From Vapnik to empirical process theory

A non-asymptotic view

François Portier

ENSAI

ENSAI École nationale de la statistique et de l'analyse de l'information

CREST
CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

Probabilistic background

Reminder on probability

The data will be **random variables**. We rely the probabilistic framework.

1. A random variable $Z = (Y, X)$ valued in $\mathcal{Z} = \{0, 1\} \times \mathbb{R}^d$ is characterized by its distribution P defined as, for any $A \subset \mathcal{Z}$

$$P(A) = \mathbb{P}(Z \in A)$$

2. The expectation of Z is

$$\mathbb{E}Z = \int z \, dP(z)$$

3. The variance of Z is

$$\mathrm{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}Z)^2]$$

4. For a function $g : \mathcal{Z} \to \mathbb{R}$, define

$$P(g) = \mathbb{E}[g(Z)]$$

Some context

The general goal

Suppose that the data $Z_1, \ldots, Z_n$ is an independent and identically distributed collection of random variable. Define the empirical measure

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{Z_i}$$

We have the following claim:

$$P_n \text{ is a good approximation of } P$$

supported by many well-known results (central limit theorem, Hoeffding bound...)

The motivation

To derive statistical guarantees on **classification algorithms** for $Z = (Y, X)$, i.e.,

"**learning from the data** how to predict **label** $Y \in \{0, 1\}$ given **covariates** X"

Some context

The main tool: **deviation inequalities**

For any probability measure P satisfying some conditions, any function g satisfying other conditions, we have, for any $n \geq 1$, any $t > 0$,

$$\mathbb{P}(|P_n(g) - P(g)| > t) \leq \delta = Bound(t, n, P, g)$$

or for any $n \geq 1$, any $\delta > 0$, with probability (at least) $1 - \delta$,

$$|P_n(g) - P(g)| \leq t = Bound(\delta, n, P, g)$$

The main tool: **uniform** deviation inequalities

A particular attention is given to uniform deviation with respect to some given class $\mathcal{G}$. Specifically we consider deviation inequalities with

$$\sup_{g \in \mathcal{G}} |P_n(g) - P(g)|.$$

The obtained bound will depend on the complexity of $\mathcal{G}$.

Agenda

Basic deviation inequalities (1 hour)

- A brief presentation of the Chernoff method
- The Chernoff multiplicative deviation bound
- Some subGaussian bounds (expectation and deviation - analysis of the variance)
- Exercise 3, 4, 5 from the lecture notes

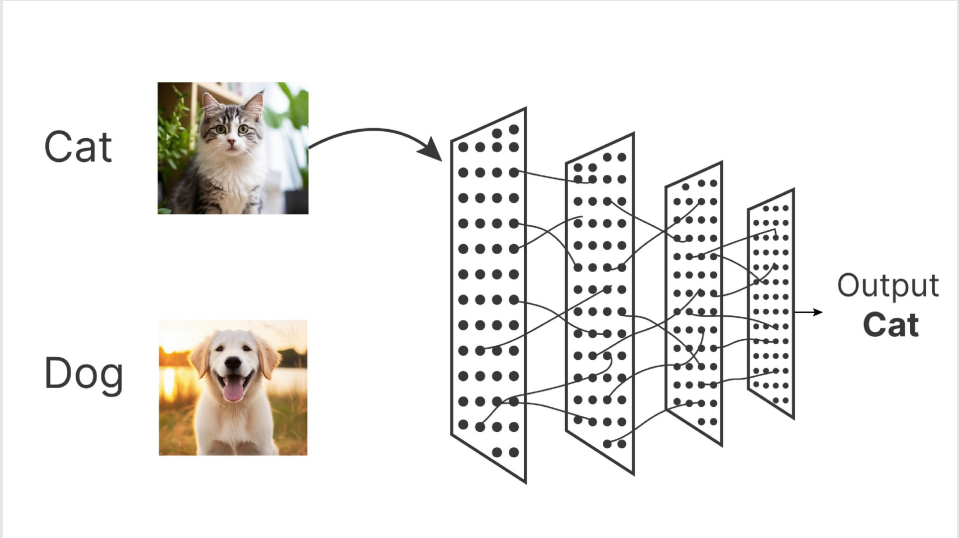
Vapnik deviation inequalities (2 hours)

- Symmetrization (expectation, convex transform and deviation)
- The first Vapnik inequality
- The second Vapnik inequality (with relative deviation - without proof - see lecture notes)
- Shattering coefficient and Vapnik dimension - definition and examples

What is Classification?

The goal is to **predict** a discrete **label** $Y \in \{0, 1\}$ for a given input $X \in \mathbb{R}^d$

- Email spam detection (spam or not spam)
- Image recognition (cat or dog)
- Disease diagnosis (positive or negative)



A **supervised learning** task: learning by examples

Classification framework

Objective

Let g be a classifier, i.e., $g : \mathbb{R}^d \to \{0, 1\}$. The missclassification risk of g is defined as

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

The above is the **standard metric** for classification. Our goal is to approximate

$$g^* = \arg\min_{g \in \mathcal{G}_{all}} R(g)$$

where $\mathcal{G}_{all}$ is the set of all classifiers.

Why is it difficult?

- **The risk function R is unknown** and should be estimated from the data
- Optimizing over $\mathcal{G}_{all}$ is not possible

The Bayes classifier

Define the conditional probability

$$\eta^*(x) = \mathbb{P}(Y = 1|X = x) \qquad (x \in \mathbb{R}^d)$$

Define the Bayes classifier as

$$g^*(x) = \mathbb{1}_{\eta^*(x)>1/2} \qquad (x \in \mathbb{R}^d)$$

Theorem 1 (The Bayes risk is the smallest)

We have

$$R(g) \geq R(g^*) \quad \text{for all classifier } g$$

This provides support for the following **plug-in classification rule**:

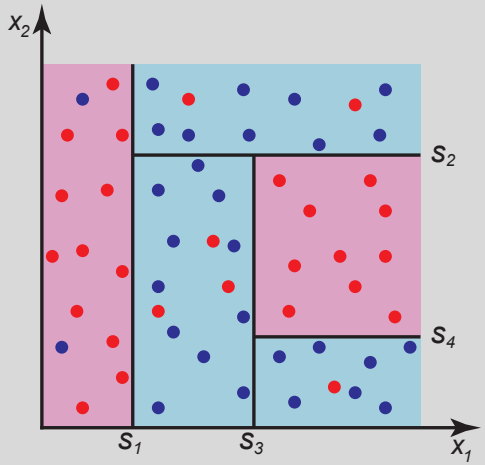
$$\hat{g} = \mathbb{1}_{\hat{\eta}_n>1/2},$$

in which: one first estimates **conditional probability** $\hat{\eta}_n$ and then do **majority vote**. This is done in classification trees, nearest neighbors, Nadaraya-Watson or Logistic regression. This is not the case for margin-based empirical risk minimizers (e.g., SVM).

