Lecture Notes:
Beyond Empirical Risk minimization
Local Averages and K-Nearest Neighbors

Alessandro Rudi, Pierre Gaillard

April 2019

1 Beyond Empirical Risk Minimization: Local Averages

In this lecture we start from a different characterization of the target function f^*. We have seen that it is defined globally as $f^* : X \to Y$ satisfying

$$R(f^*) = \inf_{f:X\to Y} R(f),$$

(the infimum is over the measurable functions from X to Y) that is

$$f^* \in \arg\min_{f:X\to Y} R(f).$$

Assume without loss of generality that

$$f^* = \arg\min_{f:X\to Y} R(f). \tag{1}$$

We can provide a pointwise characterization of f^* as follows

Theorem 1. *When Y is a compact set and ℓ is continuous, then*

$$f^*(x) = \arg\min_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x], \tag{2}$$

almost everywhere, where $\mathbb{E}[q(y) \mid x]$ denotes the conditional expectation of $q(y)$ given x, with $q : Y \to \mathbb{R}$.

Proof. We sketch the proof as follows. Denote by $\widetilde{f}$ the function in Eq. 2. Note that by definition

$$\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x] = \inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x],$$

almost everywhere. Then, by noting that for any function $f : X \to Y$

$$\mathbb{E}[\ell(f(x), y) \mid x] \geq \inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x] = \mathbb{E}[\ell(\widetilde{f}(x), y) \mid x],$$

we have for any $f : X \to Y$

$$R(\widetilde{f}) = \mathbb{E}[\ell(\widetilde{f}(x), y)] = \mathbb{E}_x[\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x]] = \mathbb{E}_x[\inf_{y' \in Y} \mathbb{E}[\ell(y', y) \mid x]] \tag{3}$$

$$\leq \mathbb{E}_x[\inf_{y' \in Y} \mathbb{E}[\ell(y', y) \mid x]] \leq \mathbb{E}_x[\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x]] = R(f). \tag{4}$$

So $R(\widetilde{f}) = \inf_{f:X\to Y} R(f)$. To conclude the proof we need to prove that $\widetilde{f}$ is measurable, which is rather technical and out of the scope of the lecture (see [1]). □

2 Learning via Local Averages

While the characterization in Eq. (1) suggested approaches like empirical risk minimization (we have seen it in the previous lecture), the characterization in terms of Eq. (2) gave rise to the so called *local average methods*. Denoting by $\rho(y|x)$ the conditional probability of y given x, and by $\widehat{\rho}(y|x)$ an estimator for $\rho(y|x)$, local averages estimators are of the form

$$\widehat{f}(x) = \arg\min_{y' \in Y} \int \ell(y', y) d\widehat{\rho}(y|x).$$

To study the excess risk for this estimator we perform the following analysis. Denote by $E(y', x)$ the function $E(y', x) = \int \ell(y', y) d\rho(y|x)$ and by $\widehat{E}(y', x)$ the function $\widehat{E}(y', x) = \int \ell(y', y) d\widehat{\rho}(y|x)$, then

$$\begin{aligned} R(\widehat{f}) - R(f^*) &= \mathbb{E}_x \left[E(\widehat{f}(x), x) - E(f^*(x), x) \right] \\ &= \mathbb{E}_x \left[E(\widehat{f}(x), x) - \widehat{E}(\widehat{f}(x), x) \right] + \mathbb{E}_x \left[\widehat{E}(\widehat{f}(x), x) - E(f^*(x), x) \right]. \end{aligned}$$

Now note that

$$\mathbb{E}_x \left[E(\widehat{f}(x), x) - \widehat{E}(\widehat{f}(x), x) \right] \leq \mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

Moreover, since $\widehat{E}(\widehat{f}(x), x) = \inf_{y' \in Y} \widehat{E}(y', x)$ and $E(f^*(x), x) = \inf_{y' \in Y} E(y', x)$, then

$$\mathbb{E}_x \left[\widehat{E}(\widehat{f}(x), x) - E(f^*(x), x) \right] = \mathbb{E}_x \left[\inf_{y' \in Y} \widehat{E}(y', x) - \inf_{y' \in Y} E(y', x) \right] \leq \mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

So finally

$$R(\widehat{f}) - R(f^*) \leq 2\mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

3 Estimators for the conditional expectation

Assume here to have $X \subseteq \mathbb{R}^d$, that $Y \subset \mathbb{R}$ and that $\rho(y|x), \rho(y, x), \rho(x)$ are probability densities. We characterize $\rho(y|x)$ as

$$\rho(y|x) = \frac{\rho(y, x)}{\rho(x)}.$$

Usually estimators for the conditional probability have the following form

$$\widehat{\rho}(y|x) = \frac{\widehat{\rho}(y,x)}{\widehat{\rho}(x)},$$

where $\widehat{\rho}(y,x)$ and $\widehat{\rho}(x)$ are estimators for $\rho(y,x)$ and $\rho(x)$. Now we introduce some methods to estimate probability densities.

