# Machine learning - Master ICFP 2019-2020

## Local averaging methods

Francis Bach

January 31, 2020

These notes are based on notes from Alessandro Rudi and Pierre Gaillard.

## 1 Introduction - review

- Training set: observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, of inputs/outputs, features/variables are independent and identically distributed (i.i.d.) random variables with common distribution $P$.
- $\mathcal{X}$ can de diverse, $\mathcal{Y}$ is typically $\{0, 1\}$ (binary classification) or $\mathbb{R}$ (regression).
- We consider a fixed (testing) distribution $P$ on $\mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$; $\ell(y, z)$ is the loss of predicting $z$ while the true label is $y$. **We assume that the testing distribution is the same as the training distribution.**
- Risk, or generalization performance of a prediction function $f : \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{R}(f) = \mathbb{E}\Big[\ell(Y, f(X))\Big].$$

  Be careful with the randomness or lack thereof of $f$: $\hat{f}_n$ depends on the training data and not on the testing data, and thus $\mathcal{R}(\hat{f}_n)$ is random because of the dependence on the training data $\mathcal{D}_n$.

  The function $\mathcal{R}$ depends on the distribution $P$ on $(X, Y)$. We sometimes use the notation $\mathcal{R}_P(f)$ to make it explicit.

  - Binary classification: $\mathcal{Y} = \{0, 1\}$ (or often $\mathcal{Y} = \{-1, 1\}$), and $\ell(y, z) = 1_{y \neq z}$ ("0-1" loss). Then $\mathcal{R}(f) = \mathbb{P}(f(X) \neq Y)$.
  - Regression: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$ (square loss). Then $\mathcal{R}(f) = \mathbb{E}(Y - f(X))^2$.
- Target function = Bayes predictor $f^* \in \arg\min \mathcal{R}(f) = \mathbb{E}\Big[\ell(Y, f(X))\Big]$.

  **Proposition 1 (Bayes predictor)** *The risk is minimized at a* Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ *satisfying for all* $x \in \mathcal{X}$, $f^*(x) \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}(\ell(Y, z)|X = x)$. *The Bayes risk* $\mathcal{R}^*$ *is the risk of all Bayes predictors and is equal to*

$$\mathcal{R}^* = \mathbb{E}_{x \sim P_X} \inf_{z \in \mathcal{Y}} \mathbb{E}(\ell(Y, z)|X = x).$$

Note that (a) the Bayes predictor is not unique, but that all lead to the same Bayes risk, and (b) that the Bayes risk is usually non zero (unless the dependence between $x$ and $y$ is deterministic).

  - For binary classification: $\mathcal{Y} = \{0, 1\}$ and $\ell(y, z) = 1_{y \neq z}$, the Bayes predictor is $f^*(x) \in \arg\max_{i \in \{1,\dots,k\}} \mathbb{P}(Y = i | X = x)$.
  - For regression: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$, the Bayes predictor is $f^*(x) = \mathbb{E}(Y|X = x)$.

- Goal of supervised machine learning: estimate $f^*$, knowing only the data $\mathcal{D}_n$ *and* the loss $\ell$.

**Definition 1 (Excess risk)** *The excess risk of a function from $f : \mathcal{X} \to \mathcal{Y}$ is equal to $\mathcal{R}(f) - \mathcal{R}^*$ (it is always non-negative).*

## 2 Local averaging methods

- In local averaging methods, we aim at approximating the target function directly without any form of optimization. This will be done by approximating the conditional distribution $P(y|X = x) = P(y|x)$.
- We then replace $f^*(x) \in \arg\min_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \ell(y, z) dP(y|x)$ by $\hat{f}(x) \in \arg\min_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \ell(y, z) d\widehat{P}(y|x)$.

To study the excess risk for this estimator we perform the following analysis. Denote by $M(x, z)$ the function $M(x, z) = \int \ell(y, z) dP(y|x)$, that is, the true conditional loss of predicting $z$ at $x$, and by $\widehat{M}(x, z)$ the function $\widehat{M}(x, z) = \int \ell(y, z) d\widehat{P}(y|x)$, its approximation, then

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

Now note that

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

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

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

So finally

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

Note the similarity with the estimation error in the previous lecture. We have:

$$M(x, z) - \widehat{M}(x, z) = \int \ell(y, z)(dP(y|x) - d\widehat{P}(y|x)).$$

So we need to show that $dP(y|x) - d\widehat{P}(y|x)$ is small.

We are going to see two main types of estimators of $P(y|x)$:

- Nadaraya-Watson estimators
- Nearest-neighbors estimators

# 3 Kernel estimation

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

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

Usually estimators for the conditional probability have the following form

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

where $\widehat{P}(y,x)$ and $\widehat{P}(x)$ are estimators for $P(y,x)$ and $P(x)$. 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., $q(x) = \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2}$) and $q_\tau(x) = \tau^{-d} q(x/\tau)$, for $\tau > 0$, a scaled version. We will typically choose $\tau$ tending to zero.