The big picture (Györfi et al., 2006)

Local averaging methods

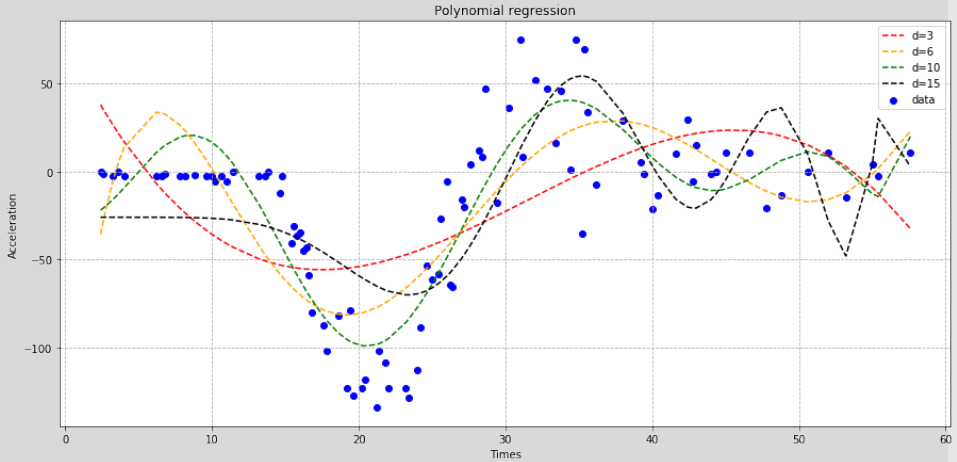
- Nadaraya-Watson (NW)
- nearest neighbor (k-NN)
- partitioning methods



Risk minimization (global - not local - modeling)

- Logistic regression
- RKHS methods (SVM, splines)
- Neural networks

(Often conducted with penalization)



In blue might be considered as plug-in approaches

NW and k-NN

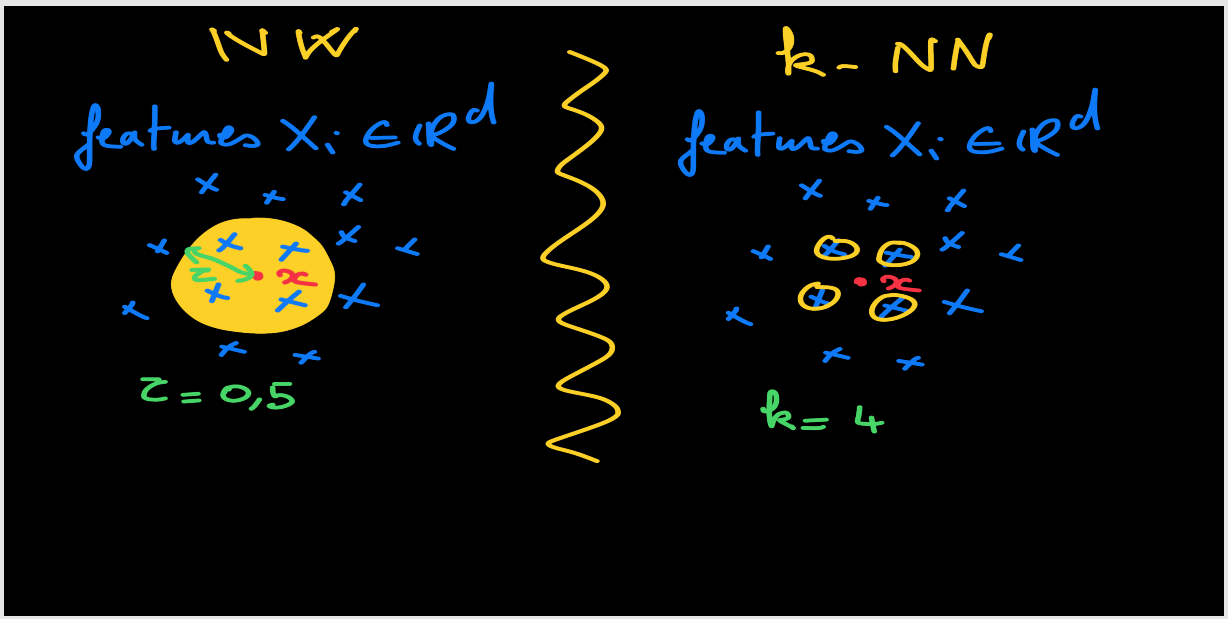
$x \in \mathbb{R}^d$, $\|\cdot\|$ is a norm on $\mathbb{R}^d$, $B(x,\tau)$ is the closed ball,

NW (1964)

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

k-NN (1951)

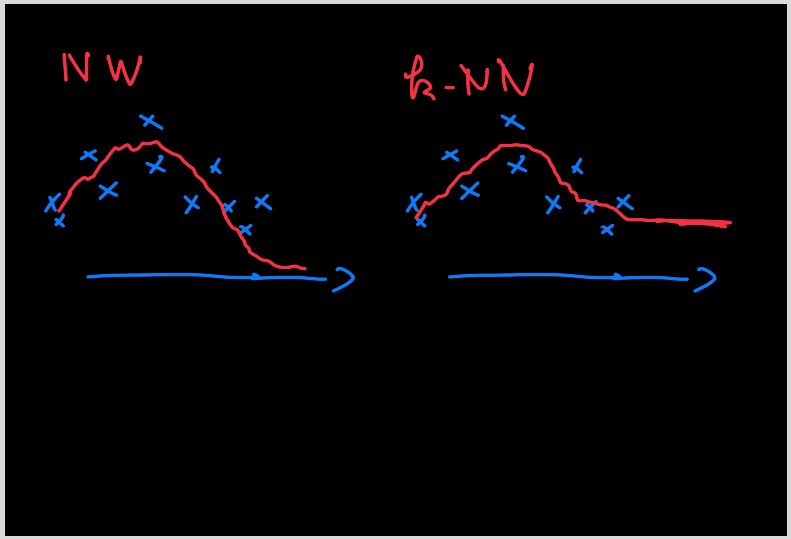
- Let $N_k(x)$ denote the indexes of k-NN to x among $\{1,\dots,n\}$
- $\hat{\eta}_n(x) = \frac{1}{k}\sum_{i\in N_k(x)} Y_i$



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

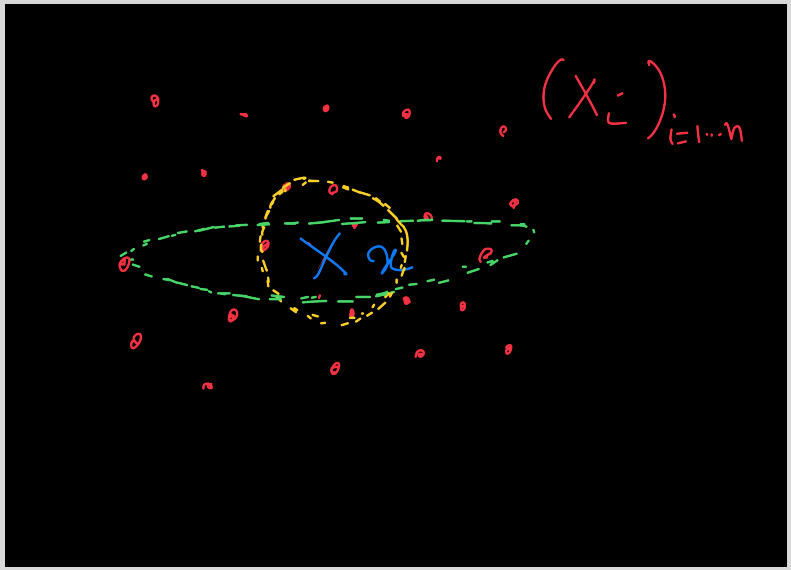
- intuitive yet powerful methods as both match the optimal convergence rate (Biau and Devroye, 2015; Györfi et al., 2006)
- kNN is bandwidth adaptive in particular, free from boundary problems; adapts to covariate space (Kpotufe, 2011)
- can be enhanced with metric learning (Weinberger et al., 2006); 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