3.1 Density estimation

A classical way to estimate probability density is to approximate it via convolutions of the empirical distribution. Let q be a probability density (i.e. $q(x) = e^{-\|x\|^2}$) and $\tau^{-d} q_\tau(x) = q(x/\tau)$, for $\tau > 0$. Let moreover $x_1, \ldots, x_n$ sampled i.i.d. from ρ. We define the estimator as

$$\widehat{\rho}(x) = \frac{1}{n} \sum_{i=1}^{n} q_\tau(x - x_i).$$

By denoting by $\widehat{\rho}_n$ the probability $\widetilde{\rho}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ (where δ is the Dirac's delta) and by $\star$ the convolution operator (i.e. $(f \star g)(x) = \int f(y) g(x-y) dy$) we have

$$\rho \quad \approx \quad \rho \star q_\tau \quad \approx \quad \widetilde{\rho}_n \star q_\tau \quad = \quad \widehat{\rho}(x).$$

In particular

Lemma 1. *Let $|\rho(x) - \rho(y)| \leq C\|x - y\|$ for any x, y, then*

$$|\rho(x) - (\rho \star q_\tau)(x)| \leq CT\tau,$$

where $T := \int \|z\| q(z) dz$. (The integrals are assumed on $\mathbb{R}^d$).

Proof. Since $\int q_\tau(x-y)dy = \int q_\tau(y)dy = 1$, we have

$$\begin{aligned}|\rho(x) - (\rho \star q_\tau)(x)| &= |\tau^{-d} \int (\rho(x) - \rho(y)) q((x-y)/\tau) dy| \leq \tau^{-d} \int |\rho(x) - \rho(y)| q((x-y)/\tau) dy \\ &\leq C\tau^{-d+1} \int \|x - y\|/\tau \; q((x-y)/\tau) dy = C\tau^{-d+1} \int \|u/\tau\| q(u/\tau) du = C\tau \int \|z\| q(z) dz,\end{aligned}$$

where the last step is due to the change of variable $u/\tau \in \mathbb{R}^d \mapsto z \in \mathbb{R}^d$. □

Lemma 2. *For any $v \in X$, we have*

$$\mathbb{E}|(\rho \star q_\tau)(v) - \frac{1}{n}\sum_{i=1}^n q_\tau(v - x_i)|^2 \leq \frac{Q\tau^{-d}}{n},$$

where $Q = \max_t q(t)$.

Proof. Define the random variable $z = q_\tau(v - x)$, with x distributed according to ρ. Now note that

$$\mathbb{E}z = \int q_\tau(v-x) d\rho(x) = \int q_\tau(v-x)\rho(x)dx = \rho \star q_\tau.$$

Let $z_1, \dots, z_n$ defined as $z_i = q_\tau(v - x_i)$, since $x_1, \dots, x_n$ are independently and identically distributed according to ρ, then $z_1, \dots, z_n$ are independent copies of z and

$$\mathbb{E}|\frac{1}{n}\sum_{i=1}^{n}(z_i - \mathbb{E}z)|^2 = \frac{1}{n}\mathbb{E}(z_1 - \mathbb{E}z)^2$$

Now

$$\mathbb{E}(z - \mathbb{E}z)^2 \leq \mathbb{E}z^2 = \int q_\tau(v-x)^2\rho(x)dx \leq (\max_t q_\tau(t)) \int q_\tau(v-x)\rho(x)dx = \max_t q_\tau(t) = \tau^{-d}\max_t q(t).$$

□

Finally

Theorem 2. *Let ρ such that $|\rho(x) - \rho(y)| \leq C\|x - y\|$, then for any $v \in X$*

$$\left(\mathbb{E}\left|\rho(v) - \frac{1}{n}\sum_{i=1}^{n} q_\tau(v - x_i)\right|^2\right)^{1/2} \leq CT\ \tau\ +\ \sqrt{\frac{Q\tau^{-d}}{n}}.$$

Proof. The result is obtained combining the two lemmas above □

The estimator for $\rho(x, y)$ can be derived in the same way, using $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^{d'}$ with $d' = d + p$ where d is the dimension of the euclidian space containing X and p the dimension of the space containing Y.

4 k-Nearest Neighbours

We would like to classify objects, described with vectors x in $\mathbb{R}^d$, among $L+1$ classes $\mathcal{Y} := \{0, \dots, L\}$ in an automatic fashion. To do so, we have at hand a labelled data set of n data points $(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$ for $1 \leq i \leq n$. The data is assumed to be the realization of i.i.d. random variables (X_i, Y_i) following a distribution ν. The goal of this lesson is to build a classifier, i.e., a function

$$g : \mathbb{R}^d \to \mathcal{Y}$$

which minimizes the probability of mistakes: $\mathbb{P}_{(X,Y)\sim\nu}\{g(X) \neq Y\}$. The latter can be rewritten as the expected risk $R(g) := \mathbb{E}_{(X,Y)\sim\nu}[\mathbb{1}_{g(X)\neq Y}]$ of the 0-1 loss.

The k-nearest neighbor classifier works as follows. Given a new input $x \in \mathbb{R}^d$, it looks at the k nearest points x_i in the data set $D_n = (x_i, y_i)$ and predicts a majority vote among them. The k-nearest neighbor classifier is quite popular because it is simple to code and to understand; it has nice theoretical guarantees as soon as k is appropriately chosen and performs reasonably well in low dimensional spaces. In this notes, we will investigate the following questions:

- consistency: does k-NN has the smallest possible probability of error when the number of data grows?
- how to choose k?

