Mathematical Foundations of Statistical Learning

Christophe Giraud

Université Paris Sud and Ecole Polytechnique

Abstract. The goal of a classification algorithm is to predict at best the class of an object from some observations of this object. A typical example is the spam filter of our mailbox, which predicts (more or less fairly) whether a mail is a spam or not. We introduce in these notes the main basic concepts of the theory of supervised statistical classification and some of the most popular classification algorithms. We highlight along the way the importance of some mathematical tools including symmetrization, convexification, concentration inequalities, contraction and reproducing kernel Hilbert spaces.

Contents

1 Introduction

The early 21st century has seen a dramatic increase of the use of mathematics both in private companies and in academic labs. This raise of the importance of mathematics goes in pair with the explosion of data production and computing power. In the industry, mathematical modeling can appear at every stage of the life of a product. From the technical conception, with intensive numerical simulation, via the production, with the optimization of resources and fluxes, to the marketing and the distribution with forecasts based on the analysis of huge data bases. In academic labs, mathematical modeling becomes more and more crucial, in particular in biology and medicine, where scientists have to handle massive data sets produced by the virtue of recent biotechnological developments.

Automatic classification is perhaps one of the most invasive uses of mathematics. The goal of automatic classification is to predict at best the class y of an object x from some observations. A typical example is the spam filter of our mailbox, which predicts (more or less fairly) whether a mail is a spam or not. It is omnipresent in our daily life, by filtering the spams in our mailbox, reading automatically post-code on our letters or recognizing faces on photos that we post in social networks. It is also very important in sciences, e.g. in medicine for early diagnosis of diseases from high-throughput data or for *in silico* exploration of candidate drugs.

We introduce in these notes the main basic concepts of (supervised) statistical learning. We describe in Section 2 the mathematical modeling of a generic classification problem. In Section 3 we analyze the prediction accuracy of a universal classification algorithm and in Section 4 we derive from this theoretical algorithm some practical and popular algorithms. The appendices gather some technical results involved in the definition and analysis of the classification algorithm.

2 Mathematical modeling

For the sake of simplicity, we will restrict in these notes to the case where we have only two classes (as for the spam filter). The problem of automatic classification can be modeled as follows. Let $\mathcal{X}$ be some measurable space. We observe conjointly a data point $X \in \mathcal{X}$ and a label $Y \in \{-1+1\}$. Our aim is to find a function $h : \mathcal{X} \to \{-1, +1\}$, called *classifier*, such that $h(X)$ predicts at best the label Y.

Assume that the couple $(X, Y) \in \mathcal{X} \times \{-1+1\}$ is sampled from a distribution $\mathbb{P}$. For a classifier $h : \mathcal{X} \to \{-1, +1\}$ the probability of misclassification is

$$L(h) = \mathbb{P}(Y \neq h(X)).$$

Since $|Y - h(X)| \in \{0, 2\}$, we have

$$L(h) = \frac{1}{4}\,\mathbb{E}\left[(Y - h(X))^2\right] = \frac{1}{4}\,\mathbb{E}\left[(Y - \mathbb{E}\left[Y|X\right])^2\right] + \frac{1}{4}\,\mathbb{E}\left[(\mathbb{E}\left[Y|X\right] - h(X))^2\right].$$

Therefore $L(h)$ is minimal for the so-called Bayes classifier

$$h_*(X) = \text{sign}(\mathbb{E}[Y|X]) \quad \text{where} \quad \text{sign(x)} = \mathbf{1}_{x>0} - \mathbf{1}_{x\leq 0} \quad \text{for } x \in \mathbb{R}.$$

When the distribution $\mathbb{P}$ is known, we simply use the Bayes classifier h_* in order to have the smallest possible probability of misclassification. Unfortunately, the distribution $\mathbb{P}$ is usually unknown, so we cannot compute the Bayes classifier h_*.

In practice, we only have access to some training data $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. with distribution $\mathbb{P}$ and our goal is to build from this training data a classifier $\widehat{h} : \mathcal{X} \to \{-1, +1\}$ such that $L(\widehat{h}) - L(h_*)$ is as small as possible.

3 Empirical risk minimization

Since $\mathbb{P}$ is unknown we cannot compute the probability of misclassification $L(h)$. We can compute instead the empirical probability of misclassification

$$\widehat{L}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} = \widehat{\mathbb{P}}_n(Y \neq h(X)),$$

where $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$. To a set $\mathcal{H}$ of classifiers, we call *dictionary*, we can associate the so-called empirical risk minimization classifier

$$\widehat{h}_{\mathcal{H}} \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{L}_n(h). \tag{1}$$

The definition of this classifier is natural, yet we face two issues: which dictionary $\mathcal{H}$ should be chosen and how does $\widehat{h}_{\mathcal{H}}$ behave compared to h_*? These two issues are of course strongly connected. Decomposing the difference between the misclassification probabilities $L(\widehat{h}_{\mathcal{H}})$ and $L(h_*)$, we find

$$0 \;\leq\; L(\widehat{h}_{\mathcal{H}}) - L(h_*) \;=\; \underbrace{\min_{h\in\mathcal{H}} L(h) - L(h_*)}_{\text{approximation error}} + \underbrace{L(\widehat{h}_{\mathcal{H}}) - \min_{h\in\mathcal{H}} L(h)}_{\text{stochastic error}}.$$

