Nearest neighbor process

François Portier

ENSAI, CREST

April, 2022

Regression background

Regression

- (X, Y) a random vector with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$
- If $\mathbb{E}[Y^2] < \infty$, there exists $h^* : \mathbb{R}^d \to \mathbb{R}$ such that for all $h : \mathbb{R}^d \to \mathbb{R}$

$$\mathbb{E}[(Y - h^*(X))^2] \leq \mathbb{E}[(Y - h(X))^2]$$

- h^* is the **conditional expectation** of Y given X:

  the "best prediction" of Y we can get from X
- **GOAL:**

  Estimating h^*

  (which is unknown as it depends on the underlying probability measure)

Estimation from data

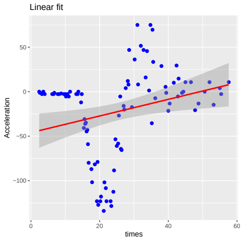
- $(X, Y), (X_i, Y_i)_{i \in \{1,\dots,n\}}$ **iid** random vectors
- The estimate of h^* (that depends on the data) is

$$\hat{h} : \mathbb{R}^d \to \mathbb{R}$$

The big picture (Györfi et al., 2006)

Global modeling methods

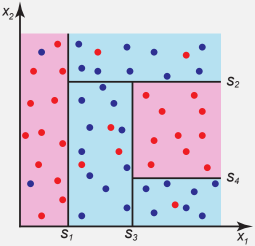
- Polynomial regression
- Spline approximation
- RKHS methods



NB: Often conducted with penalization

Local averaging methods

- Nadaraya-Watson (NW)
- nearest neighbor (k-NN)
- extreme value estimates (when conditioning upon large values)
- partitioning methods



NW and k-NN

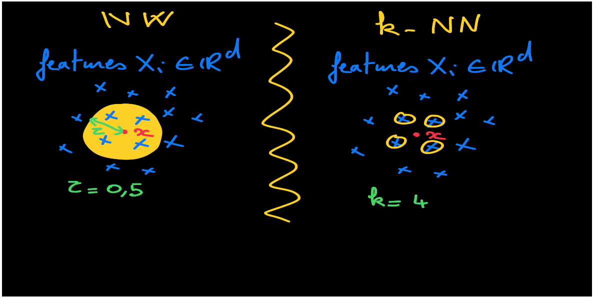
$x \in \mathbb{R}^d$, $\|\cdot\|$ is a norm on $\mathbb{R}^d$, $B(x,\tau)$ is the closed ball,

NW (1964)

- Let $\tau > 0$
- $\hat{h}^{(NW)}(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{B(x,\tau)}(X_i)}{\sum_{i=1}^n \mathbb{1}_{B(x,\tau)}(X_i)}$

k-NN (1951)

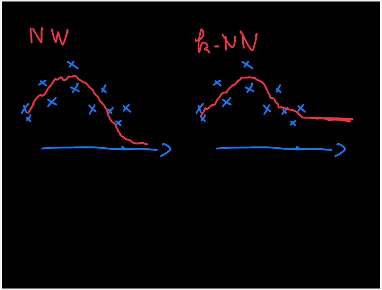
- Let $N_k(x)$ denote the k-NN of x among $\{X_1, \dots, X_n\}$
- $\hat{h}^{(NN)}(x) = \frac{1}{k} \sum_{i \in N_k(x)} Y_i$



Both part of Stone (1977)'s theorem framework: $\sum_{i=1}^n Y_i w_{n,i}(x)$ where $\sum_{i=1}^n w_{n,i}(x) = 1$