There are plenty of other possible interesting questions. How should we choose the metric (invariance properties,...)? Can we get improved performance by using different weights between neighbors (see Kernel methods)? Is it possible to improve the computational complexity (by reducing the data size or keeping some data in memory,...). These questions are however beyond the scope of these lecture notes and we refer the interested reader to the book [3].

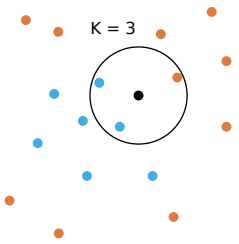

Figure 1: k-nearest neighbors with two classes (orange and blue) and $k = 3$. The new input (i.e., the black point) is classified as blue which corresponds to the majority class among its three nearest neighbors.

Assumptions and notation For simplicity, we assume the binary case: $L = 1$ and $\mathcal{Y} = \{0, 1\}$. For each $l \in \{0, 1\}$, we denote by μ_l the law of X given $Y = l$ under ν and p_l the marginal distribution of Y:

$$\mu_l = X \mid \{Y = l\} \qquad \text{and} \qquad p_l = P_{(X,Y)\sim\nu}\{Y = l\}\,.$$

We also assume that μ_l is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$. We denote by f_l its density. And for each $x \in \mathbb{R}^d$, we denote by

$$\eta(x) := \mathbb{P}_{(X,Y)\sim\nu}\big\{Y = 1 | X = x\big\}\,. \tag{5}$$

In the following except when stated otherwise the expectation and probability are according to $(X, Y) \sim \nu$. For clarity, we will omit the subscript $(X, Y) \sim \nu$ in $\mathbb{E}$ and $\mathbb{P}$. In some cases, if the classifier is random, for instance because it was build on the random data set (X_i, Y_i) the expectation might also be taken with respect to the classifier itself. But it will be explicited.

Lemma 3. *For any classifier $g : \mathbb{R}^d \to \mathcal{Y}$, $R(g) = \mathbb{E}_{(X,Y)\sim\nu}\big[\eta(X)\mathbb{1}_{g(X)=0} + (1 - \eta(X))\mathbb{1}_{g(X)=1}\big]$.*

Proof.

$$\begin{aligned}
R(g) &= \mathbb{E}\big[\mathbb{1}_{g(X)\neq Y}\big] = \mathbb{E}\big[\mathbb{E}\big[\mathbb{1}_{g(X)\neq Y}|X\big]\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbb{1}_{g(X)\neq Y}|X, Y = 1\big]\mathbb{P}\{Y = 1|X\} + \mathbb{E}\big[\mathbb{1}_{g(X)\neq Y}|X, Y = 0\big]\mathbb{P}\{Y = 0|X\}\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbb{1}_{g(X)\neq 1}|X, Y = 1\big]\eta(X) + \mathbb{E}\big[\mathbb{1}_{g(X)\neq 0}|X, Y = 0\big](1 - \eta(X))\big] \\
&= \mathbb{E}\big[\mathbb{1}_{g(X)=0}\eta(X) + \mathbb{1}_{g(X)=1}(1 - \eta(X))\big]\,.
\end{aligned}$$

□

4.1 The (optimal) Bayes classifier

It is worth to notice that a random classifier sampling $g(X) = 0$ and $g(X) = 1$ with probability $1/2$ has an expected risk $1/2$. Hence, we will only focus on non-trivial classifiers that outperform this expected error. If the function η was known, one could define the Bayes classifier as follows:

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

Lemma 4. *The risk of the Bayes classifier is*

$$R^* := R(g^*) = \mathbb{E}\big[\min\{\eta(X), 1-\eta(X)\}\big].$$

Furthermore, for any classifier g we have

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

The above lemma implies that the Bayes classifier is optimal and $R^* = \min_{g:\mathbb{R}^d \mapsto \{0,1\}} R(g)$. The goal of this lesson is to build a classifier that gets close to R^*. We call such estimator consistent.

Definition 1 (Consistency)**.** *We say that an estimator $\hat{g}_n$ is consistent if*

$$\mathbb{E}_{(X_i,Y_i)\sim\nu}\big[R(\hat{g}_n)\big] \underset{n\to+\infty}{\longrightarrow} R^*.$$

Proof. Applying Lemma 3, we get from the definition of g^*

$$\begin{aligned} R^* &= \mathbb{E}\big[\eta(X)\mathbb{1}_{g^*(X)=0} + (1-\eta(X))\mathbb{1}_{g^*(X)=1}\big] \\ &= \mathbb{E}\big[\eta(X)\mathbb{1}_{\eta(x)<1/2} + (1-\eta(X))\mathbb{1}_{\eta(X)\geq 1/2}\big] \\ &= \mathbb{E}\big[\min\{\eta(X), 1-\eta(X)\}\big],. \end{aligned}$$