The first term measures the quality of the approximation of h_* by some classifier $h \in \mathcal{H}$. This approximation error is purely deterministic and enlarging the dictionary $\mathcal{H}$ can only reduce it. The second term measures the error made by minimizing over $h \in \mathcal{H}$ the empirical misclassification probability $\widehat{L}_n(h)$ instead of the true misclassification probability $L(h)$. This term is stochastic and it tends to increase when $\mathcal{H}$ increases. This phenomenon is illustrated in Figure 1. In this illustration in $\mathcal{X} = \mathbb{R}^2$, the classifiers of the dictionary $\mathcal{H}_{\mathsf{lin}} = \{h(x) = \text{sign}(\langle w, x\rangle) : \|w\| = 1\}$ are not flexible enough and they produce poor classification. In this case the approximation error is large. On the other hand the classifiers of the dictionary $\mathcal{H}_{\mathsf{poly}} = \{h(x) = 2\,\mathbf{1}_A(x) - 1 : A \text{ polygon in } \mathcal{X}\}$ are very flexible and can always reproduce exactly the classification of the data $(X_i, Y_i)_{i=1,\dots,n}$. The empirical error

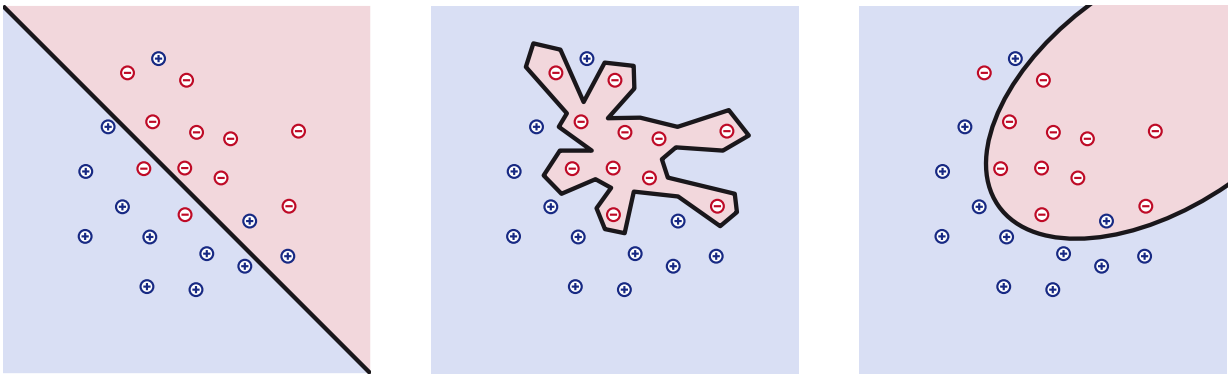

Figure 1: Examples of classification produced by different dictionaries. Left: with the linear classifiers $\mathcal{H}_{\mathsf{lin}}$. Center: with the polygon classifiers $\mathcal{H}_{\mathsf{poly}}$. Right: with classifiers based on quadratic forms.

$\widehat{L}_n(\widehat{h}_{\mathcal{H}_{\mathsf{poly}}})$ is then always 0, but $\widehat{h}_{\mathcal{H}_{\mathsf{poly}}}$ tends to produce poor classification of new data (X, Y) and the stochastic term $L(\widehat{h}_{\mathcal{H}}) - \min_{h\in\mathcal{H}} L(h)$ is large. The last example, based on a less flexible set of quadratic classifiers produces a better result.

To choose a good dictionary $\mathcal{H}$, we shall then find a good balance between the approximation properties of $\mathcal{H}$ and its size. The first step towards a principle for choosing the dictionary $\mathcal{H}$ is to assess the misclassification probability of the empirical risk minimizer $\widehat{h}_{\mathcal{H}}$.

3.1 Misclassification probability of $\widehat{h}_{\mathcal{H}}$

As mentioned above, increasing the size of $\mathcal{H}$ tends to increase the stochastic error $L(\widehat{h}_{\mathcal{H}}) - \min_{h\in\mathcal{H}} L(h)$. Actually, it is not really the size of the dictionary that matters, but rather its flexibility in terms of classification. For example, we cannot classify correctly the three labeled points $\{((0,1),+1),((1,1),-1),((1,0),+1)\}$ with a classifier in $\mathcal{H}_{\mathsf{lin}}$. Conversely, for any set of labeled points $(x_i, y_i)_{i=1,\dots,n}$, there exists $h \in \mathcal{H}_{\mathsf{poly}}$ such that $h(x_i) = y_i$.

In order to capture this classification flexibility, we introduce the shattering coefficient

$$\mathbb{S}_n(\mathcal{H}) = \max_{(x_1,\dots,x_n)\in\mathcal{X}^n} \operatorname{card}\{(h(x_1),\dots,h(x_n)) : h \in \mathcal{H}\}\ , \tag{2}$$

which gives the maximal number of different labeling of n points that the classifiers in $\mathcal{H}$ can produce. For example, since n points can be arbitrarily labelled with classifiers in $\mathcal{H}_{\mathsf{poly}}$, we have $\mathbb{S}_n(\mathcal{H}_{\mathsf{poly}}) = 2^n$. On the contrary, the number of possible labeling of n points with classifiers in $\mathcal{H}_{\mathsf{lin}}$ is more limited. Actually Proposition 1 in Section 3.3 ensures that $\mathbb{S}_n(\mathcal{H}_{\mathsf{lin}}) \le (n+1)^2$. Next theorem provides an upper-bound of the stochastic error and a confidence interval for the misclassification probability $L(\widehat{h}_{\mathcal{H}})$ in terms of the shattering coefficient.