Stylized facts about k-NN and NW

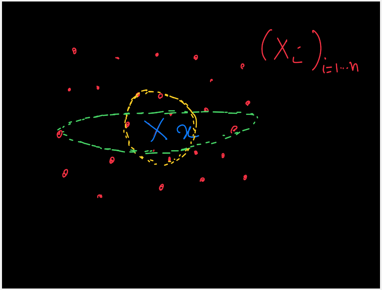
- intuitive yet powerful methods
  $\Rightarrow$ both match the optimal convergence rate (Einmahl and Mason, 2000; Jiang, 2019)
- kNN is bandwidth adaptive
  $\Rightarrow$ free from boundary problems; adapts to covariate space (Kpotufe, 2011)
- can be enhanced with metric learning (Weinberger et al., 2006); parallelization (Qiao et al., 2019); bagged version (Biau et al., 2010)
- can be used in residual variance (Devroye et al., 2018) and sparse gradient (Ausset et al., 2021) estimation



Different behavior at the boundary

Stylized facts about k-NN and NW

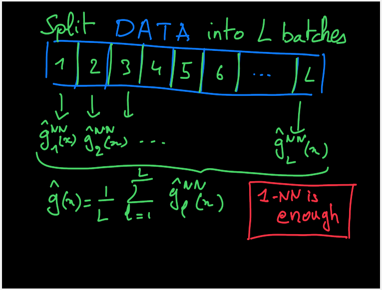
- intuitive yet powerful methods
  $\Rightarrow$ both match the optimal convergence rate (Einmahl and Mason, 2000; Jiang, 2019)
- kNN is bandwidth adaptive
  $\Rightarrow$ free from boundary problems; adapts to covariate space (Kpotufe, 2011)
- can be enhanced with metric learning (Weinberger et al., 2006); parallelization (Qiao et al., 2019); bagged version (Biau et al., 2010)
- can be used in residual variance (Devroye et al., 2018) and sparse gradient (Ausset et al., 2021) estimation



Metric learning with kNN

Stylized facts about k-NN and NW

- intuitive yet powerful methods
  $\Rightarrow$ both match the optimal convergence rate (Einmahl and Mason, 2000; Jiang, 2019)
- kNN is bandwidth adaptive
  $\Rightarrow$ free from boundary problems; adapts to covariate space (Kpotufe, 2011)
- can be enhanced with metric learning (Weinberger et al., 2006); parallelization (Qiao et al., 2019); bagged version (Biau et al., 2010)
- can be used in residual variance (Devroye et al., 2018) and sparse gradient (Ausset et al., 2021) estimation



Recursive kNN

Local averaging methods

Tools to analyze k-NN

- Bias-variance decomposition
- The k-NN radius
- The k-NN variance

k-NN process

- Motivation
- Main result

Bias-Variance decomposition

Definition

The k-NN radius is $\hat{\tau}_{k,x} = \inf\{\tau \geq 0 \,:\, \sum_{i=1}^{n} \mathbb{1}_{B(x,\tau)}(X_i) \geq k\}$

We consider

$$\hat{h}^{(NN)}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{B(x,\hat{\tau}_{k,x})}(X_i)}{\sum_{i=1}^{n} \mathbb{1}_{B(x,\hat{\tau}_{k,x})}(X_i)}$$

(always defined even when ties occurs)

Decomposition

$$\hat{h}^{(NN)}(x) - h^*(x) = \underbrace{\sum_{i=1}^{n} (Y_i - h^*(X_i)) w_{n,i}(x)}_{\text{the variance}} + \underbrace{\sum_{i=1}^{n} (h^*(X_i) - h^*(x)) w_{n,i}(x)}_{\text{the bias}}$$

If h^* is L-Lipschitz,

$$|\text{the bias}| \leq L\hat{\tau}_{k,x}$$

Useful results 1 (for k-NN radius)

Lemma (Chernoff bound)

Let $(Z_i)_{i\geq 1}$ *be a sequence of i.i.d. random variables valued in* $\{0,1\}$*. Set* $\mu = n\mathbb{E}[Z_1]$ *and* $S = \sum_{i=1}^n Z_i$*. For any* $\delta \in (0,1)$ *and all* $n \geq 1$*, we have with probability* $1-\delta$*:*

$$S \geq \left(1 - \sqrt{\frac{2\log(1/\delta)}{\mu}}\right)\mu.$$

Property 1 (k-NN radius)