Furthermore, let $g : \mathbb{R}^d \to \mathcal{Y}$, then

$$\begin{aligned} R(g) - R^* &= \mathbb{E}\big[\eta(X)(\mathbb{1}_{g(X)=0} - \mathbb{1}_{g^*(X)=0}) + (1-\eta(X))(\mathbb{1}_{g^*(X)=1} - \mathbb{1}_{g^*(X)=1})\big] \\ &= \mathbb{E}\big[(2\eta(X)-1)(\mathbb{1}_{g(X)=0} - \mathbb{1}_{g^*(X)=0})\big] \\ &= \mathbb{E}\big[(2\eta(X)-1)\mathbb{1}_{g(X)\neq g^*(X)}\mathrm{sign}(1-2\mathbb{1}_{g^*(X)=0})\big] \end{aligned}$$

But $\mathrm{sign}(1-2\mathbb{1}_{g^*(X)=0}) = \mathrm{sign}(1-2\mathbb{1}_{\eta(X)\leq 1/2}) = \mathrm{sign}(2\eta(X)-1)$ which concludes the proof. □

Therefore, if η was known, one could compute the optimal classifier g^*. However, η is unknown and one should thus estimate it.

4.2 Plug-in estimator

Let $\hat{\eta}_n$ be an estimator of η, i.e., $\hat{\eta}_n$ is a function of the observation $D_n = (X_i, Y_i)_{1\leq i\leq n}$ which takes values in the functions from $\mathbb{R}^d$ to $[0,1]$. We will omit in the following the dependence of $\hat{\eta}_n$ in the data D_n. From $\hat{\eta}_n$, we can build the plug-in estimator as follows:

$$\hat{g}_n(x) = \begin{cases} 1 & \text{if } \hat{\eta}_n(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

Hopefully, if $\hat{\eta}_n$ is close enough to η the estimator $\hat{g}_n$ will be also close to g^* and will have a small risk. This is formalized be the following Lemma.

Lemma 5. *If $\hat{g}_n$ is defined in* (6)*, then* $R(\hat{g}_n) - R^* \leq 2\mathbb{E}_{(X,Y)\sim\nu}\big[\big|\eta(X) - \hat{\eta}_n(X)\big|\ \big|D_n\big]$.

Proof. From Lemma 4, we have

$$R(\hat{g}_n) - R^* = 2\mathbb{E}\big[|\eta(X) - 1/2|\mathbb{1}_{\hat{g}_n(X)\neq g^*(X)}|D_n\big]$$

To prove the Lemma it suffices to show that for all $x \in \mathbb{R}^d$ $|\eta(x) - 1/2|\mathbb{1}_{\hat{g}_n(x)\neq g^*(x)} \leq |\eta(x) - \hat{\eta}_n(x)|$. Let $x \in \mathbb{R}^d$. We can assume that $\mathbb{1}_{\hat{g}_n(x)\neq g^*(x)} \neq 0$ which implies that $\hat{\eta}_n(x) - 1/2$ and $\eta(x) - 1/2$ have opposite sign. In particular this yields

$$|\eta(x) - 1/2| \leq |\eta(x) - 1/2| + |1/2 - \hat{\eta}_n(x)| = |\eta(x) - \hat{\eta}_n(x)|$$

which concludes the proof. □

The above Lemma shows first, if $\hat{\eta}_n = \eta$, then the plug-in classifier $\hat{g}_n$ is the Bayes optimal classifier. Second, if $\hat{\eta} \approx \eta$, then $\hat{g}_n$ is close to g^*. Therefore, if we could build from the data an estimator $\hat{\eta}_n$ of η such that for all $x \in \mathbb{R}^d$

$$\mathbb{E}_{(X_i,Y_i)\sim\nu}\big[|\eta(x) - \hat{\eta}_n(x)|\big] \underset{n\to+\infty}{\longrightarrow} 0$$

then the associated plugin classifier $\hat{g}_n$ would be consistent (see Definition 1). The reverse if not true: estimating η is harder then estimating g^*. We will show that this is the case for the k-nearest neighbors if the number of neighbors grows appropriately. This is not the case for fixed numbers of neighbors.

5 The k-nearest neighbors classifier (kNN)

The kNN classifiers classifies a new input x with the majority class among its k-nearest neighbors (see Figure 1). More formally, we denote by $X_{(i)}(x)$ the i-th nearest neighbor of $x \in \mathbb{R}^d$ (using the Euclidean distance) among the inputs X_i, $1 \leq i \leq n$. We have for all $x \in \mathbb{R}^d$

$$\big\|x - X_{(1)}(x)\big\| \leq \big\|x - X_{(2)}(x)\big\| \leq \cdots \leq \big\|x - X_{(n)}(x)\big\|$$

and $X_{(i)}(x) \in \{X_1, \ldots, X_n\}$ for all $1 \leq i \leq n$. We denote by $Y_i(x)$ the class associated with $X_i(x)$. We can then define

$$\hat{\eta}_n^k(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) = \frac{1}{k}\sum_{i=1}^{n} Y_{(i)}\mathbb{1}_{X_i\in\{X_{(1)}(x),\ldots,X_{(k)}k(x)\}}$$

and $\hat{g}_n^k$ the kNN classifier is the plugin estimator defined in (6). We denote by

$$R_{kNN} := \lim_{n\to\infty} \mathbb{E}_{(X_i,Y_i)\sim\nu}\big[R(\hat{g}_n^k)\big]$$

the asymptic risk of the k-nearest neighbor classifier.