Let moreover $x_1, \ldots, x_n$ be sampled i.i.d. from $P$. We define the estimator as

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

By denoting by $\widehat{P}_n$ the probability measure $\widehat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ (where $\delta$ 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

$$P \;\approx\; P \star q_\tau \;\approx\; \widehat{P} \star q_\tau \;=\; \widehat{P}_n,$$

which we now make precise. In particular

**Lemma 1 (Bias)** *Assume $|P(x) - P(x')| \leq C\|x - x'\|$ for any $x, x'$, then*

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

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

**Proof** Since $\int q_\tau(x - x') dx' = \int q_\tau(x') dx' = 1$, we have

$$|P(x) - (P \star q_\tau)(x)| = |\tau^{-d} \int (P(x) - P(x')) q((x-x')/\tau) dx'| \leq \tau^{-d} \int |P(x) - P(x')| q((x-x')/\tau) dx'$$
$$\leq C\tau^{-d+1} \int \|x - x'\|/\tau \; q((x-x')/\tau) dx' = C\tau^{-d+1} \int \|u/\tau\| q(u/\tau) du = C\tau \int \|z\| q(z) dz,$$

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 (Variance)** *For any $v \in X$, we have*

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

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

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

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

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 $P$, 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

$$\mathbb{E}(z - \mathbb{E}z)^2 \leq \mathbb{E}z^2 = \int q_\tau(v - x)^2 P(x)dx \leq (\max_t q_\tau(t))\|P\|_\infty \int q_\tau(v - x)dx = \|P\|_\infty \max_t q_\tau(t),$$

which is equal to $\tau^{-d}\|P\|_\infty \max_t q(t)$. ■

Finally, we can combine the two elements to get:

**Theorem 1** *Let $P$ such that $|P(x) - P(x')| \leq C\|x - x'\|$, then for any $v \in \mathcal{X}$*

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

**Proof** The result is obtained combining the two lemmas above. ■

Thus, in order to balance the two terms, we need $\tau^2 \sim \tau^{-d}/n$, thus $\tau \sim n^{-1/(d+2)}$, with an overall convergence rate less than $n^{-1/(d+2)}$. We thus see the *curse of dimensionality.*

Note that this is only a bound for a single $v$, and that extra work is needed to get uniform guarantees.

The estimator for $P(x, y)$ 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 $\mathcal{X}$ and $p$ the dimension of the space containing $\mathcal{Y}$. We now give examples where $\mathcal{Y} \subset \mathbb{R}$.

### 3.2 Nadaraya-Watson estimator

We consider the Gaussian kernel $q(x) = \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2}$, where we use the fact it is non-negative pointwise (as opposed to positive-definiteness in later lectures, see `https://francisbach.com/cursed-kernels/`).

**Regression.** We have

$$\widehat{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} q_\tau(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\tau^d} \exp(-\frac{\|x - x_i\|^2}{2\tau^2}),$$

and, for $\mathcal{Y} = \mathbb{R}$, with $q^1_\tau(y - y') = \frac{1}{\sqrt{2\pi}\tau} e^{-(y-y')^2/2}$ a kernel on $\mathbb{R}$:

$$\widehat{P}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} q_\tau(x - x_i) q^1_\tau(y - y_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\tau^d} \exp(-\frac{\|x - x_i\|^2}{2\tau^2}) \frac{1}{(2\pi)^{1/2}\tau} \exp(-\frac{(y - y_i)^2}{2\tau^2}).$$

Thus

$$\widehat{P}_n(y|x) = \sum_{i=1}^{n} w_i(x) \frac{1}{(2\pi)^{1/2}\tau} \exp(-\frac{(y - y_i)^2}{2\tau^2})$$

with

$$w_i(x) = \frac{\exp(-\frac{\|x-x_i\|^2}{2\tau^2})}{\sum_{j=1}^{n} \exp(-\frac{\|x-x_j\|^2}{2\tau^2})}.$$

This is a mixture model and $\hat{f}(x)$ is the conditional expectation, equal to $\sum_{i=1}^{n} w_i(x) y_i$ .

**Classification.** We can define the estimator accordingly, and we get $\hat{f}(x) = 1$ if $\sum_{i=1}^{n} w_i(x) y_i > 1/2$, which is often referred to as the "plug-in" estimator.

# 4 $k$-Nearest Neighbours

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 (for classification) or simply the averaged response (for regression). 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. several questions are typically investigated:

- 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 [1].

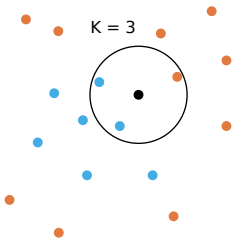


Figure 1: $k$-nearest neighbors with two classes (orange and blue) and $k = 3$. The new input (i.e., the black point) is classified as blue which corresponds to the majority class among its three nearest neighbors.

# 5 The $k$-nearest neighbors classifier (kNN)

The kNN classifier 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 $x \in \mathbb{R}^d$

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

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

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

and $\hat{g}_n^k$ the $k$NN classifier is the corresponding plugin estimator defined by $\hat{g}_n^k(x) = 1$ if $\eta_n^k(x) > 1/2$, and zero otherwise.

This applies to regression as well.

## 5.1 Consistent nearest neighbors making $k \to \infty$

**Theorem 2 (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 $\nu$, we have*

$$\lim_{n\to\infty} \mathcal{R}(\hat{f}_{k(n)NN}) := \mathcal{R}^* .$$

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 $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, 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 3 (Stone 1977)** *Let $(g_n)_{n\geq 0}$ a WAP such that for all distribution $\nu$ the weights $w_{n,i}$ satisfy*

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

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

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

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

Let us make some remarks about the conditions:

1. is a technical condition
2. 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.
3. says that no point should have a too important weight. The number of points in the local neighborhood of $X$ should increase to $\infty$.

### 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 or Nadaraya-Watson estimators.

## References


[1] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1996.

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

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