Theorem 1 Control of the stochastic error

For any $t > 0$, with probability at least $1 - e^{-t}$ we have

$$L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h) \leq 4 \sqrt{\frac{2 \log(2\, \mathbb{S}_{\mathcal{H}}(n))}{n}} + \sqrt{\frac{2t}{n}} \tag{3}$$

and

$$\left| L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) \right| \leq 2 \sqrt{\frac{2 \log(2\, \mathbb{S}_{\mathcal{H}}(n))}{n}} + \sqrt{\frac{t}{2n}} \tag{4}$$

Proof. Next lemma shows that the left-hand terms in (3) and (4) can be upper bounded in terms of maximum difference over $\mathcal{H}$ between the empirical misclassification probability and the true misclassification probability

$$\widehat{\Delta}_n(\mathcal{H}) = \sup_{h \in \mathcal{H}} \left| \widehat{L}_n(h) - L(h) \right|. \tag{5}$$

Lemma 1.1

We have the upper-bounds

$$L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h) \leq 2\, \widehat{\Delta}_n(\mathcal{H}) \quad \textit{and} \quad \left| L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) \right| \leq \widehat{\Delta}_n(\mathcal{H}).$$

Proof of Lemma 1.1. For any $h \in \mathcal{H}$, we have $\widehat{L}_n(\widehat{h}_{\mathcal{H}}) \leq \widehat{L}_n(h)$ and therefore

$$\begin{aligned} L(\widehat{h}_{\mathcal{H}}) - L(h) &= L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) + \widehat{L}_n(\widehat{h}_{\mathcal{H}}) - L(h) \\ &\leq L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) + \widehat{L}_n(h) - L(h) \\ &\leq 2\, \widehat{\Delta}_n(\mathcal{H}). \end{aligned}$$

Since this inequality is true for any $h \in \mathcal{H}$, the first bound of Lemma 1.1 follows. The second bound is obvious. □

In order to prove Theorem 1, it remains to prove that we have

$$\widehat{\Delta}_n(\mathcal{H}) \leq 2 \sqrt{\frac{2 \log(2\, \mathbb{S}_{\mathcal{H}}(n))}{n}} + \sqrt{\frac{t}{2n}}$$

with probability at least $1 - e^{-t}$. We split the proof of this bound into two lemmas.

Lemma 1.2

With probability at least $1 - e^{-t}$, we have

$$\widehat{\Delta}_n(\mathcal{H}) \leq \mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H}) \right] + \sqrt{\frac{t}{2n}}\ .$$

Proof of Lemma 1.2. We have $\widehat{\Delta}_n(\mathcal{H}) = F((X_1, Y_1), \ldots, (X_n, Y_n))$ with

$$\begin{aligned} F \quad : \quad & (\mathcal{X} \times \{-1, +1\})^n \quad \to \quad \mathbb{R} \\ & ((x_1, y_1), \ldots, (x_n, y_n)) \mapsto \frac{1}{n} \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \mathbf{1}_{y_i \neq h(x_i)} - L(h) \right| . \end{aligned}$$

For any $(x_1, y_1), \ldots, (x_n, y_n), (x_i', y_i') \in \mathcal{X} \times \{-1, +1\}$, we have

$$\left| F((x_1, y_1), \ldots, (x_i', y_i'), \ldots, (x_n, y_n)) - F((x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)) \right| \leq \frac{1}{n} ,$$

so according to McDiarmid concentration inequality (see Theorem 4 in Appendix B), with probability at least $1 - e^{-2ns^2}$, we have $\widehat{\Delta}_n(\mathcal{H}) \leq \mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right] + s$. Lemma 1.2 follows by setting $s = \sqrt{t/(2n)}$. □

It remains to bound the expectation of $\widehat{\Delta}_n(\mathcal{H})$ in terms of $\mathbb{S}_{\mathcal{H}}(n)$.

Lemma 1.3

For any dictionary $\mathcal{H}$ we have the upper-bound

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right] \leq 2\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathcal{H}}(n))}{n}} .$$

Proof of Lemma 1.3. The proof of Lemma 1.3 is based on a classical and elegant symmetrization argument.

The first step of the symmetrization is to represent the misclassification probability $L(h)$ as the expectation of an empirical misclassification probability

$$L(h) = \mathbb{P}(Y \neq h(X)) = \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right] ,$$

where $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ is independent of $(X_i, Y_i)_{i=1,\ldots,n}$ and is identically distributed. In the following, $\widetilde{\mathbb{E}}$ will refer to the expectation with respect to the variables $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ and $\mathbb{E}$ will refer to the expectation with respect to the variables $(X_i, Y_i)_{i=1,\ldots,n}$. According to Jensen's and Fatou's inequalities, we have

$$\begin{aligned} \mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right] &= \mathbb{E}\left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} - \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right] \right|\right] \\ &\leq \mathbb{E}\widetilde{\mathbb{E}}\left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right) \right|\right] . \end{aligned}$$

The second step is to capitalize on the symmetry of the variables $\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}$. We introduce n i.i.d. random variables $(\sigma_i)_{i=1,\ldots,n}$ independent of $(X_i, Y_i, \widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ with distribution

$\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$. By symmetry, we notice that $\left(\sigma_i\big(\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\big)\right)_{i=1,\ldots,n}$ has the same distribution as $\left(\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right)_{i=1,\ldots,n}$, so we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right] &\leq \mathbb{E}\widetilde{\mathbb{E}}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(\mathbf{1}_{Y_i\neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i\neq h(\widetilde{X}_i)}\right)\Big|\right] \\
&\leq 2\mathbb{E}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{Y_i\neq h(X_i)}\Big|\right] \\
&\leq 2\max_{y\in\{-1,+1\}^n}\max_{x\in\mathcal{X}^n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{y_i\neq h(x_i)}\Big|\right],
\end{aligned}$$

where the second inequality simply follows from the triangular inequality. For any $(x, y) \in \mathcal{X}^n \times \{-1, +1\}^n$, let us define the set

$$\mathcal{V}_\mathcal{H}(x, y) = \left\{(\mathbf{1}_{y_1\neq h(x_1)}, \ldots, \mathbf{1}_{y_n\neq h(x_n)}) : \ h \in \mathcal{H}\right\}.$$