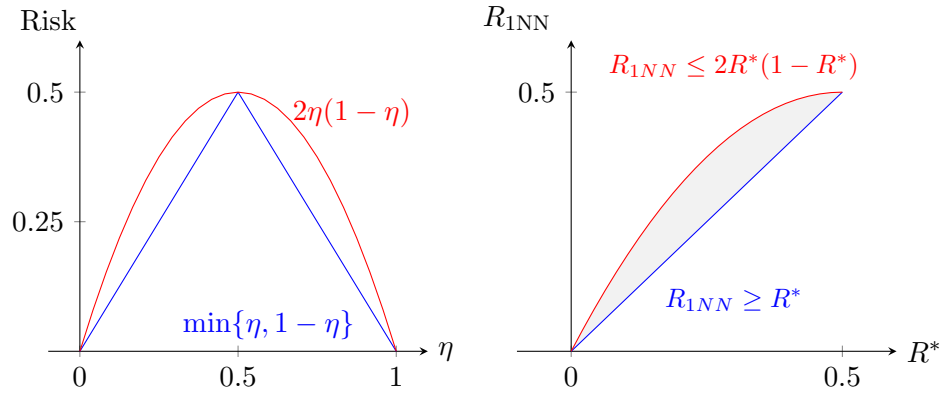

Figure 2: [left] Risk of the 1-nearest neighbor and optimal risk according to η. [right] The risk of the 1-nearest neighbor lies in the dotted area in-between the blue curve (optimal risk) and the red curve (upper-bound of Theorem 3).

5.1 The nearest neighbor classifier

Theorem 3 (Inconsistency of the 1-nearest neighbor)**.** *The asymptotic risk of the 1-nearest neighbor satisfies* $R^* \leq R_{kNN} = \mathbb{E}\big[2\eta(X)(1-\eta(X))\big] \leq 2R^*(1-R^*)$.

Sketch of proof of Theorem 3. We do not provide the complete proof here but only a sketch with the main idea. We refer the curious reader to [3] for the rigorous argument. Let $(X,Y) \sim \nu$ be some new input. From (5), knowing X the label Y follows a Bernoulli distribution with parameter $\eta(X)$. When the number n of data points increases the nearest neighbor of X gets closer to X (this has to be made rigorous since X is a random variable). Thus by continuity of η, given X when $n \to \infty$, we also have $Y_{(1)}(X) \sim \mathcal{B}(\eta(X))$. Therefore,

$$\lim_{n\to\infty} \mathbb{E}_{(X_i,Y_i)\sim\nu}\big[R(\hat{g}_n^1)\big] = \mathbb{P}\big\{Y_{(1)}(X) \neq Y\big\}$$

where $Y_{(1)}(X), Y \sim \mathcal{B}(\eta(X))$ are independent given X. The probability of error is thus

$$\begin{aligned}
\mathbb{P}\big\{Y_{(1)}(X) \neq Y\big\} &= \mathbb{E}_{(X,Y)\sim\nu}\big[\mathbb{P}\big\{Y_{(1)}(X) \neq Y \big| X\big\}\big] \\
&= \mathbb{E}_{(X,Y)\sim\nu}\big[P\{Y=1, Y_{(1)}(X) \neq 1|X\} + \mathbb{P}\big\{Y \neq 1, Y_{(1)}(x) = 1|X\big\}\big] \\
&= \mathbb{E}_{(X,Y)\sim\nu}\big[\mathbb{P}\{Y=1|X\}\mathbb{P}\big\{Y_{(1)}(x) \neq 1|X\big\} + \mathbb{P}\big\{Y \neq 1|X\big\}\mathbb{P}\{Y_{(1)}(x) = 1|X\big\}\big] \\
&= \mathbb{E}_{(X,Y)\sim\nu}\big[2\eta(X)\big(1-\eta(X)\big)\big]\,.
\end{aligned}$$

This concludes the first equality of the Theorem. As for the second, denoting $R(X) := \min\{\eta(X), 1-\eta(X)\}$, we have

$$\mathbb{E}\big[\eta(X)\big(1-\eta(X)\big)\big] = \mathbb{E}\big[R(X)(1-R(X)\big] \overset{\text{Concavity}}{\leq} \mathbb{E}[R(X)]\big(1-\mathbb{E}[R(X)]\big) = R^*(1-R^*)\,.$$

□

The 1-nearest neighbor is therefore not consistent as shown in Figure 2 as soon as the optimal risk is not trivial: $R^* \notin \{0, 1/2\}$. This result was first proved by [2] with assumptions on ν and η and by [4] without any assumption. It is worth to stress that this result is completely distribution free (independent of ν and η). The smoothness of ν and η does not matter for the limit, it only changes the rate of convergence.

5.2 Inconsistency of the k-NN classifier (fixed k)

Therefore, a single neighbor is not sufficient to approach the optimal risk R^*. Actually, we could prove a similar result for any fixed number of neighbors. It is convenient to let k be odd to avoid ties. We refer to [3] for the proof.