Let $x \in \mathbb{R}^d$ be a continuity point of f_X such that $f_X(x) > 0$. If $k \to \infty$ and $k/n \to 0$,

$$\hat{\tau}_{k,x} = O_P((k/n)^{1/d})$$

Useful results 2 (for the variance)

sub-Gaussian random variable

A centered random variable ϵ is sub-Gaussian whenever

$$\mathbb{E}[\exp(\lambda\epsilon)] \leq \exp(\lambda^2 v/2) \qquad \forall\lambda \in \mathbb{R}$$

where $v > 0$ is called the sub-Gaussian factor

Lemma (subGaussian concentration inequality)

(i) If ϵ is subGaussian, $\mathbb{P}(\epsilon > t) \leq \exp(-t^2/(2v))$

(ii) If (ϵ_i) are iid subGaussians with factor v, then $\sum_i w_i\epsilon_i$ is subGaussian with factor $v\sum_i w_i^2$.

Property 2 (k-NN variance)

Suppose that $(\epsilon, X), (\epsilon_i, X_i)_{i=1,\dots,n}$ is iid such that ϵ is subGaussian with variance σ^2 and $\epsilon \perp X$. Then we have that with probability $1-\delta$:

$$\left| \frac{\sum_{i=1}^{n} \epsilon_i \mathbb{1}_{B(x,\hat{\tau}_{k,x})}(X_i)}{\sum_{i=1}^{n} \mathbb{1}_{B(x,\hat{\tau}_{k,x})}(X_i)} \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{k}}$$

Suppose the following is fulfilled

- $x \in \mathbb{R}^d$ is a continuity point of f_X such that $f_X(x) > 0$.
- The function g is Lipschitz
- For each i, $\epsilon_i = Y_i - h^*(X_i)$ is subGaussian with variance σ^2 and is independent from X_i

Proposition (k-NN rate)

If $k \to \infty$ and $k/n \to 0$

$$|\hat{h}^{(NN)}(x) - h^*(x)| = O_{\mathbb{P}}\left(\sqrt{\frac{1}{k}} + (k/n)^{1/d}\right)$$

The optimal bound $n^{-1/(2+d)}$ is reached whenever $k = n^{2/(2+d)}$ (similar to NW)

Proposition (asymptotic variance) (Mack, 1981)

- NW $\quad \dfrac{\sigma^2(x)\int K^2 \mathrm{d}\lambda}{(n\tau^d)f(x)}$
- k-NN $\quad \dfrac{2\sigma^2(x)\int K^2 \mathrm{d}\lambda}{k}$

Textbook: Biau and Devroye (2015)

Empirical process theory

- Let $(Z_i)_{i\geq 1}$ be a sequence of iid random variables with distribution μ on $\mathcal{Z}$
- Let $\mathcal{G}$ be a collection of functions $g : \mathcal{Z} \to \mathbb{R}$
- Let $\ell^\infty(\mathcal{G})$ be the space of bounded functions defined on $\mathcal{G}$

Definition

The empirical process is an element of $\ell^\infty(\mathcal{G})$ defined as

$$\mathbb{G}_n(g) = \sqrt{n}(\mu_n(g) - \mu(g)), \qquad (g \in \mathcal{G})$$

where $\mu_n(g) = n^{-1} \sum_{i=1}^n g(X_i)$ and $\mu(g) = \int g d\mu$.

Leading question: what is the behavior of the process $\{\mathbb{G}_n(g)\}_{g\in\mathcal{G}}$?