The last upper-bound on $\mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right]$ can be written as

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathcal{H})\right] \leq \frac{2}{n} \times \max_{y\in\{-1,+1\}^n}\max_{x\in\mathcal{X}^n}\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}_\mathcal{H}(x,y)}|\langle\sigma, v\rangle|\right],$$

where $\langle x, y\rangle$ is the canonical scalar product on $\mathbb{R}^n$. We notice that for any $y \in \{-1, +1\}^n$ there is a bijection between $\mathcal{V}_\mathcal{H}(x, y)$ and the set $\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$. As a consequence we have the upper-bound

$$\max_{y\in\{-1,+1\}^n}\max_{x\in\mathcal{X}^n}\operatorname{card}(\mathcal{V}_\mathcal{H}(x, y)) \leq \mathbb{S}_n(\mathcal{H}).$$

In view of the last two inequalities, it simply remains to prove that

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}}|\langle\sigma, v\rangle|\right] \leq \sqrt{2n\log\left(2\operatorname{card}(\mathcal{V})\right)}, \quad \text{for any finite} \quad \mathcal{V} \subset \{-1, 0, +1\}^n \tag{6}$$

in order to conclude the proof of Lemma 1.3. Let us prove (6). Writing $\mathcal{V}^\# = \mathcal{V} \cup -\mathcal{V}$, Jensen's inequality ensures that for any $s > 0$

$$\begin{aligned}
\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}}|\langle\sigma, v\rangle|\right] &= \mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}^\#}\langle\sigma, v\rangle\right] \leq \frac{1}{s}\log\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}^\#}e^{s\langle\sigma,v\rangle}\right] \\
&\leq \frac{1}{s}\log\left(\sum_{v\in\mathcal{V}^\#}\mathbb{E}_\sigma\left[e^{s\langle\sigma,v\rangle}\right]\right). \qquad (7)
\end{aligned}$$

Combining the facts that the σ_i are independent, $(e^x + e^{-x}) \leq 2e^{x^2/2}$ for all $x \in \mathbb{R}$ and $v_i^2 \in \{0, 1\}$ for all $v \in \mathcal{V}^\#$, we have

$$\begin{aligned}
\mathbb{E}_\sigma\left[e^{s\langle\sigma,v\rangle}\right] &= \prod_{i=1}^n \mathbb{E}_\sigma\left[e^{sv_i\sigma_i}\right] = \prod_{i=1}^n \frac{1}{2}(e^{sv_i} + e^{-sv_i}) \\
&\leq \prod_{i=1}^n e^{s^2v_i^2/2} \leq e^{ns^2/2}.
\end{aligned}$$

Plugging this inequality in (7), we obtain

$$\mathbb{E}_\sigma \left[\sup_{v \in \mathcal{V}} |\langle \sigma, v \rangle| \right] \leq \frac{\log(\mathrm{card}(\mathcal{V}^\#)}{s} + \frac{ns}{2} \qquad \text{for any } s > 0.$$

The right-hand side is minimal for $s = \sqrt{2\log(\mathrm{card}(\mathcal{V}^\#))/n}$, which gives the upper bound

$$\mathbb{E}_\sigma \left[\sup_{v \in \mathcal{V}} |\langle \sigma, v \rangle| \right] \leq \sqrt{2n \log(\mathrm{card}(\mathcal{V}^\#))} \; .$$

We finally obtain (6) by noticing that $\mathrm{card}(\mathcal{V}^\#) \leq 2\,\mathrm{card}(\mathcal{V})$. The proof of Lemma 1.3 is complete. □

The bounds (3) and (4) are obtained by combining the three lemmas. □

3.2 Dictionary selection

Let us consider a collection $\{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$ of classifiers dictionaries. We would like to select among this collection, the dictionary $\mathcal{H}_*$ with the smallest misclassification probability $L(\widehat{h}_{\mathcal{H}_*})$. The so-called *oracle* dictionary $\mathcal{H}_*$ depends on the unknown distribution $\mathbb{P}$, so it is not accessible to the statistician. In the following, we will build on Theorem 1 in order to design a data-driven procedure for selecting a dictionary $\mathcal{H}_{\widehat{m}}$ among the collection $\{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, with performances similar to those of $\mathcal{H}_*$.

The oracle dictionary $\mathcal{H}_*$ is obtained by minimizing the misclassification probability $L(\widehat{h}_{\mathcal{H}})$ over $\mathcal{H} \in \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$. A first idea is to select $\mathcal{H}_{\widehat{m}}$ by minimizing over the collection $\{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$ the empirical misclassification probability $\widehat{L}_n(\widehat{h}_{\mathcal{H}})$. This selection procedure will not give good results since for any $\mathcal{H} \subset \mathcal{H}'$ we always have $\widehat{L}_n(\widehat{h}_{\mathcal{H}'}) \leq \widehat{L}_n(\widehat{h}_{\mathcal{H}})$, so the procedure will tend to select the largest possible dictionary. For designing a good selection procedure, we have to take into account the fluctuations of $\widehat{L}_n(\widehat{h}_{\mathcal{H}})$ around $L(\widehat{h}_{\mathcal{H}})$. The bound (4) in Theorem 1 give us a control of these fluctuations. Building on this bound we have the following result.