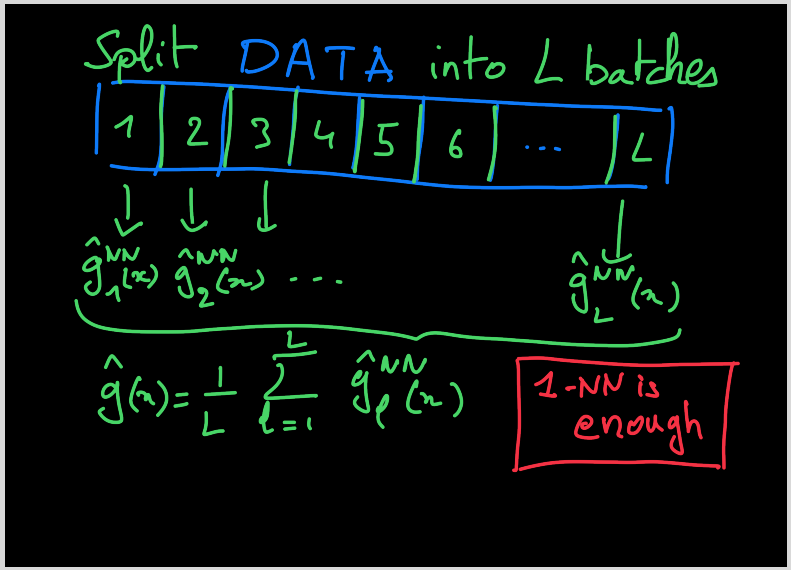
- intuitive yet powerful methods as both match the optimal convergence rate (Biau and Devroye, 2015; Györfi et al., 2006)
- kNN is bandwidth adaptive in particular, free from boundary problems; adapts to covariate space (Kpotufe, 2011)
- can be enhanced with metric learning (Weinberger et al., 2006); 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 as both match the optimal convergence rate (Biau and Devroye, 2015; Györfi et al., 2006)
- kNN is bandwidth adaptive in particular, 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

Classification and regression

Theorem 2 (classification is easier than regression)

We have

$$0 \leq R(g) - R(g^*) = \mathbb{E}[|2\eta^*(X) - 1| \mathbb{1}_{g(X) \neq g^*(X)}]$$

Supposing that $g = \mathbb{1}_{\eta > 1/2}$

$$R(g) - R(g^*) \leq 2\mathbb{E}[|\eta(X) - \eta^*(X)|]$$

The above suggests that classification may enjoy better rate than regression

Theorem 3 (classification is easier than regression)

When $R(g^*) = 0$ and $g = \mathbb{1}_{\eta > 1/2}$, we have

$$R(g) \leq 4\mathbb{E}[|\eta(X) - \eta^*(X)|^2]$$

Brief review on L_2-rates

Györfi, Kohler, Krzyżak & Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer.
Chapter 6 gives a thorough treatment of k-NN regression, showing that under a $s \leq 1$-Hölder condition on η and with $k \asymp n^{2s/(2s+d)}$, one gets the minimax rate

$$\mathbb{E}\Big[|\hat{\eta}_n(X) - \eta^*(X)|^2\Big] = O\Big(n^{-2s/(2s+d)}\Big)\,.$$

Biau & Devroye (2015). *Lectures on the Nearest Neighbor Method.* Springer.
Chapter 14 shows that even for $s = 2$, one gets the minimax rate

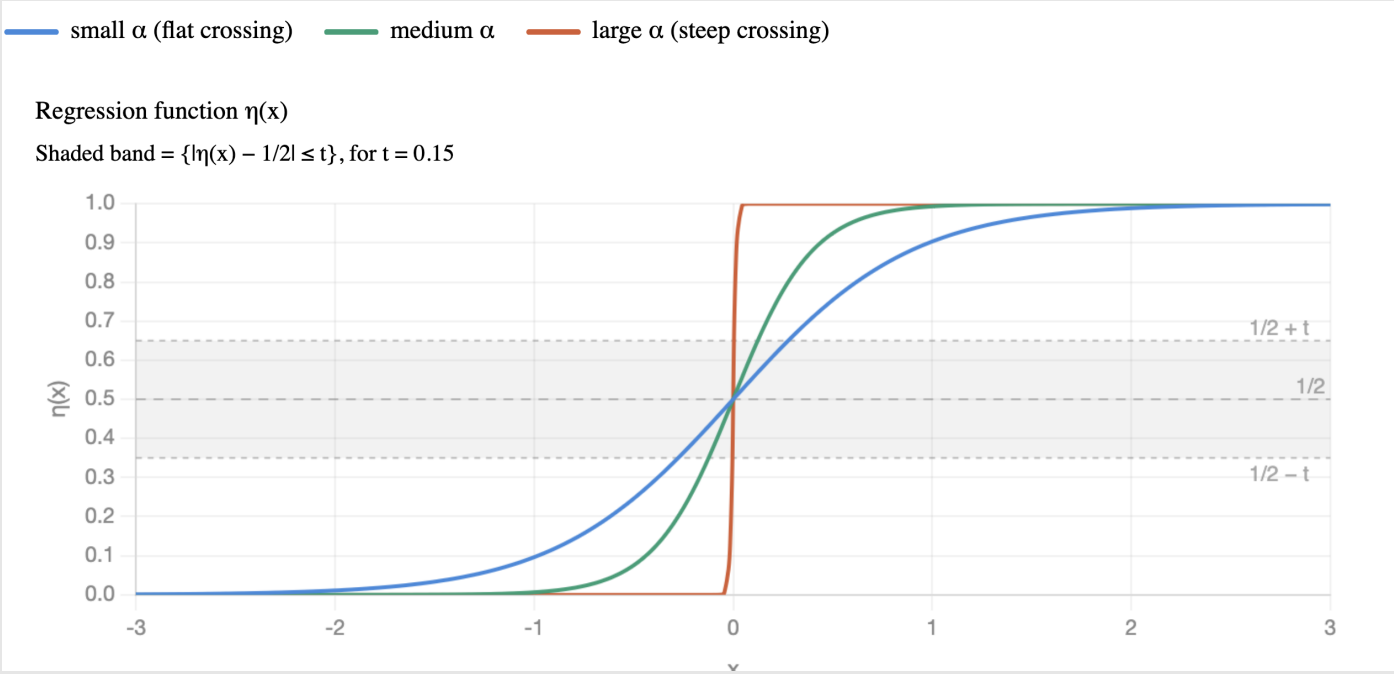
$$\mathbb{E}\Big[|\hat{\eta}_n(X) - \eta^*(X)|^2\Big] = O\Big(n^{-2s/(2s+d)}\Big)\,.$$

Margin assumption

The margin assumption is given bellow.

