Machine Learning 2:
k-nearest neighbors, deviation bounds

Master 2 Computer Science

Aurélien Garivier

2018-2019

UNIVERSITÉ DE LYON

ENS DE LYON

Table of contents

The result of last lecture on the nearest-neighbor classifier

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

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

A3. η is c-Lipschitz continuous:

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

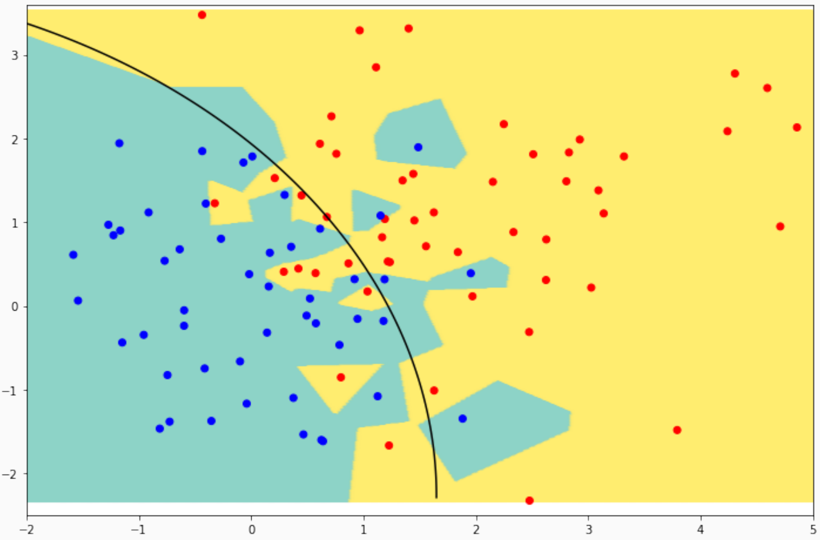
.

Theorem

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

$$L_D\left(\hat{h}_m^{NN}\right) \leq 2L_D^* + \frac{3c\sqrt{d}}{m^{1/(d+1)}}$$

Numerically



What does the analysis say?

- Where is the analysis loose? (sanity check: uniform $\mathcal{D}_X$)
- *finite sample* bound: explicit, non-asympototic
- The second term $\frac{3c\sqrt{d}}{m^{1/(d+1)}}$ is *distribution-free*
- Does not give the trajectorial decreasing rate of the risk
- Exponential bound d (cannot be avoided...)

  $\Longrightarrow$ *curse of dimensionality*
- Is is better than a simple grid approach?

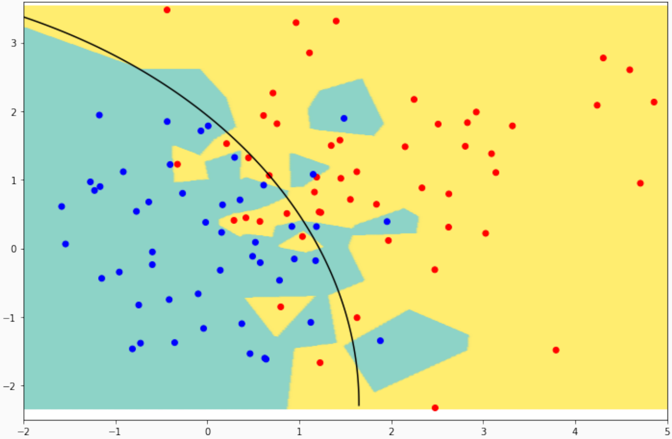
  $\Longrightarrow$ *adaptivity* to the dimension of manifold supporting data

- How to improve the classifier?

  $\Longrightarrow$ k-nearest neighbors

More neighbors are better?

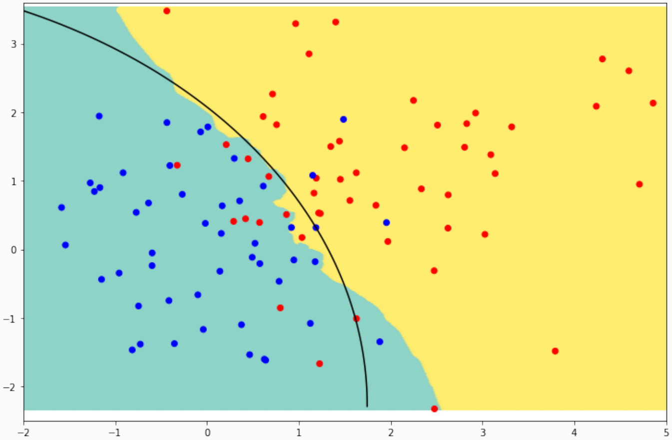
In general, yes in the sense that for m large enough, larger k is better.



But one can find counter-examples: $\forall k \geq 3, \forall m \geq k,\ L\big(\hat{h}_m^{kNN}\big) \geq \big(\hat{h}_m^{NN}\big)$.

More neighbors are better?

In general, yes in the sense that for m large enough, larger k is better.



But one can find counter-examples: $\forall k \geq 3, \forall m \geq k,\ L\left(\hat{h}_m^{kNN}\right) \geq \left(\hat{h}_m^{NN}\right)$.

Deviation Bound for Bernoulli Variables

Remember: Jensen's Inequality

Let $\mathcal{X}$ be a convex set and $\phi : x \to \mathbb{R}$ be a convex function.

Basic: For all $x, x' \in \mathcal{X}$, $\phi(tx + (1-t)x') \leq t\phi(x) + (1-t)\phi(x')$.