Theorem 2 Dictionary selection

Let us consider the dictionary selection procedure

$$\widehat{m} = \operatorname*{argmin}_{m=1,\ldots,M} \left\{ \widehat{L}_n(\widehat{h}_{\mathcal{H}_m}) + \mathrm{pen}(\mathcal{H}_m) \right\}, \quad \textit{with} \quad \mathrm{pen}(\mathcal{H}) = 2\sqrt{\frac{2\log(2\,\mathbb{S}_n(\mathcal{H}))}{n}} \, .$$

Then, for any $t > 0$, with probability at least $1 - e^{-t}$ we have

$$L(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) \leq \min_{m=1,\ldots,M} \left\{ \inf_{h \in \mathcal{H}_m} L(h) + 2\,\mathrm{pen}(\mathcal{H}_m) \right\} + \sqrt{\frac{2\log(M) + 2t}{n}} \, . \tag{8}$$

Before proving Theorem 2 let us comment the bound (8). Since $\min_{h\in\mathcal{H}} L(h) \leq L(\widehat{h}_{\mathcal{H}})$, we obtain with probability $1-e^{-t}$

$$L(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) \leq L(\widehat{h}_{\mathcal{H}_*}) + 2\,\mathrm{pen}(\mathcal{H}_*) + \sqrt{\frac{2\log(M)+2t}{n}}\,.$$

In particular, we can compare the misclassification probability of the selected classifier with the misclassification probability of the best classifier among the collection $\left\{\widehat{h}_{\mathcal{H}_1},\ldots,\widehat{h}_{\mathcal{H}_M}\right\}$.

We also notice that the bound (8) increases as $\sqrt{2\log(M)/n}$ with the number M of candidate dictionaries. Finally, the results remain valid if we take $\mathrm{pen}(\mathcal{H})$ larger than $2\sqrt{2\log(2\,\mathbb{S}_n(\mathcal{H}))/n}$.

Proof of Theorem 2. We recall the notation $\widehat{\Delta}_n(\mathcal{H}) = \sup_{h\in\mathcal{H}} \left|\widehat{L}_n(h) - L(h)\right|$. According to Lemma 1.2 and Lemma 1.3 we have with probability at least $1-e^{-t}$

$$\widehat{\Delta}_n(\mathcal{H}_m) \leq \mathrm{pen}(\mathcal{H}_m) + \sqrt{\frac{\log(M)+t}{2n}}, \quad \textit{simultaneously} \text{ for all } m=1,\ldots,M. \tag{9}$$

Therefore, according to Lemma 1.1 and the selection criterion we have with probability at least $1-e^{-t}$

$$\begin{aligned} L(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) &\leq \widehat{L}_n(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) + \mathrm{pen}(\mathcal{H}_{\widehat{m}}) + \sqrt{\frac{\log(M)+t}{2n}} \\ &\leq \min_{m=1,\ldots,M}\left\{\widehat{L}_n(\widehat{h}_{\mathcal{H}_m}) + \mathrm{pen}(\mathcal{H}_m)\right\} + \sqrt{\frac{\log(M)+t}{2n}}\,. \end{aligned} \tag{10}$$

To conclude, we only need to control the size of $\widehat{L}_n(\widehat{h}_{\mathcal{H}_m})$ in terms of $\inf_{h\in\mathcal{H}_m} L(h)$. This can be done directly by combining (3) and (4), but the resulting bound is not tight.

In order to compare $\widehat{L}_n(\widehat{h}_{\mathcal{H}_m})$ to $\inf_{h\in\mathcal{H}_m} L(h)$, let us notice that for any $h\in\mathcal{H}_m$ we have

$$\widehat{L}_n(\widehat{h}_{\mathcal{H}_m}) \leq \widehat{L}_n(h) \leq L(h) + \widehat{\Delta}_n(\mathcal{H}_m),$$

so taking the infimum over $h\in\mathcal{H}_m$ we obtain for all $m=1,\ldots,M$

$$\widehat{L}_n(\widehat{h}_{\mathcal{H}_m}) \leq \inf_{h\in\mathcal{H}_m} L(h) + \widehat{\Delta}_n(\mathcal{H}_m).$$

Combining this bound with (9) and (10), we obtain

$$L(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) \leq \min_{m=1,\ldots,M}\left\{\inf_{h\in\mathcal{H}_m} L(h) + 2\,\mathrm{pen}(\mathcal{H}_m)\right\} + 2\sqrt{\frac{\log(M)+t}{2n}}\,.$$

The proof of Theorem 2 is complete. □

Remark: Combining (9) and Lemma 1.1, we obtain the confidence interval for the misclassification probability

$$\mathbb{P}\left(L(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) \in \left[\widehat{L}_n(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) - \delta(\widehat{m},t), \widehat{L}_n(\widehat{h}_{\mathcal{H}_{\widehat{m}}}) + \delta(\widehat{m},t)\right]\right) \geq 1-e^{-t}$$

$$\text{with} \quad \delta(\widehat{m},t) = \mathrm{pen}(\mathcal{H}_{\widehat{m}}) + \sqrt{\frac{\log(M)+t}{2n}}\,.$$

3.3 Vapnik-Chervonenkis dimension

Computing the shattering coefficient $\mathbb{S}_n(\mathcal{H})$ can be tricky in practice. Nevertheless, a nice combinatorial property of the shattering coefficients provides a simple upper-bound for $\mathbb{S}_n(\mathcal{H})$, which depends on $\mathcal{H}$ only through a single quantity, the so-called Vapnik-Chervonenkis dimension of $\mathcal{H}$.