- Answer1: When $\mathcal{G}$ is **not too large** $\mathbb{E}[\sup_{g\in\mathcal{G}} |\mathbb{G}_n(g)|] = O(\sigma_{\mathcal{G}})$
- Usefulness1: Provide theoretical guarantee on (nonparametric) estimate such as Quantile, Copulas, Kaplan-Meier, NW
- Answer2: When $\mathcal{G}$ is **not too large** $\{\mathbb{G}_n(g)\}_{g\in\mathcal{G}}$ converges weakly in the space $\ell^\infty(\mathcal{G})$
- Usefulness2: Provide distribution of meaningful statistical object (see next)

van der Vaart and Wellner (1996); Nolan and Pollard (1987); Massart (1990); Giné and Guillou (2002)

Illustrative example: independence testing

Framework

Testing if two random variables $Z^{(1)}$ and $Z^{(2)}$ are independent, that is

$$H_0 : Z^{(1)} \perp Z^{(2)} \quad \Leftrightarrow \quad \|F_{1,2} - F_1 F_2\|_\infty = 0$$

where $F_{1,2}$ the joint cdf and F_J each marginal's cdf.

Empirical process results

Consider

$$\mathcal{G} = \left\{ (Z^{(1)}, Z^{(2)}) \mapsto \mathbb{1}_{Z^{(1)} \leq z^{(1)}} \mathbb{1}_{Z^{(2)} \leq z^{(2)}} \ : \ z = (z^{(1)}, z^{(2)}) \in \mathbb{R}^2 \right\}$$

The class $\mathcal{G}$ being **sufficiently small**, we have

$$\left\{ \sqrt{n} (\hat{F}_{1,2}(z^{(1)}, z^{(2)}) - F_{1,2}(z^{(1)}, z^{(2)})) \right\}_{(z^{(1)}, z^{(2)}) \in \mathbb{R}^2}$$

(where the $\hat{F}_{1,2}$ is the estimated cdf) converges weakly to a Gaussian process $\mathbb{W}$

Consequence: Under H_0, $\sqrt{n} \|\hat{F}_{1,2} - \hat{F}_1 \hat{F}_2\|_\infty \rightsquigarrow \|\mathbb{W}\|_\infty$

Classically, independence testing is based on **copula** (Fermanian et al., 2004; Segers, 2012)

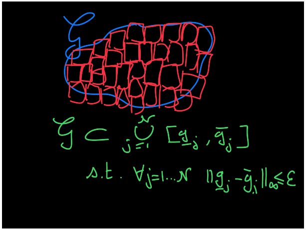
Weak convergence *via* bracketing entropy (what is it to be small?)

- Let $\underline{f}$ and $\overline{f}$ be two functions in $L_2(\mu)$ bracket

$$[\underline{f}, \overline{f}] = \{g \in L_2(\mu) \,:\, \underline{f} \le g \le \overline{f}\}$$

- A bracket $[\underline{f}, \overline{f}]$ such that $\|\underline{f} - \overline{f}\|_{L_2(\mu)} \le \epsilon$ is called an ϵ-bracket.

$\mathcal{N}_{[\,]}(\mathcal{G}, L_2(\mu), \epsilon)$ is the smallest $\mathcal{N}$ such that:



there exists an $(L_2(\mu), \epsilon)$-bracketing of cardinal $\mathcal{N}$

Bracketing condition

for any positive sequence $(\delta_n)_{n\ge1}$ going to 0, it holds that

$$\int_0^{\delta_n} \sqrt{\log\left(\mathcal{N}_{[\,]}\left(\mathcal{G}, L_2(P), \epsilon\|G\|_{L_2(P)}\right)\right)}\,\mathrm{d}\epsilon \to 0 \quad \text{as } n \to \infty,$$

where G is an envelope for $\mathcal{G}$, i.e., $|g(z)| \le G(z)$

Weak convergence *via* bracketing entropy

Theorem (van der Vaart and Wellner, 1996)

Under the bracketing condition, it holds that $\{\mathbb{G}_n(g)\}_{g\in\mathcal{G}}$ converges weakly in $\ell^\infty(\mathcal{G})$ to a Gaussian process with covariance function $\mu(g_1 g_2) - \mu(g_1)\mu(g_2)$.