(MA) There exist constants $C_0 > 0$ and $\alpha \geq 0$ such that

$$P_X\left(0 < \left|\eta^*(X) - \tfrac{1}{2}\right| \leq t\right) \leq C_0\, t^\alpha, \quad \forall\, t > 0.$$



Fast learning rates for plug-in classifiers

Let $\hat{\eta}_n$ be an estimator of the regression function η^*. Suppose the following pointwise concentration bound: For some constants $C_1 > 0$, $C_2 > 0$, for some positive sequence $(a_n)_{n\geq 1}$, for all $n \geq 1$, all $\delta > 0$, and for almost all x with respect to P_X, we have

$$\mathbb{P}(|\hat{\eta}_n(x) - \eta^*(x)| > t) \leq C_1 \exp\left(-C_2\, a_n\, t^2\right).$$

Lemma (Audibert and Tsybakov (2007), Lemma 3.1)

Consider the plug-in classifier $\hat{g}_n = \mathbf{1}_{\{\hat{\eta}_n \geq 1/2\}}$. Then

$$\mathbb{E}R(\hat{g}_n) - R(g^*) \leq C\, a_n^{-\frac{1+\alpha}{2}}, \quad \forall\, n \geq 1,$$

for some constant $C > 0$ depending only on α, C_1, and C_2.

The above shows that classification may enjoy better rate, in $a_n^{-(1+\alpha)/2}$, than regression in $a_n^{-1/2}$. The analysis relies on a point-wise deviation inequality that is different in nature to results given in previous references.

A point-wise deviation inequality for the k-NN method

The leading assumption

(iid) Let $(Z, Z_1, \ldots, Z_n)$ is a collection of independent and identically distributed random elements with common distribution P on $\mathcal{Z} = \{0,1\} \times \mathbb{R}^d$.

(Xx) The distribution of X has a density f such that $f(y) \geq b > 0$ for all $y \in B(x, \tau_0)$

(E) $Y - \eta^*(X)$ is independent from X and is subGaussian with factor σ^2.

Let $N_k(x)$ denote the index set of the k-NN to point $x \in \mathbb{R}^d$ among $X_1, \ldots, X_n$. Define

$$\hat{\eta}_n(x) := k^{-1} \sum_{i \in N_k(x)} Y_i$$

and suppose that $\eta^*(x) := P(Y|X = x)$ is L-Lipschitz on $B(x, \tau_0)$.

A point-wise deviation inequality for the k-NN method

Theorem

Under the previous assumptions, namely (iid), (Xx) and (E), if $1 \leq k \leq n$ and $0 < \delta \leq 1/2$ are such that

$$8 \log(1/\delta) \leq 2k \leq nbV_d \tau_0^d,$$

we have with probability $1 - 2\delta$,

$$|\hat{\eta}_n(x) - \eta^*(x)| \leq \sqrt{\frac{2\sigma^2}{k} \log(2/\delta)} + L \left(\frac{2k}{nbV_d} \right)^{1/d}$$

Proof done on the blackboard (30 minutes)

Corollary

Let $a_n = n^{2/(d+2)}$. There is $A > 0$ such that, for all $n \geq 1$ and $t > 0$, it holds

$$\mathbb{P}(|\hat{\eta}_n(x) - \eta^*(x)| > t) \leq 4 \exp\left(-a_n t^2/(2A)^2\right)$$

Proof on the blackboard.

Fast learning rates for k-NN

New assumption

(Xx) The distribution of X has a density f such that $f(y) \geq b > 0$ for all $y \in B(x, \tau_0)$

is extended into

(X) The distribution of X admits a density f on $S \subset \mathbb{R}^d$ such that $b := \inf_{y \in S} f(y) > 0$ and for all $\tau \leq \tau_0$, $\int_{S \cap B(x,\tau_0)} d\lambda \geq c \int_{B(x,\tau_0)} d\lambda$.

Consider the following k-NN classifier

$$\hat{g}_n(x) = \mathbb{1}_{\{\hat{\eta}_n(x) > 1/2\}}$$

Theorem

Under (iid), (X), (E) and (MA), if η is Lipschitz, there exists $C > 0$ such that

$$0 \leq \mathbb{E}[R(\hat{g}_n)] - R(g^*) \leq C n^{-(1+\alpha)/(d+2)}$$

Related work

Chaudhuri & Dasgupta (2014). Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 27.
Derives sharp rates for $\mathbb{E}[|\hat{\eta}_n(X) - \eta^*(X)|]$ combining Hölder smoothness of η and the Tsybakov margin condition, showing the k-NN estimator achieves minimax rates.

Döring, Györfi, Walk; (2018). Rate of Convergence of k-Nearest-Neighbor Classification Rule. JMLR. See the introduction for references and precise comments on fast-rates in k-NN classification.

Cannings, Berrett & Samworth (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Annals of Statistics*, 48(3), 1789–1814.
Proves that a local (adaptive) choice of k achieves the minimax rate $O(n^{-4/(d+4)})$ for the excess risk under Hölder and margin conditions, with the bound going through $\mathbb{E}[|\hat{\eta}_n(X) - \eta^*(X)|]$.

Uniform in k

The following is inspired from Einmahl and Mason (2005); Dony et al. (2006); Giné and Nickl (2009); Goldenshluger and Lepski (2011a) where various uniform in bandwidth bound are given for kernel smoothing based estimation procedure.
As claimed in Dony et al. (2006) such result implies that "*if one chooses the bandwidth depending on the data or the location x, as is usually done in practice, one has the same order of convergence as in the case of a deterministic bandwidth sequence*"
The next result is considering the uniformity with respect to k for $\hat{\eta}^{(NW)}$.

Theorem

Suppose (iid), (X) and (E) are fulfilled. Let $n \geq 1$ and $x \in S$. There is $A > 0$ such that, for all $n \geq 1$ and $t > 0$, it holds

$$\mathbb{P}(|\hat{\eta}_n(x) - \eta^*(x)| > t) \leq 4n \exp\left(-(k^{-1/2} + (k/n)^{1/d})^2 t^2/(2A)^2\right)$$

This can be obtained as a corollary of the previous bound.

Uniform in x for k-NN

1. The goal is to obtain an inequality valid with probability $1-\delta$,

$$\sup_{x\in\mathbb{R}^d} |\hat{\eta}_n(x) - \eta^*(x)| \leq \text{Bound}(n, \delta, d)$$

   as sharp as possible. This reflects our goal of achieving uniformly good performance, and not only in regions where P places significant mass. This is more challenging, since the L_2-risk does not penalize errors in regions with little or no probability mass. As a result, different approaches may exhibit substantially different behaviors in such regions. This part is inspired from a long line of research (see e.g., Stute (1982); Einmahl and Mason (2000); Giné and Guillou (2002)) investigating the rate of convergence of local estimator such as Parzen-Rosenblat or Nadaraya-Watson.
2. Uniform confidence interval (Bickel and Rosenblatt, 1973; Giné and Nickl, 2010; Chernozhukov et al., 2014) is a clear motivation for studying uniform bound.
3. Adaptive bandwidth choice Einmahl and Mason (2005); Dony et al. (2006); Goldenshluger and Lepski (2011b)

The next is related to Jiang (2019); Portier (2025) where uniform in x bound are obtained. We have the following non-asymptotic inequality.