Theorem 4. *Let $k \geq 1$ be odd and fixed. Then, the asymptotic risk of the k-nearest neighbor satisfies*

$$\begin{aligned} R_{kNN} &= \mathbb{E}_X \left[\sum_{j=0}^{k} \binom{k}{j} \eta(X)^j (1-\eta(X))^{k-j} \big(\eta(X) \mathbb{1}_{j<k/2} + (1-\eta(X)) \mathbb{1}_{j>k/2} \big) \right] \\ &= R^* + \mathbb{E}\Big[\big| 2\eta(X) - 1 \big| \mathbb{P}\Big\{ \text{Binomial}(k, \min\{\eta(X), 1-\eta(X)\}) > \frac{k}{2} \Big| X \Big\} \Big] . \end{aligned}$$

Sketch of proof of Theorem 4. Similarly to Theorem 3, we only provide an idea of the proof. Let $(X, Y) \sim \nu$ be a new data point. When the number of data goes to infinity, the nearest neighbors $X_{(1)}(X), \ldots, X_{(k)}(X)$ of X get closer to X (to be proved rigorously) and given X their labels $Y_{(1)}(X), \ldots, Y_{(k)}(X)$ are i.i.d. Bernoulli random variables with parameter $\eta(X)$. The k-NN classifier predicts

$$\hat{g}_n^k = \begin{cases} 1 & \text{if} \quad Y_{(1)}(X) + \cdots + Y_{(k)}(X) > \frac{k}{2} \\ 0 & \text{if} \quad Y_{(1)}(X) + \cdots + Y_{(k)}(X) < \frac{k}{2} \end{cases} .$$

The asymptotic probability of error of the k-NN classifier is thus

$$\begin{aligned} R_{kNN} &= \lim_{n \to \infty} \mathbb{P}\big\{ \hat{g}_n^k \neq Y \big\} \\ &= \mathbb{P}\left\{ Y_{(1)}(X) + \cdots + Y_{(k)}(X) < \frac{k}{2}, Y = 1 \right\} + \mathbb{P}\left\{ Y_{(1)}(X) + \cdots + Y_{(k)}(X) > \frac{k}{2}, Y = 0 \right\} \\ &= \mathbb{E}_X \Bigg[\underbrace{\mathbb{P}\{Y=1|X\}}_{\eta(X)} \mathbb{P}\Big\{ \underbrace{Y_{(1)}(X) + \cdots + Y_{(k)}(X)}_{\text{Binomial}(k,\eta(X))} > \frac{k}{2} | X \Big\} \\ &\qquad + \underbrace{\mathbb{P}\{Y=0|X\}}_{1-\eta(X)} \mathbb{P}\Big\{ \underbrace{Y_{(1)}(X) + \cdots + Y_{(k)}(X)}_{\text{Binomial}(k,\eta(X))} < \frac{k}{2} \Big\} \Bigg] , \end{aligned}$$

where given X, $Y_{(1)}(X) + \cdots + Y_{(k)}(X), Y$ are i.i.d. independent Bernoulli random variables with parameter $\eta(X)$. This proves the first equality.

$$R_{kNN} = \mathbb{E}_X \big[\alpha(\eta(X)) \big]$$

where

$$\alpha(p) := p\, \mathbb{P}\Big\{ \text{Binomial}(k,p) < \frac{k}{2} \Big\} + (1-p)\, \mathbb{P}\Big\{ \text{Binomial}(k,p) > \frac{k}{2} \Big\} .$$

If $p < 1/2$, then $p < 1 - p$ and

$$\begin{aligned}\alpha(p) &= p\left(1 - \mathbb{P}\Big\{\text{Binomial}(k,p) > \frac{k}{2}\Big\}\right) + (1-p)\,\mathbb{P}\Big\{\text{Binomial}(k,p) > \frac{k}{2}\Big\} \\ &= p + (1-2p)\,\mathbb{P}\Big\{\text{Binomial}(k,p) > \frac{k}{2}\Big\}\,.\end{aligned}$$

Following the same calculation for $p > 1/2$ yields

$$\alpha(p) = \min\{p, 1-p\} + |2p-1|\mathbb{P}\Big\{\text{Binomial}\big(k, \min\{p,1-p\}\big) > \frac{k}{2}\Big\}$$

which concludes the proof using that $R^* = \mathbb{E}_X\big[\min\{\eta(X), 1-\eta(X)\}\big]$. □

The previous Theorem may provide nice inequalities on R_{kNN} as shown by the next corollary.

Corollary 1. *We have $R^* \le \cdots \le R_{5NN} \le R_{3NN} \le R_{1NN} \le 2R^*(1-R^*)$. Furthermore, let $k \ge 1$ be odd and fixed. Then, the asymptotic risk of the k-NN classifier satisfies*

$$R_{kNN} \le R^* + \frac{1}{\sqrt{ke}}\,.$$

Proof. The first inequalities are because $\mathbb{P}\Big\{\text{Binomial}\big(k,p\big) > \frac{k}{2}\Big\}$ decreases in k for $p < 1/2$. Let $0 \le p \le 1/2$ and $B \sim \text{Binomial}(k,p)$. Then,

$$\begin{aligned}(1-2p)\,\mathbb{P}\Big\{B > \frac{k}{2}\Big\} &= (1-2p)\,\mathbb{P}\left\{\frac{B-kp}{k} > \frac{1}{2} - p\right\} \\ &\overset{(*)}{\le} (1-2p)e^{-2k(1/2-p)^2} \\ &\le \sup_{0\le u\le 1} u e^{-ku^2/2} \\ &= \frac{1}{\sqrt{ke}}\,,\end{aligned}$$

where $(*)$ is by the Okamoto-Hoeffding inequality that we recall below (see [3, Thm 8.1]).