By convention we set $\mathbb{S}_0(\mathcal{H}) = 1$. We call VC-dimension of $\mathcal{H}$ the integer $d_{\mathcal{H}}$ defined by

$$d_{\mathcal{H}} = \sup\left\{d \in \mathbb{N} : \mathbb{S}_d(\mathcal{H}) = 2^d\right\} \ \in \ \mathbb{N} \cup \{+\infty\}\,.$$

It corresponds to the maximum number of points in $\mathcal{X}$ that can be arbitrarily classified by the classifiers in $\mathcal{H}$. Next proposition gives an upper-bound of the shattering coefficient $\mathbb{S}_n(\mathcal{H})$ in terms of the VC-dimension $d_{\mathcal{H}}$.

Proposition 1 Sauer's lemma

Let $\mathcal{H}$ be a set of classifiers with finite VC-dimension $d_{\mathcal{H}}$. For any $n \in \mathbb{N}$ we have

$$\mathbb{S}_n(\mathcal{H}) \leq \sum_{i=0}^{d_{\mathcal{H}}} C_n^i \leq (n+1)^{d_{\mathcal{H}}} \quad \textit{with} \quad C_n^i = \begin{cases} \frac{n!}{i!\,(n-i)!} & \textit{for } n \geq i \\ 0 & \textit{for } n < i. \end{cases}$$

Proof. We first prove the inequality

$$\mathbb{S}_k(\mathcal{H}) \leq \sum_{i=0}^{d_{\mathcal{H}}} C_k^i \tag{11}$$

for any $\mathcal{H}$ with finite VC-dimension $d_{\mathcal{H}}$, by induction on k.

Let us consider the case $k = 1$. If $d_{\mathcal{H}} = 0$, it means that no point can be shattered so all points can be labelled in only one way. Therefore $\mathbb{S}_1(\mathcal{H}) = 1$ which is equal to C_1^0. If $d_{\mathcal{H}} \geq 1$, we have $\mathbb{S}_1(\mathcal{H}) = 2$ which is also equal to $C_1^0 + C_1^1$.

Assume now that (11) is true for all $k \leq n-1$. Let us consider $\mathcal{H}$ with finite VC-dimension $d_{\mathcal{H}}$. As mentioned above, when $d_{\mathcal{H}} = 0$ all points can only be labelled in one way so $\mathbb{S}_k(\mathcal{H}) = 1$ and (11) is true for all k. We assume now that $d_{\mathcal{H}} \geq 1$. Let $x_1, \ldots, x_n$ be n points in $\mathcal{X}$ and define

$$\mathcal{H}(x_1, \ldots, x_n) = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}\,.$$

The set $\mathcal{H}(x_1, \ldots, x_n)$ depends only on the values of h on $\{x_1, \ldots, x_n\}$, so we can replace $\mathcal{H}$ by $\mathcal{F} = \left\{h|_{\{x_1,\ldots,x_n\}} : h \in \mathcal{H}\right\}$ in the definition of $\mathcal{H}(x_1, \ldots, x_n)$. Since $d_{\mathcal{F}}$ is not larger than $d_{\mathcal{H}}$, we can assume with no loss of generality, that $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{H} = \mathcal{F}$ in the following. Let us consider the set

$$\mathcal{H}' = \left\{h \in \mathcal{H} : h(x_n) = 1 \text{ and } h' = h - 2 \times \mathbf{1}_{\{x_n\}} \in \mathcal{H}\right\}.$$

Since $\mathcal{H}(x_1, \ldots, x_n) = \mathcal{H}'(x_1, \ldots, x_n) \cup (\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)$ we have

$$\text{card}\left(\mathcal{H}(x_1, \ldots, x_n)\right) \leq \text{card}\left(\mathcal{H}'(x_1, \ldots, x_n)\right) + \text{card}\left((\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)\right). \tag{12}$$

Let us bound apart the cardinality of $\mathcal{H}'(x_1, \ldots, x_n)$ and the cardinality of $(\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)$.

1. First, we note that $\mathrm{card}\,(\mathcal{H}'(x_1,\ldots,x_n)) = \mathrm{card}\,(\mathcal{H}'(x_1,\ldots,x_{n-1}))$ since $h(x_n)=1$ for all $h\in\mathcal{H}'$. Second, we observe that the VC-dimension $d_{\mathcal{H}'}$ of $\mathcal{H}'$ is at most $d_{\mathcal{H}}-1$. Actually, if d points $x_{i_1},\ldots,x_{i_d}$ of $\mathcal{X}=\{x_1,\ldots,x_n\}$ are shattered by $\mathcal{H}'$, then $x_n\notin\{x_{i_1},\ldots,x_{i_d}\}$ since $h(x_n)=1$ for all $h\in\mathcal{H}'$. Furthermore, the set $\{x_{i_1},\ldots,x_{i_d},x_n\}$ is shattered by $\mathcal{H}$ due to the definition of $\mathcal{H}'$, so $d+1\le d_{\mathcal{H}}$, which implies $d_{\mathcal{H}'}\le d_{\mathcal{H}}-1$. Applying (11) with $k=n-1$ we obtain that
$$\mathrm{card}\left(\mathcal{H}'(x_1,\ldots,x_n)\right) = \mathrm{card}\left(\mathcal{H}'(x_1,\ldots,x_{n-1})\right) \le \sum_{i=0}^{d_{\mathcal{H}}-1} C_{n-1}^i. \tag{13}$$