Probabilistic version: If $\phi : \mathcal{X} \to \mathbb{R}$ is convex and if X is a random variable with range in $\mathcal{X}$, then $\phi\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[\phi(X)\big]$.

Conditional version: If X and Y are random variables and the range of X in included in $\mathcal{X}$, if $\phi(X)$ is integrable then $\phi\big(\mathbb{E}[X|Y]\big) \leq \mathbb{E}\big[\phi(X)|Y\big]$.

Example: For a real-valued random variable X with finite expectation, $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$ and thus $\mathbb{V}ar[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$.

Make a picture. Think about equality case.

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:*

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(- n \operatorname{kl}(x, \mu)\big) \ , \tag{1}$$

where $\operatorname{kl}(p,q) = p \log \dfrac{p}{q} + (1-p) \log \dfrac{1-p}{1-q}$. *Same for left deviations.*

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

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

(iii) *Hoeffding's bound for Bernoulli variables:*

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(- 2n(x-\mu)^2 \big) \ . \tag{2}$$

(iv) *Inequalities (1) and (2) hold for arbitrary independent random variables with range* $[0,1]$ *and expectation* μ.

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

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

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

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

Bennett's and Bernstein's inequalities

Let $(X_i)_{1\le i\le n}$ 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 $\mathbb{E}[X_i^2] \le \sigma^2$ for all i and let $\phi(u) = (1+u)\log(1+u) - u$. Then, for all $x > 0$,

$$\mathbb{P}\big(\bar X \ge \bar\mu + x\big) \le \exp\left(-n\,\sigma^2\phi\left(\frac{x}{\sigma^2}\right)\right) \le \exp\left(-\frac{n\,x^2/2}{1+x/3}\right) .$$

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

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

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

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

$$\mathbb{P}\left(\bar X_k < \frac{\mu}{2}\right) \le \exp\left(-k\left(\frac{3}{2}\log\frac{3}{2} - \frac{1}{2}\right)\mu\right) \le \exp\left(-\frac{3k\mu}{28}\right) .$$

k-nearest neighbours

Definition

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

$$d\left(X_{(1)}(X), X\right) \leq d\left(X_{(2)}(X), X\right) \leq \cdots \leq d\left(X_{(m)}(X), X\right) .$$

Risk bound

Let $\mathcal{C}_\epsilon$ be an ϵ-covering of $\mathcal{X}$:

$$\forall x \in X, \exists x' \in C_\epsilon : d(x, x') \leq \epsilon \, .$$

Excess risk for k-nearest-neighbours

If η is c-Lipschitz continuous: $\forall x, x' \in \mathcal{X}, \left|\eta(x) - \eta(x')\right| \leq c\, d\big(x, x'\big)$, then for all $k \geq 2$ and all $m \geq 1$:

$$\begin{aligned} L\big(\hat{h}^{kNN}\big) - L\big(h^*\big) &\leq \frac{1}{\sqrt{k\,e}} + \frac{2k|\mathcal{C}_\epsilon|}{m} + 4c\epsilon \\ &\leq \frac{1}{\sqrt{k\,e}} + (2+4c)\left(\frac{\alpha k}{m}\right)^{\frac{1}{d+1}} \quad \begin{cases} \text{for } \epsilon = \left(\frac{\alpha k}{m}\right)^{\frac{1}{d+1}} & , \\ \text{if } |\mathcal{C}_\epsilon| \leq \alpha \epsilon^{-d} \end{cases} \\ &\leq (3+4c)\left(\frac{\alpha}{m}\right)^{\frac{1}{d+3}} \quad \text{for } k = \left(\frac{m}{\alpha}\right)^{\frac{2}{d+3}} \, . \end{aligned}$$

Sketch of the analysis

$$
\begin{aligned}
L(\hat{h}_m^{kNN}) - L(h^*) &= \mathbb{E}\left[\left|2\eta(X)-1\right| \mathbb{1}\{\hat{h}_m^{kNN} \neq h^*(x)\}\right] \\
&\leq \mathbb{P}\left(d(X, X_{(k)}) > 2\epsilon\right) + \mathbb{E}\left[\left|2\eta(X)-1\right| \mathbb{1}\{\hat{h}_m^{kNN} \neq h^*(x)\} \mathbb{1}\{d(X, X_{(k)}) \leq 2\epsilon\}\right]
\end{aligned}
$$

- $\mathbb{P}\left(d(X, X_{(k)}) > 2\epsilon\right) \leq \sum_{c \in \mathcal{C}_\epsilon} \mathbb{P}(X \in c, N_c < k) \leq \frac{2k\left|\mathcal{C}_\epsilon\right|}{m}$
- For x such that $\eta(x) \leq 1/2 - 2c\epsilon$,

$$
P\left(\hat{h}_m^{kNN}(x) = 1 \middle| X = x, d(X, X_{(k)}) \leq 2\epsilon\right) \leq \exp\left(-\frac{k}{2}\left(2\eta(x) + 4c\epsilon - 1\right)^2\right) .
$$

Same for $\eta(x) \geq 1/2 + 2c\epsilon$. And for $1/2 - 2c\epsilon \leq \eta(x) \leq 1/2 + 2c\epsilon$ the probability is upper-bounded by 1. In all cases, on $\{d(X, X_{(k)}) \leq 2\epsilon\}$:

$$
\left|2\eta(X)-1\right| P\left(\hat{h}_m^{kNN}(X) \neq h^*(X)\right) \leq 4c\epsilon + \sup_{u \geq 0} u \exp(-ku^2/2) = 4c\epsilon + \frac{1}{\sqrt{ke}} .
$$