Lemma 6 (Okamoto-Hoeffding inequality)**.** *Let $X_1, \ldots, X_n$ be independant bounded random variables such that $X_i \in [a_i, b_i]$ almost surely. Then, for all $\varepsilon > 0$*

$$\mathbb{P}\left\{S_n - \mathbb{E}[S_n] \ge \varepsilon\right\} \le e^{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}\,,$$

where $S_n = \sum_{i=1}^n X_i$.

□

Therefore the asymptotic error of the k-NN classifier decreases with k but is not consistent: for any fixed k, it does not converge to the optimal risk R^*. The idea is thus to make $k \to \infty$ with n.

5.3 Consistent nearest neighbors making $k \to \infty$

Theorem 5 (Stone 1964)**.** *If $k(n) \to \infty$ and $\frac{k(n)}{n} \to 0$ then the $k(n)$-NN classifier is universally consistent: for all distribution ν, we have*

$$R_{k(n)NN} := \lim_{n\to\infty} \mathbb{E}_{(X_i,Y_i)\sim\nu}\big[R(\hat{g}_n^k)\big] = R^* \,.$$

Historically, this is the first universally consistent algorithm. The proof is not trivial and comes from a more general result (Stone's Theorem) on "Weighted Average Plug-in" classifiers (WAP).

Definition 2 (Weighted Average Plug-in classifier (WAP))**.** *Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, a WAP classifier is a plug-in estimator $\hat{g}_n$ associated to an estimator of the form*

$$\hat{\eta}_n(x) = \sum_{i=1}^n w_{n,i}(x) Y_i$$

where the weights $w_{n,i}(x) = w_{n,i}(x, X_1, \ldots, X_n)$ are non negative and sum to one.

This is the case of the k-NN classifier which satisfies

$$w_{n,i}(x) = \begin{cases} \frac{1}{k} & \text{if } X_i \text{ is a } k\text{NN of } x \\ 0 & \text{otherwise} \end{cases} \,.$$

Theorem 6 (Stone 1977)**.** *Let $(g_n)_{n\geq 0}$ a WAP such that for all distribution ν the weights $w_{n,i}$ satisfy*

a) *it exists $c > 0$ s.t. for all non-negative measurable function f with $\mathbb{E}[f(X)] < \infty$,*

$$\mathbb{E}\left[\sum_{i=1}^n w_{n,i}(X) f(X_i)\right] \leq c\mathbb{E}\big[f(X)\big] \,;$$

b) *for all $a > 0$, $\mathbb{E}\left[\sum_{i=1}^n w_{n,i}(X) \mathbb{1}_{\|X_i - X\| > a}\right] \underset{n\to+\infty}{\longrightarrow} 0$*

c) $\mathbb{E}\big[\max_{1\leq i\leq n} w_{n,i}(X)\big] \underset{n\to+\infty}{\longrightarrow} 0$

Let us make some remarks about the conditions:

a) is a technical condition

b) says that the weights of points outside of a ball around X should vanish to zero. Only the X_i in a smaller and smaller neighborhood of X should contribute.

c) says that no point should have a too important weight. The number of points in the local neighborhood of X should increase to ∞.

Proof of Theorem 6. From Lemma 5 together with Cauchy-Schwarz, it suffices to show that $\mathbb{E}\big[(\eta(X) - \hat{\eta}_n(X))^2\big] \underset{n\to+\infty}{\longrightarrow} 0$. Let us introduce

$$\tilde{\eta}_n(x) := \sum_{i=1}^n w_{n,i}(x) \underbrace{\eta(X_i)}_{\text{instead of } Y_i \text{ in } \hat{\eta}_n}$$

in which we replaced Y_i in $\hat{\eta}_n$ with $\eta(X_i)$ which we recall:

$$\hat{\eta}_n(x) = \sum_{i=1}^n w_{n,i}(x) Y_i \qquad \text{and} \qquad \eta(x) = \sum_{i=1}^n w_{n,i}(x)\eta(x)\,.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\mathbb{E}\big[(\eta(X) - \hat{\eta}_n(X))^2\big] \leq 2\underbrace{\mathbb{E}\big[(\eta(X) - \tilde{\eta}_n(X))^2\big]}_{(1)} + 2\underbrace{\mathbb{E}\big[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2\big]}_{(2)}\,.$$

We will upper-bound (1) and (2) independently.