Theorem

Suppose (iid), (X) and (SG) are fulfilled. Let $n \geq 1$, $1 \leq k \leq n$ and $\delta \in (0,1)$, be such that

$$16d \log(12n/\delta) \leq k \leq \tau_0^d nbcV_d/2.$$

We have with probability $1-\delta$,

$$\sup_{x \in \mathbb{R}^d} |\hat{\eta}_n(x) - \eta^*(x)| \leq \sqrt{\frac{2\sigma^2 \log(dn^d/\delta)}{k}} + L\left(\frac{2k}{nbcV_d}\right)^{1/d}.$$

Proof is similar to the pointwise upper bound except that Vapnik is used in place of Chernoff bound.

Conclusion

Several non-asymptotic error bounds have been obtained for k-NN methods. They include point-wise, uniform in x and/or k. They all are based on Vapnik inequalities.

The main limitation is they are not suited to kernel functions as Vapnik's theory is valid for indicator of sets.

Suppose that all elements $g \in \mathcal{G}$ are such that $0 \leq g(x) \leq U$. Show that

$$\mathbb{E}[\sup_{g\in\mathcal{G}} n^{-1} \sum_{i=1}^{n} \{g(Z_i) - P(g(Z))\}] \leq 2U\mathbb{E}[\sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \eta_i \mathbb{1}_{\{U_i < g(Z_i)\}}]$$

for some collection $(U_i)_{1\leq i\leq n}$ of independent and identically distributed random variables. A Vapnik inequality (of the first type) for such classes can be obtained. This type of classes is refereed to as *VC subgraph classes*.

Risk minimization

Objective

Based on examples $Z_1, \dots, Z_n$ distributed as $Z = (Y, X)$, Aim is to estimate

$$\arg\min_{g \in \mathcal{G}_{all}} \left\{ R(h) := \mathbb{P}(Y \neq g(X)) \right\}$$

The optimal classifier is the Bayes classifier $g^* = \mathbb{1}_{\{\eta^* > 1/2\}}$.

(surrogate) risk minimization

A surrogate problem is introduced as

$$\arg\min_{g \in \mathcal{G}_{all}} \left\{ R_\ell(g) := \mathbb{E}[\ell(Y, g(X))] \right\}$$

where ℓ is a loss function (preferably smooth and convex).

Some differences with the regression framework

- Classification: The loss that is used in training ℓ is not the same as the 0-1 loss.
- Regression; Square, Huber loss, quantile loss. Output Y is unbounded.

What are the challenges of empirical risk minimization in classification?

Four difficulties

1. The risk for training is not the same as the misclassification risk. One needs to work with surrogate losses $\ell(y, g(x))$ leading to

$$R_\ell(g) := \underset{g \in \mathcal{G}_{all}}{\arg\min} \left\{ R_\ell(g) := \mathbb{E}(\ell(Y, g(X))) \right\}$$

2. **Both risk functions R and R_ℓ are unknown** and should be estimated from the data
3. **An optimization set $\mathcal{G}$ needs to be selected**. The oracle predictor (minimizer of the risk) is given by

$$g^* = \underset{g \in \mathcal{G}_{all}}{\arg\min} \left\{ R(g) := \mathbb{P}(g(X) \neq Y) \right\}$$

   Optimizing over $\mathcal{G}_{all}$ is not possible. Need for flexible but small enough spaces.
4. **There is an optimization problem to solve** once $\mathcal{G}$ is set.

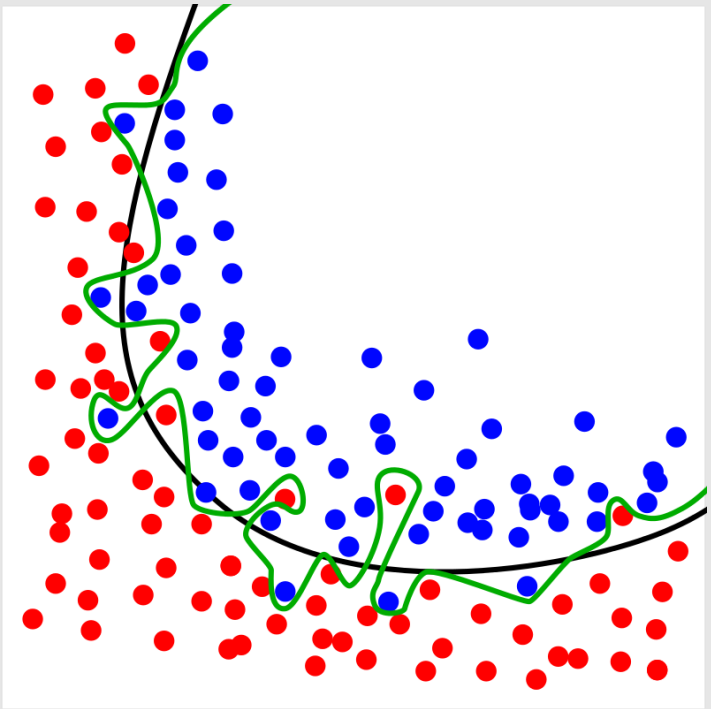
Point 4 will not be considered in this course

Overfitting

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} R_n(g)$$

Definition

Overfitting happens when $\mathcal{G}$ is so large that the output is perfectly predicted by the predictor. In case of overfitting $\hat{R}_n(\hat{g}_n)$ is really close to 0 but $R(\hat{g}_n)$ is not. **No generalization on new data.**



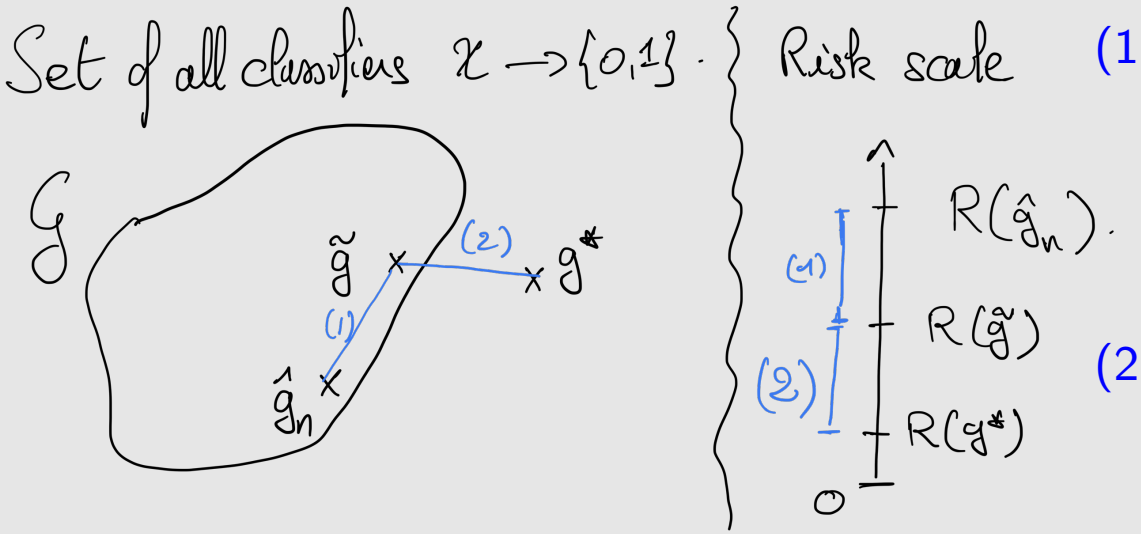
Classification in which overfitting has happened (green classifier). The empirical risk is $\hat{R}_n(\hat{g}_n) = 0$ whereas $R(\hat{g}_n)$ is certainly larger. The bound in the following inequalities will be large because of the model complexity.

