sqrt{\log(b)}}^\infty 2b \exp\left(-\frac{t^2}{a^2}\right) dt \\
&\leq a\sqrt{\log(b)} + 2b \int_{a\sqrt{\log(b)}}^\infty \frac{t}{a\sqrt{\log(b)}} \exp\left(-\frac{t^2}{a^2}\right) dt \\
&= a\sqrt{\log(b)} + \frac{2b}{a\sqrt{\log(b)}} \times \frac{a^2}{2} \exp\left(-\frac{\left(a\sqrt{\log(b)}\right)^2}{a^2}\right) \\
&= a\sqrt{\log(b)} + \frac{a}{\sqrt{\log(b)}} .
\end{aligned}$$

NB: cutting at $a\sqrt{\log(2b)}$ gives a better but less nice inequality for our use.

Application: Finite VC-dim classes are agnostically learnable

It suffices to prove that finite VC-dim implies the uniform convergence property. From Sauer's lemma, for all $m \geq d/2$ we have $\tau_{\mathcal{H}}(2n) \leq (2en/d)^d$. With the previous theorem, this yields that with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{H}} \left| L_D(h) - L_S(h) \right| \leq \frac{1 + \sqrt{d \log \left(2en/d\right)}}{\delta \sqrt{n/2}} \leq \frac{1}{\delta} \sqrt{\frac{8d \log(2en/d)}{n}}$$

as soon as $\sqrt{d \log \left(2en/d\right)} \geq 1$. To ensure that this is at most ϵ, one may choose

$$n \geq \frac{8d \log(n)}{(\delta\epsilon)^2} + \frac{8d \log(2e/d)}{(\delta\epsilon)^2} \ .$$

By the following lemma, it is sufficient that

$$n \geq \frac{32d \log \left(\frac{4d}{(\delta\epsilon)^2}\right)}{(\delta\epsilon)^2} + \frac{16d \log \left(\frac{2e}{d}\right)}{(\delta\epsilon)^2} \ .$$

UNIVERSITÉ DE LYON

ENS DE LYON

Technical Lemma

Lemma

Let $a > 0$. Then

$$x \geq 2a\log(a) \quad \Longrightarrow \quad x \geq a\log(x)\ .$$

Proof: For $a \leq e$, true for every $x > 0$. Otherwise, for $a \geq \sqrt{e}$ we have $2a\log(a) \geq a$ and thus for every $t \geq 2a\log(a)$, as $f : t \mapsto t - a\log(t)$ is increasing on $[a, \infty)$, $f(t) \geq f(2a\log(a)) = a\log(a) - a\log(2\log(a)) \geq 0$, since for every $a > 0$ it holds that $a \geq 2\log(a)$.

Lemma

Let $a \geq 1, b > 0$. Then

$$x \geq 4a\log(2a) + 2b \quad \Longrightarrow \quad x \geq a\log(x) + b\ .$$

Proof: It suffices to check that $x \geq 2a\log(x)$ (given by the above lemma) and that $x \geq 2b$ (obvious since $4a\log(2a) \geq 0$).

UNIVERSITÉ DE LYON

ENS DE LYON