(1) For simplicity, to bound this term we assume η to be absolutely continuous: let $\varepsilon > 0$, it exists $a > 0$ such that $\|x - x'\| \leq a \Rightarrow (\eta(x) - \eta(x'))^2 \leq \varepsilon$. Then,

$$\begin{aligned}
(1) &= \mathbb{E}\bigg[\Big(\sum_{i=1}^n w_{n,i}(X)\big(\eta(X) - \eta(X_i)\big)\Big)^2\bigg] \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}\bigg[\sum_{i=1}^n w_{n,i}(X)\big(\eta(X) - \eta(X_i)\big)^2\bigg] \\
&= \mathbb{E}\bigg[\sum_{i=1}^n w_{n,i}(X)\big(\eta(X) - \eta(X_i)\big)^2 \mathbb{1}_{\|X_i - X\| \leq \varepsilon}\bigg] + \mathbb{E}\bigg[\sum_{i=1}^n w_{n,i}(X)\big(\eta(X) - \eta(X_i)\big)^2 \mathbb{1}_{\|X_i - X\| \geq \varepsilon}\bigg] \\
&\leq \varepsilon + \mathbb{E}\bigg[\sum_{i=1}^n w_{n,i}(X)\big(\eta(X) - \eta(X_i)\big)^2 \mathbb{1}_{\|X_i - X\| \geq \varepsilon}\bigg] \\
&\leq \varepsilon + \underbrace{\mathbb{E}\bigg[\sum_{i=1}^n w_{n,i}(X) \mathbb{1}_{\|X_i - X\| \geq \varepsilon}\bigg]}_{\underset{n\to+\infty}{\longrightarrow} 0 \quad \text{from Assumption (b)}}\,.
\end{aligned}$$

Therefore (1) converges to 0 as $n \to \infty$. If η is not absolutely continuous, the result still holds using Assumption (a) but the proof is harder (see [3], p99).

(2) For the second term, using that $\mathbb{E}[\eta(X_i)] = Y_i$, only the diagonal terms in the sum remain

$$
\begin{aligned}
(2) &= \mathbb{E}\Big[\Big(\sum_{i=1}^n w_{n,i}(X)\big(Y_i - \eta(X_i)\big)\Big)^2\Big] \\
&= \sum_{i=1}^n \sum_{j\neq i} w_{n,i}(X) w_{n,j}(X) \underbrace{\mathbb{E}\Big[\big(Y_i - \eta(X_i)\big)\big(Y_j - \eta(X_j)\big)\Big]}_{=0} + \mathbb{E}\Big[\sum_{i=1}^n w_{n,i}(X)^2\big(Y_i - \eta(X_i)\big)^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^n w_{n,i}(X)^2\big(Y_i - \eta(X_i)\big)^2\Big] \\
&\leq \mathbb{E}\Big[\sum_{i=1}^n w_{n,i}(X)^2\Big] \\
&\leq \mathbb{E}\Big[\underbrace{\sum_{i=1}^n w_{n,i}(X)}_{=1} \max_{1\leq j\leq n} w_{n,i}(X)\Big] \\
&\leq \mathbb{E}\Big[\max_{1\leq j\leq n} w_{n,i}(X)\Big] \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{from Assumption (c)}\,.
\end{aligned}
$$

□

Let us now conclude with the proof of the consistency of the k nearest neighbors when $k \to \infty$.

Proof of Theorem 5. First, we recall the definition of the weights $w_{n,i}(x)$ for the kNN classifier:

$$
w_{n,i}(x) = \frac{\mathbb{1}_{X_i \in X_{(1)}(x),\ldots,X_{(k)}(x)}}{k} = \begin{cases} \frac{1}{k} & \text{if } X_i \text{ belong to the } k \text{ nearest neighbors of } x \\ 0 & \text{otherwise} \end{cases}.
$$

It suffices to show that they satisfy the three assumptions of Theorem 6 (Stone's theorem):

c) for all x, $\max_{1\leq i\leq n} w_{n,i}(x) = \frac{1}{k(n)} \underset{n\to+\infty}{\longrightarrow} 0$ so that assumption (c) holds.

b) let $a > 0$, recall that $X_{(k)}(x)$ is the k-th nearest neighbor of x. We use that almost surely the distance of the k-nearest neighbor of X with X goes to zero when $k/n \to 0$:$\|X - X_{(k)}\| \underset{n\to+\infty}{\longrightarrow} 0$ when $\frac{k}{n} \to 0$ (see [3] for details). This yields $\mathbb{P}\{\|X - X_{(k)}(X)\| > a\} \to 0$ which entails

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^n w_{n,i}(X)\mathbb{1}_{\|X_i - X\|>a}\Big] \\
&\qquad \leq \mathbb{E}\left[\sum_{i=1}^n w_{n,i}(X)\mathbb{1}_{\|X_i - X\|>a}\mathbb{1}_{\|X_i - X_{(k)}(X)\|>a}\right] + \mathbb{E}\left[\sum_{i=1}^n w_{n,i}(X)\mathbb{1}_{\|X_i - X\|>a}\mathbb{1}_{\|X_i - X_{(k)}(X)\|<a}\right] \\
&\qquad \leq 0 + \mathbb{P}\big\{\|X_i - X_{(k)}(X)\| < a\big\} \to 0
\end{aligned}
$$

a) Technical. See [3], Lemma 5.3.

□

Conclusion

The k-nearest neighbors are universally consistent if $k \to \infty$ and $k/n \to 0$. Stone's theorem is actually more general and applies to other rules such as histograms.

References

[1] D Aliprantis Charalambos and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

[2] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[3] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[4] Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.