2. When $h,h'\in\mathcal{H}\setminus\mathcal{H}'$ fulfill $h(x_i)=h'(x_i)$ for $i=1,\ldots,n-1$, they also fulfill $h(x_n)=h'(x_n)$ by definition of $\mathcal{H}'$. Therefore, we have as above $\mathrm{card}\,((\mathcal{H}\setminus\mathcal{H}')(x_1,\ldots,x_n)) = \mathrm{card}\,((\mathcal{H}\setminus\mathcal{H}')(x_1,\ldots,x_{n-1}))$. Furthermore $d_{\mathcal{H}\setminus\mathcal{H}'}$ is not larger than $d_{\mathcal{H}}$ since $\mathcal{H}\setminus\mathcal{H}'\subset\mathcal{H}$ so equation (11) with $k=n-1$ gives
$$\mathrm{card}\left((\mathcal{H}\setminus\mathcal{H}')(x_1,\ldots,x_n)\right) = \mathrm{card}\left((\mathcal{H}\setminus\mathcal{H}')(x_1,\ldots,x_{n-1})\right) \le \sum_{i=0}^{d_{\mathcal{H}}} C_{n-1}^i\,. \tag{14}$$

Combining (12), (13) and (14), we obtain that
$$\mathrm{card}\,(\mathcal{H}(x_1,\ldots,x_n)) \le \sum_{i=1}^{d_{\mathcal{H}}} C_{n-1}^{i-1} + \sum_{i=0}^{d_{\mathcal{H}}} C_{n-1}^i = \sum_{i=0}^{d_{\mathcal{H}}} C_n^i\,,$$
since $C_{n-1}^i + C_{n-1}^{i-1} = C_n^i$ for $i\ge 1$. As a consequence (11) is true for $k=n$ and the induction is complete.

The second upper-bound of the proposition is obtained by
$$\sum_{i=0}^{d} C_n^i \le \sum_{i=0}^{d} \frac{n^i}{i!} \le \sum_{i=0}^{d} C_d^i\, n^i = (1+n)^d.$$
The proof of Proposition 1 is complete. □

Let us give some examples of VC-dimension for some simple dictionaries on $\mathcal{X}=\mathbb{R}^d$. The proofs are let as exercises.

Example 1: linear classifiers.
The VC-dimension of the set $\mathcal{H}=\{h(x)=\mathrm{sign}(\langle w,x\rangle) : \|w\|=1\}$ of linear classifiers is d.

Example 2: affine classifiers.
The VC-dimension of the set $\mathcal{H}=\{h(x)=\mathrm{sign}(\langle w,x\rangle+b) : \|w\|=1,\ b\in\mathbb{R}\}$ of affine classifiers is $d+1$.

Example 3: hyper-rectangle classifiers.
The VC-dimension of the set $\mathcal{H}=\left\{h(x)=2\,\mathbf{1}_A(x)-1 : A \text{ hyper-rectangle of } \mathbb{R}^d\right\}$ of hyper-rectangle classifiers is $2d$.

Example 4: convex polygon classifiers.
The VC-dimension of the set $\mathcal{H}=\left\{h(x)=2\,\mathbf{1}_A(x)-1 : A \text{ convex polygon of } \mathbb{R}^d\right\}$ of convex polygon classifiers is $+\infty$ (consider n points on the unit sphere : for any subset of these points you can choose their convex hull as convex polygon).

4 From theoretical to practical classifiers

4.1 Empirical risk convexification

The empirical risk classifiers analyzed in the previous section have some very nice statistical properties, but they cannot be used in practice because of their computational cost. Actually, there is no efficient way to minimize (1) since neither $\mathcal{H}$ nor $\widehat{L}_n$ are convex. Some of the most popular classification algorithms are obtained by a simple convex relaxation of the minimization problem (1). The empirical misclassification probability $\widehat{L}_n$ is replaced by some convex surrogate and the set of classifiers $\mathcal{H}$ is replaced by some convex functional set $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Let us consider some convex set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$. A function $f \in \mathcal{F}$ is not a classifier, but we can use it for classification by classifying the data points according to the sign of f. In other words we can associate to f the classifier $\operatorname{sign}(f)$. The empirical misclassification probability of this classifier can be written as

$$\widehat{L}_n(\operatorname{sign}(f)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \operatorname{sign}(f)(X_i)<0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i)<0\}}.$$

Let us replace this empirical misclassification probability $\widehat{L}_n$ by some convex surrogate, which is more amenable to numerical computations. A simple and efficient way to obtain a convex criterion is to replace the loss function $z \to \mathbf{1}_{z<0}$ by some convex function $z \to \ell(z)$. Building on this simple idea, we will focus in the following on classifiers obtained by the procedure

$$\widehat{h}_{\mathcal{F}} = \operatorname{sign}(\widehat{f}_{\mathcal{F}}) \quad \text{where} \quad \widehat{f}_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{L}_n^{\ell}(f) \quad \text{with} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i)). \tag{15}$$

This classifier can be computed efficiently since both $\mathcal{F}$ and $\widehat{L}_n^{\ell}$ are convex. Many popular classifiers are obtained by solving (15) with some specific choices of $\mathcal{F}$ and ℓ, see Section 4.3 and 4.4 for some examples.

Some popular convex loss ℓ

It is natural to consider a convex loss function ℓ which is non-increasing and non-negative. Usually, we also ask that $\ell(z) \geq \mathbf{1}_{z<0}$ for all $z \in \mathbb{R}$ since in this case we can give an upper-bound on the misclassification probability, see Theorem 3. Some classical loss functions are

- the exponential loss $\ell(z) = e^{-z}$
- the logit loss $\ell(z) = \log_2(1 + e^{-z})$
- the hinge loss $\ell(z) = (1 - z)_+ \quad (\text{with } (x)_+ = \max(0, x))$

see Figure 2 for a plot of these three functions.