Overfitting implies a large variance error

Bias-variance trade-off

The Bayes classifier is defined as

$$g^* \in \arg\min_{g \in \mathcal{G}_{\text{all}}} R(g).$$



(1) Statistical error (variability due to randomness in data). This is the **variance** component.

(2) Modeling error (gap due to model choice). This is the **bias** component.

Increasing the size of $\mathcal{G}$ is not a solution

- When $\mathcal{G}$ increases the bias diminishes
- When $\mathcal{G}$ increases the variance increases (see the oracle inequality)

Model selection through penalization

Definition

It consists in minimizing

$$\hat{R}_n(g) + \lambda \text{pen}(g)$$

(instead of $\hat{R}_n(g)$), where the function $h \mapsto \text{pen}(g) \in \mathbb{R}_{\geq 0}$ measures the complexity of the model in the sense that:

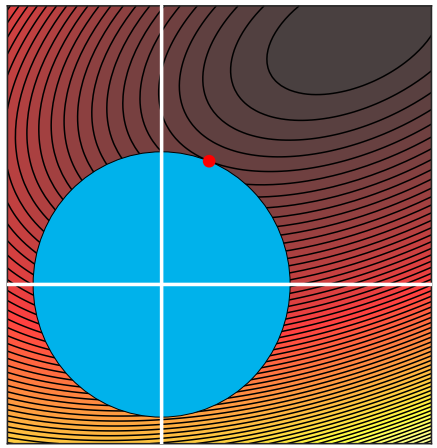
- a complex model g has a large image value, i.e., $\text{pen}(g)$ is large
- a simple model is such that $\text{pen}(g) \simeq 0$

The Lagrangian formulation is

$$\min_{g \,:\, \text{pen}(g) \leq C} \hat{R}_n(g)$$

for some $C > 0$ related to λ in this way:

- Large value of λ implies small C, i.e., small model
- Small value of λ implies large C, i.e., complex model



Level curve of $g \mapsto \hat{R}_n(g)$

region $\text{pen}(g) \leq C$

From finite dictionary...

As claimed in Boucheron et al. (2005): "Before seeing the data, we do not know which function the algorithm will choose"

Lemma 1

$$R(\hat{h}_n) - R(h^*) \leq 2 \sup_{g \in \mathcal{G}} |\hat{R}_n(g) - R(g)|$$

While the above is a coarse upper bound, it is still well used (Bartlett et al., 2021) as it leads to the minimax rate for some distribution classes.

Lemma 2 (Hoeffding inequality)

Under $(Z_i)_{i \geq 1}$ be an iid sequence of random variables valued in $[0, 1]$ (or in $[-1, 0]$), then for all $n \geq 1$ and all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \{Z_i - E[Z_1]\} > t \right) \leq \exp(-2t^2/n).$$

From finite dictionary...

Theorem (oracle inequality for finite dictionary)

Under (iid), suppose that $\mathcal{G} = \{g_1, \dots, g_M\}$, then for all $\delta \in (0,1)$, it holds with probability $1-\delta$,

$$R(\hat{g}_n) \leq \inf_{g \in \mathcal{G}} R(g) + \sqrt{\frac{2\log((M+1)/\delta)}{n}}$$

Influence of the different parameters: n and M are playing opposite role in the quality of the bound. Complex model requires more data. Also one has, with probability $1-\delta$, for all g

$$R(g) \leq \hat{R}_n(g) + \sqrt{\frac{2\log(M/\delta)}{n}}$$

allowing to picture an explicit "overfitting" vs "complexity" trade-off.

... to the infinite case with Vapnik

Note that for $g \in \{0,1\}$ and $y \in \{0,1\}$,

$$\mathbb{1}_{\{g \neq y\}} = \mathbb{1}_{\{(2g-1)(2y-1) \leq 0\}} = \frac{1}{2}(-(2g-1)(2y-1)+1) :$$

We get

$$\arg\min_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{g(X_i) \neq Y_i\}} = \arg\min_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} -(2g(X_i)-1)(2Y_i-1)/2$$

We have

$$\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i \{-(2g(X_i)-1)(2Y_i-1)\}/2 \stackrel{d}{=} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i (2g(X_i)-1)/2$$

The Vapnik dimension is such that $\mathbb{S}_n(\mathcal{G}) = \mathbb{S}_n(\mathcal{G} - 1/2)$. And we obtain

$$\sup_{g \in \mathcal{G}} \{\hat{R}_n(g) - R(g)\} \leq \sqrt{8 n^{-1} \log(\mathbb{S}_n(\mathcal{G})/\delta)}$$

Theorem (the infinite case with Vapnik)

Under (iid), then for all $\delta \in (0,1)$, it holds with probability $1-\delta$,

$$R(\hat{g}_n) \leq \inf_{g \in \mathcal{G}} R(g) + \sqrt{8n^{-1}\log(2\mathbb{S}_n(\mathcal{G})/\delta)}$$

or, with probability $1-\delta$, for all classifier g,

$$R(g) \leq \hat{R}(g) + \sqrt{8n^{-1}\log(\mathbb{S}_n(\mathcal{G})/\delta)}$$

Using that $\mathbb{S}_n(\mathcal{G}) \leq (en/d_V)^{d_V}$, we get a clear decomposition with an overfitting term $\hat{R}_\ell(g)$ and a complexity term $\sqrt{d_V/n}$.

Limitation

- The Vapnik approach is limited to $0-1$-loss and optimizing binary choices g because of the Vapnik's dimension definition
- More general classes might be consider through VC-subgraphs or entropy consideration

1 Background on deviation inequalities

2 Background on classification

3 k-nearest neighbors
- (fast) learning rates for k-NN classification
- Uniform in x and k bound

4 Empirical risk minimization
- The background
- The Vapnik inequality in classification
- From logistic regression to neural network
- Surrogate losses and calibration

Logistic Regression

- Logistic regression is a standard classification algorithm
- The algorithm predicts the probability of an instance belonging to a class and assigns the label based on a threshold.
- The sigmoid function is used to map predictions to probabilities:

$$q_\beta(x) = \frac{1}{1 + \exp(-(\beta^T x))}$$

- The β coefficients are estimated minimizing the *logistic risk* which can be derived following maximum likelihood principle.

Estimation

Cross-entropy risk

Some covariates $X_1, \dots, X_n$ are observed with labels $Y_1, \dots, Y_n$. The model is now $Y_i \sim \mathcal{B}(p(x_i))$ for each i, where p is unknown. Define the empirical risk

$$\hat{R}_n(q) = -n^{-1} \sum_{i=1}^{n} Y_i \log(q(X_i))) + (1 - Y_i) \log(1 - q(X_i)))$$

One may show that minimizing $R_n(q)$ with respect to q is equivalent to minimizing an unbiased estimator of the **cross-entropy** between p and q. It is also equivalent to maximizing the **likelihood** of the observations.