Research question:

- can we obtain similar results for local averaging method?
- useful whenever we are interested in specific parts of the feature space
  - testing conditional independence
  - Conditional copula (Veraverbeke et al., 2011)
  - conditional quantile estimation (Härdle and Tsybakov, 1988)
  - *M*-smoothers (Härdle et al., 1988)

Definition of the k-NN process

k-NN

- Let $x \in \mathbb{R}^d$, $\|\cdot\|$ is a norm on $\mathbb{R}^d$; $B(x, \tau)$ is the closed ball
- μ_x is the conditional measure of Y given $X = x$, i.e.,

$$\mu_x(A) = \mathbb{P}(Y \in A | X = x)$$

- The **k-NN measure** is

$$\hat{\mu}_x^{(NN)}(A) = \frac{\sum_{i=1}^n \mathbb{1}_A(Y_i) \mathbb{1}_{B(x, \hat{\tau}_{k,x})}(X_i)}{\sum_{i=1}^n \mathbb{1}_{B(x, \hat{\tau}_{k,x})}(X_i)}$$

- The **k-NN process** defined on $\mathcal{G}$ is

$$\{\sqrt{k}(\hat{\mu}_x^{(NN)}(g) - \mu_x(g))\}_{g \in \mathcal{G}}$$

Local bracketing

For any $x \in \mathbb{R}^d$, $u > 0$, define the probability measure

$$\mu_{x,u}(A) = \frac{E(\mu_X(A)\mathbb{1}_{B(x,u^{1/d})}(X))}{E(\mathbb{1}_{B(x,u^{1/d})}(X))}$$

Local bracketing entropy

There is $\delta > 0$ such that for any positive sequence $(\delta_n)_{n\geq 1}$ going to 0, it holds that

$$\sup_{|u|\leq\delta} \int_0^{\delta_n} \sqrt{\log\left(\mathcal{N}_{[\,]}\left(\mathcal{G}, L_2(\mu_{x,u}), \epsilon\|G\|_{L_2(\mu_{x,u})}\right)\right)}\,\mathrm{d}\epsilon \to 0 \quad \text{as } n \to \infty \tag{1}$$

(same as before except that μ became $\mu_{x,u}$)

Main result

Suppose the following is fulfilled

- $f_X(x) > 0$ and that f_X is continuous at x
- $\mu_x(g)$ is Lipschitz at x (uniformly over g)
- The covariance $x \mapsto \mu_x(g_1 g_2) - \mu_x(g_1)\mu_x(g_2)$ is continuous at x

Theorem

Under the local bracketing condition, if $k \to \infty$ and $k^{(d+2)/2}/n \to 0$ we have

$$\{\sqrt{k}(\hat{\mu}_x^{(NN)}(g) - \mu_x(g))\}_{g \in \mathcal{G}}$$

converges weakly to a Gaussian process with covariance function $\mu_x(g_1 g_2) - \mu_x(g_1)\mu_x(g_2)$.

References I

The paper is available here: https://arxiv.org/abs/2110.15083

Ausset, G., S. Clémen, et al. (2021). Nearest neighbour based estimates of gradients: Sharp nonasymptotic bounds and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 532–540. PMLR.

Biau, G., F. Cérou, and A. Guyader (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory 56*(4), 2034–2040.

Biau, G. and L. Devroye (2015). *Lectures on the nearest neighbor method*, Volume 246. Springer.

Bousquet, O., S. Boucheron, and G. Lugosi (2003). Introduction to statistical learning theory. In *Summer school on machine learning*, pp. 169–207. Springer.

Devroye, L., L. Györfi, G. Lugosi, and H. Walk (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics 12*(1), 1752–1778.

Einmahl, U. and D. M. Mason (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability 13*(1), 1–37.