Training logistic regression

Empirical risk approach

The logistic regression estimator is given

$$\hat{\beta}_n \in \arg\max_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \{ Y_i \log(q_\beta(X_i)) + (1 - Y_i) \log(1 - q_\beta(X_i)) \}$$

Equivalently

$$\arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \{ \log(1 + \exp(-\tilde{Y}_i \beta^T X_i)) \}$$

with $\tilde{Y}_i = 2Y_i - 1$.

Extension

- The loss $\ell(\tilde{Y}_i \beta^T X_i) = \log(1 + \exp(-\tilde{Y}_i \beta^T X_i)$ could be modified
- Instead of $\tilde{Y} \beta^T X$ in the above one may use an activation function $\sigma(\tilde{Y} \beta^T X)$
- More generally on may use a combination of activation function $\sum_{j=1}^{K} b_j \sigma(\tilde{Y} \beta_j^T X)$

Surrogate losses: an introduction

New notation and further question

Transforming g into $f = 2g - 1$ and Y into $\tilde{Y} = (2Y - 1)$, we write the $0-1$-loss $\ell(\tilde{y}, f) = \mathbb{1}_{\{\tilde{y}f \leq 0\}}$. Motivation is then to maximize the margin $f\tilde{Y}$ empirically:

$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i)\tilde{Y}_i)$$

One question follows: **Can we consider other losses and more general function spaces than the above?**

Suppose there is a dominating loss ℓ such that

$$\mathbb{1}_{\{yf<0\}} \leq \ell(yf)$$

implying

$$R(f) \leq R_\ell(f)$$

leading to for all f

$$R(f) \leq \hat{R}_\ell(f) + R_\ell(f) - \hat{R}_\ell(\hat{f}) \leq \hat{R}_\ell(f) + \sup_f R_\ell(f) - \hat{R}_\ell(f)$$

Surrogate losses

This approach is proposed by (Bartlett et al., 2006). The loss function $\ell : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is decreasing to maximizing the margin $\tilde{y}f$ whenever minimizing $\ell(yf)$.

Definition

Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function. For $\eta \in [0, 1]$, define the conditional risk of a score $f \in \mathbb{R}$ as

$$C(f, \eta) = \eta\, \ell(f) + (1 - \eta)\, \ell(-f),$$

and its infimum over all scores as

$$C^*(\eta) = \inf_{f \in \mathbb{R}} C(f, \eta).$$

and the infimum over scores disagreeing with η

$$C(\eta) = \inf_{f \in \mathbb{R} \,:\, f(2\eta - 1) \leq 0} C(f, \eta).$$

The loss ℓ is said to be *classification calibrated* if for every $\eta \neq 1/2$,

$$\psi_\ell(\eta) := C(\eta) - C^*(\eta) > 0$$

Remark

Classification calibration is the minimal condition ensuring that a surrogate risk minimizer $\hat{f} = \arg\min_f \mathbb{E}[\ell(Yf(X))]$ (with $Y \in \{-1, +1\}$) is consistent for the Bayes classifier $g^*(x) = \mathrm{sign}(\eta^*(x) - 1/2)$.

Sufficient condition for classification calibration

Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex loss that is differentiable at 0. Then ℓ is classification calibrated if and only if $\ell'(0) < 0$.

Theorem Bartlett et al. (2006)

Suppose that ℓ is calibrated, e.g., convex such that $\ell'(0) < 0$, then for all g,

$$\psi_\ell \left(R(g) - \inf_{g \in \mathcal{G}_{all}} R(g) \right) \leq R_\ell(g) - \inf_{g \in \mathcal{G}_{all}} R(g)$$

where $\psi_\ell(\eta) := \inf_{f \leq 0} C(f, \eta) - \inf_f C(f, \eta)$.

Margin based classifier

Consider $\tilde{Y} \in \{-1, 1\}$ and score function class

$$\mathcal{F} = \{x \mapsto f(x) \in \mathbb{R}\}$$

Training of $\hat{f}$ is done maximizing empirical margin through ℓ - or minimizing the empirical risk

$$R_{n,\ell}(f) = \sum_{i=1}^{n} \ell(\tilde{Y}_i f(x_i))$$

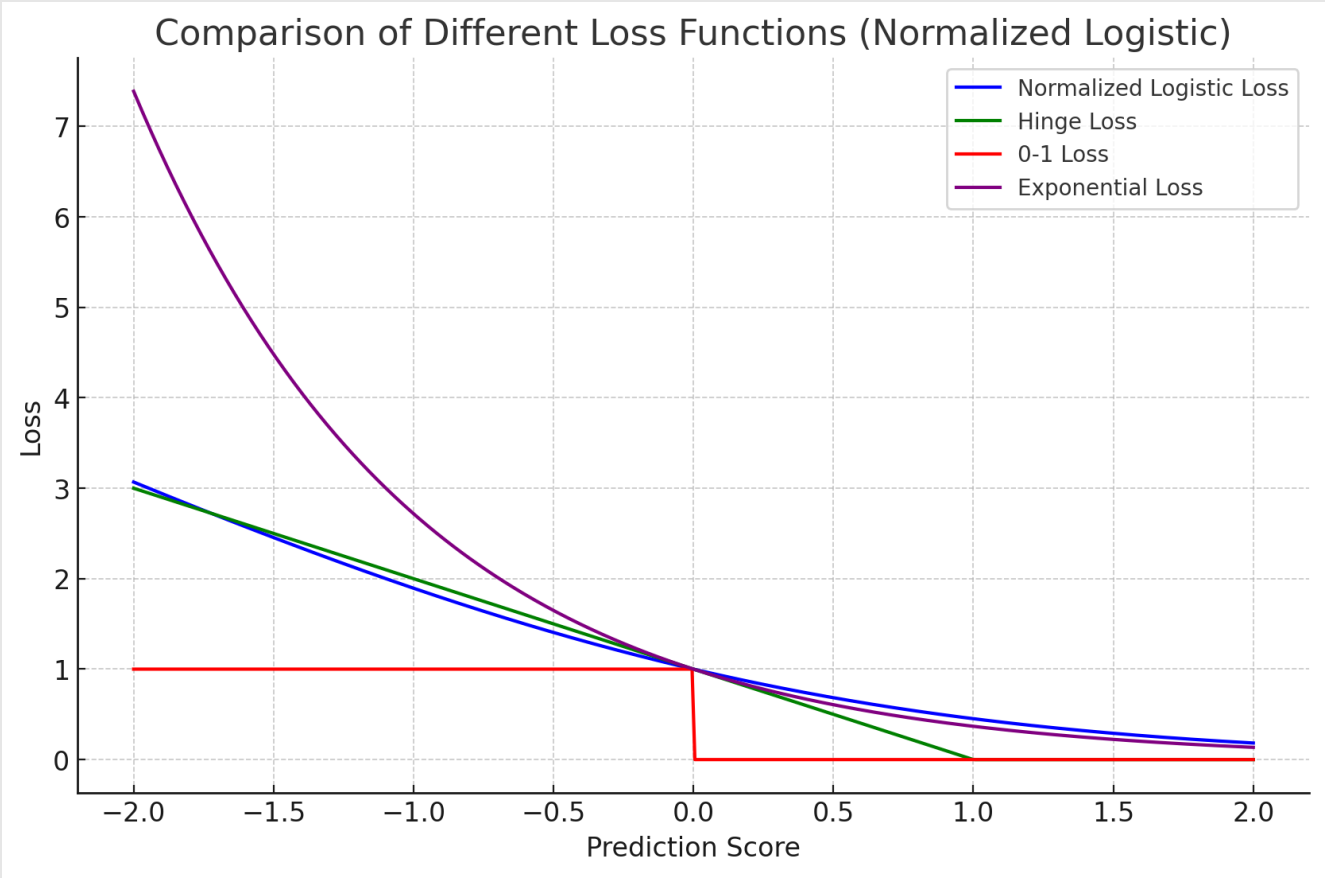
The resulting classifier is $\hat{g}(x) = 1_{\{\hat{f}(x)>0\}}$.

Advantages

- General framework where $\hat{R}_{n,\ell} \neq \hat{R}_n$ with different losses.
- Many different hypothesis class for f (linear map, neural network, kernel function...)
- penalty term $\mathrm{pen}(h)$ can be added, e.g., LASSO and SVM

Examples of losses

- 0-1 **loss** $\ell(v) = \mathbb{1}_{v<0}$
- **exponential loss** $\ell(v) = \exp(-v)$ (e.g., Adaboost)
- **logistic loss** $\ell(v) = \log(1+\exp(-v))/\log(2)$
- **hinge loss** $\ell(v) = (1-v)_+$



By McDiarmid's inequality, if all $g \in \mathcal{G}$ are valued in $[0,1]$,

$$\left| \sup_{g \in \mathcal{G}} |P_n(g) - P(g)| - \mathbb{E}[\sup_{g \in \mathcal{G}} |P_n(g) - P(g)|] \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

Theorem (the general case Bartlett et al. (2006))

Suppose (iid) is fulfilled. Let $\delta \in (0,1)$, $c \in \mathbb{R}$ and ℓ be classification calibrated such that $\ell(\mathcal{F})$ and $\ell(-\mathcal{F})$ are bounded by M. It holds with probability $1-\delta$,

$$\psi_\ell \left(R(\hat{f}_n) - \inf_{f \in \mathcal{F}_{all}} R(f) \right)$$

$$\leq \left(\inf_{f \in \mathcal{F}} R_\ell(f) - \inf_{f \in \mathcal{F}_{all}} R_\ell(f) \right) + 4Rad_n(\ell \circ \mathcal{F} - c) + 2M\sqrt{\frac{\log(1/\delta)}{2n}}$$

2 approaches to deal with neural network functions

- The contraction lemma (Bartlett and Mendelson, 2002; Golowich et al., 2018)
- The Dudley integral entropy bound (Bartlett et al., 2017)

Direct approach to bound Rademacher complexity

Contraction lemma (Ledoux and Talagrand, 1991)

If ℓ is L-Lipschitz and $\ell(0) = 0$, we have

$$Rad_n(\ell \circ \mathcal{F}) \leq 2LRad_n(\mathcal{F})$$

If $\ell_{\mathcal{F}} = \{(x, y) \mapsto \ell(yf(x))\}$ with L-Lipschitz and $\ell(0) = 0$, we have

$$Rad_n(\ell_{\mathcal{F}}) \leq 2LRad_n(y\mathcal{F}) = 2LRad_n(\mathcal{F})$$

2-Layer neural network (Bartlett and Mendelson, 2002)

For a 2-layers neural network defined on $[-1, 1]^d$

$$\mathcal{F}_B = \left\{ x \mapsto \sum_{j=1}^{K} b_j \sigma(\langle w_j, x \rangle) \, : \, \|b\|_1 \leq 1, \max_{j=1,\dots,K} \|w_j\|_1 \leq B, k \geq 1 \right\},$$

where σ is 1-Lipschitz and $\sigma(0) = 0$. We have

$$Rad_n(\mathcal{F}_B) \leq B\sqrt{\frac{8\log(2d)}{n}}.$$

We say that $\sigma : \mathbb{R} \to \mathbb{R}$ is 1-homogeneous if $\sigma(\alpha x) = \alpha\sigma(x)$ for all $x \in \mathbb{R}$ and $\alpha \geq 0$. The ReLU nonlinearity $\sigma(x) = x \vee 0$ has this property.

Theorem (Golowich et al., 2018)

Let $\bar{\sigma} : \mathbb{R} \to \mathbb{R}$ be a fixed 1-homogeneous nonlinearity, and define the componentwise version $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ via $\sigma_i(x)_j = \bar{\sigma}(x_j)$. Consider a network $f(x; theta)$, with L layers of these nonlinearities and parameters $\theta = (W_1, \ldots, W_L)$, given by

$$f(x; theta) = \sigma_L(W_L\sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1 x) \cdots)).$$

Define the class of functions on the unit Euclidean ball in $\mathbb{R}^d$,

$$\mathcal{F}_B = \{f(\cdot; theta) : \|W_i\|_F \leq B\},$$

where $\|W_i\|_F$ denotes the Frobenius norm of W_i. Then we have

$$Rad_n(\mathcal{F}_B) \lesssim \frac{\sqrt{L}B^L}{\sqrt{n}}.$$

Further topics

- Better rates are possible using margin condition Bartlett et al. (2006) or Bernstein conditions Bartlett et al. (2005)
- An analysis via the covering numbers of neural network is proposed in (Bartlett et al., 2017)
- Comparison to the hypothesis class $\mathcal{F}$ instead of $\mathcal{F}_{all}$ in the surrogate loss result can be more informative and is conducted in Mao et al. (2023)

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.

Bartlett, P., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *Annals of Statistics 33*(4), 1497–1537.

Bartlett, P., D. J. Foster, and M. J. Telgarsky (2017). Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*(473), 138–156.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res. 3*(Spec. Issue Comput. Learn. Theory), 463–482.

Bartlett, P. L., A. Montanari, and A. Rakhlin (2021). Deep learning: a statistical viewpoint. *Acta Numerica 30*, 87–201.

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.

Bickel, P. J. and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics 1*(6), 1071–1095.

Boucheron, S., O. Bousquet, and G. Lugosi (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat. 9*, 323–375.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics 42*(4), 1564–1597.

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

Dony, J., U. Einmahl, and D. M. Mason (2006). Uniform in bandwidth consistency of local polynomial regression function estimators. *Austrian Journal of Statistics 35*(2&3), 105–120.

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.

Einmahl, U. and D. M. Mason (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics 33*(3), 1380–1403.

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.

Giné, E. and R. Nickl (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields 143*(3), 569–596.

Giné, E. and R. Nickl (2010). Confidence bands in density estimation. *Annals of Statistics 38*(2), 1122–1170.

Goldenshluger, A. and O. Lepski (2011a). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Annals of Statistics 39*(3), 1608–1632.

Goldenshluger, A. and O. V. Lepski (2011b). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist. 39*, 1608–1632.

Golowich, N., A. Rakhlin, and O. Shamir (2018). Size-independent sample complexity of neural networks. In *Conference on learning theory*, pp. 297–299. PMLR.

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

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.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach spaces*, Volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Mao, A., M. Mohri, and Y. Zhong (2023, 23–29 Jul). Cross-entropy loss functions: Theoretical analysis and applications. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 23803–23828. PMLR.

Portier, F. (2025). Nearest neighbor empirical processes. *Bernoulli 31*(1), 312–332.

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.

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

Stute, W. (1982). A law of the logarithm for kernel density estimators. *The annals of Probability*, 414–422.

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.