Fermanian, J.-D., D. Radulovic, M. Wegkamp, et al. (2004). Weak convergence of empirical copula processes. *Bernoulli 10*(5), 847–860.

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, Volume 38, pp. 907–921. Elsevier.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Härdle, W., P. Janssen, and R. Serfling (1988). Strong uniform consistency rates for estimators of conditional functionals. *The Annals of Statistics 16*(4), 1428–1449.

References II

Härdle, W. and A. B. Tsybakov (1988). Robust nonparametric regression with simultaneous scale curve estimation. *The annals of statistics*, 120–135.

Jiang, H. (2019). Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 3999–4006.

Kpotufe, S. (2011). k-nn regression adapts to local intrinsic dimension. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 729–737.

Lhaut, S., A. Sabourin, and J. Segers (2021). Uniform concentration bounds for frequencies of rare events. *arXiv preprint arXiv:2110.05826*.

Mack, Y.-P. (1981). Local properties of k-nn regression estimates. *SIAM Journal on Algebraic Discrete Methods 2*(3), 311–323.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, 1269–1283.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *The Annals of Statistics*, 780–799.

Plassier, V., F. Portier, and J. Segers (2020). Risk bounds when learning infinitely many response functions by ordinary linear regression. *arXiv preprint arXiv:2006.09223*.

Qiao, X., J. Duan, and G. Cheng (2019). Rates of convergence for large-scale nearest neighbor classification. *Advances in Neural Information Processing Systems 32*, 10769–10780.

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli 18*(3), 764–782.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*(4), 595–645. With discussion and a reply by the author.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics. New York: Springer-Verlag.

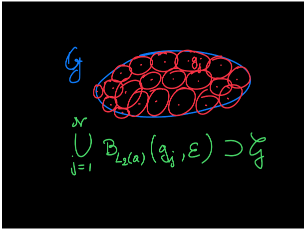
References III

Veraverbeke, N., M. Omelka, and I. Gijbels (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics 38*(4), 766–780.

Weinberger, K. Q., J. Blitzer, and L. K. Saul (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pp. 1473–1480.

Uniform bounds *via* Vapnik-Chervonenkis approach

$\mathcal{N}(\mathcal{G}, L_2(Q), \epsilon)$ is the smallest $\mathcal{N}$ such that:



there exists an $(L_2(Q), \epsilon)$-cover of cardinal $\mathcal{N}$

Definition (VC-class)

A class $\mathcal{G}$ of functions in $[-1, 1]$ is called a VC with parameters $(v > 0, A > 1)$ if for any $0 < \epsilon < 1$ and any probability measure Q, we have

$$\mathcal{N}\left(\mathcal{G}, L_2(Q), \epsilon\right) \leq (A/\epsilon)^v.$$

Successes of VC classes

- Same rate as standard empirical process results (Massart, 1990)
- Helpful in statistical learning (Bousquet et al., 2003)
- Nadaraya-Watson estimate (Nolan and Pollard, 1987; Giné and Guillou, 2002)

Uniform bound

Assumptions

- $f_X = \mathbb{1}_{[0,1]^d}$
- $K(d \vee v) \log(2An/\delta) \leq k$ where $K > 0$ universal
- $\forall (x, x') \in S_x \times S_x, g \in \mathcal{G}, \qquad |\mu_x(g) - \mu_{x'}(g)| \leq L\|x - x'\|$

Result

With probability at least $1 - \delta$:

$$\sup_{x \in S_x} |\hat{\mu}_x^{(NN)}(g) - \mu_x(g)| \leq K \left\{ \sqrt{\frac{(d \vee v)}{k} \log(2An/\delta)} + L \left(\frac{k}{nV_d} \right)^{1/d} \right\}$$

with $V_d = \lambda(B(0,1))$

Auxiliary results: Plassier et al. (2020) for the variance term, Lhaut et al. (2021) for the k-NN radius