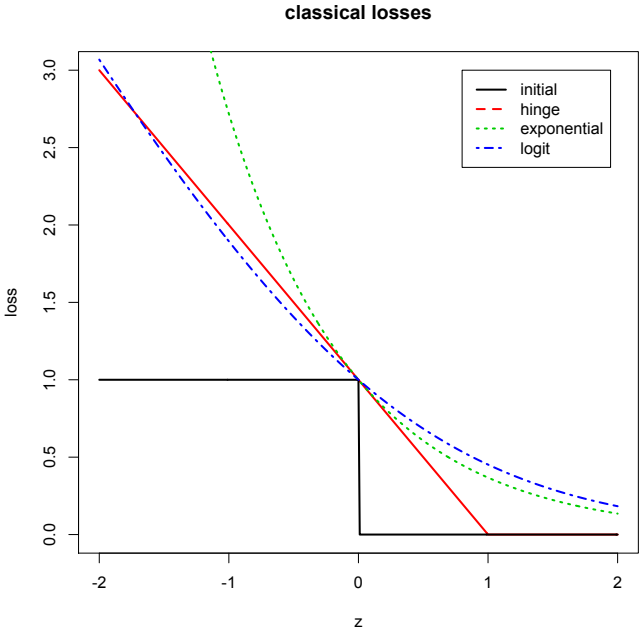

Figure 2: Plot of the exponential, hinge and logit losses

Some classical functional sets $\mathcal{F}$

The main popular convex functional sets $\mathcal{F}$ can be grouped into two class.

A first popular class of sets $\mathcal{F}$ is obtained by taking a linear combination of a finite family $\mathcal{H} = \{h_1, \ldots, h_p\}$ of classifiers

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p} \beta_j h_j(x) \text{ with } \beta_j \in \mathcal{C} \right\}, \tag{16}$$

where $\mathcal{C}$ is a convex subset of $\mathbb{R}^p$. Typical choices for $\mathcal{C}$ are the ℓ^1-ball $\{\beta \in \mathbb{R}^p : |\beta|_1 \leq R\}$, the simplex $\left\{\beta \in \mathbb{R}^p : \beta_j \geq 0,\ \sum_{j=1}^p \beta_j \leq 1\right\}$ or the whole space $\mathbb{R}^p$. This choice appears for example in boosting methods, see Section 4.4. The basic classifiers $\{h_1, \ldots, h_p\}$ are often called *weak learners*. A popular choice of weak learners is $h_j(x) = \text{sign}(x_j - t_j)$ with $t_j \in \mathbb{R}$.

A second popular class of sets $\mathcal{F}$ is obtained by taking a ball of a Reproducing Kernel Hilbert Space (RKHS). We refer to the Appendix A for a brief introduction to RKHS. More precisely, let $\mathcal{F}_k$ be a RKHS with reproducing kernel k and write $\|f\|_{\mathcal{F}}$ for the Hilbert norm of $f \in \mathcal{F}_k$. For notational simplicity, in the following we simply write $\mathcal{F}$ for $\mathcal{F}_k$. Minimizing $\widehat{L}_n^\ell$ over the ball $\{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq R\}$ is equivalent to minimizing over $\mathcal{F}$

the dual Lagrangian problem

$$\widehat{f}_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} \widetilde{L}_n^{\ell}(f) \quad \text{with} \quad \widetilde{L}_n^{\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i\, f(X_i)) + \lambda \|f\|_{\mathcal{F}}^2, \tag{17}$$

for some $\lambda > 0$. This kind of classifier appears for example in Support Vector Machine algorithms, presented in Section 4.3. It turns out that the minimizer $\widehat{f}_{\mathcal{F}}$ of (17) is of the form

$$\widehat{f}_{\mathcal{F}} = \sum_{i=1}^{n} \widehat{\beta}_i\, k(X_i, .). \tag{18}$$

Actually, let V be the linear space spanned by $k(X_1, .), \ldots, k(X_n, .)$. Decomposing $f = f_V + f_{V^\perp}$ on $V \oplus V^\perp$, we have by the reproducing property $f(X_i) = \langle f, k(X_i, .)\rangle_{\mathcal{F}} = \langle f_V, k(X_i, .)\rangle_{\mathcal{F}} = f_V(X_i)$, so the Pythagorean formula gives

$$\widetilde{L}_n^{\ell}(f_V + f_{V^\perp}) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_i f_V(X_i)\big) + \lambda\|f_V\|_{\mathcal{F}}^2 + \lambda\|f_{V^\perp}\|_{\mathcal{F}}^2\,.$$

Since λ is positive, any minimizer $\widehat{f}$ of $\widetilde{L}_n^{\ell}$ must fulfills $\widehat{f}_{V^\perp} = 0$, so it is of the form (18). Furthermore, the reproducing property ensures again that $\langle k(X_i, .), k(X_j, .)\rangle_{\mathcal{F}} = k(X_i, X_j)$ so

$$\left\| \sum_{j=1}^{n} \beta_j k(X_j, .) \right\|_{\mathcal{F}}^2 = \sum_{i,j=1}^{n} \beta_i \beta_j k(X_i, X_j).$$

The minimization problem (17) is then equivalent to

$$\widehat{f}_{\mathcal{F}} = \sum_{j=1}^{n} \widehat{\beta}_j\, k(X_j, .)$$
$$\text{with} \quad \widehat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \beta_j Y_i k(X_j, X_i) \Big) + \lambda \sum_{i,j=1}^{n} \beta_i \beta_j k(X_i, X_j) \right\}. \tag{19}$$

This formulation is of major importance in practice, since it reduces the infinite-dimensional minimization problem (17) into a n-dimensional minimization problem which can be solved efficiently. In the Section 4.3 on Support Vector Machines, we will give a more precise description of the solution of this problem when ℓ is the hinge loss.

4.2 Statistical properties

The classifier $\widehat{h}_{\mathcal{F}}$ given by (15) with $\mathcal{F}$ and ℓ convex has the nice feature to be easy to compute, but does-it have some good statistical properties?

Link with the bayes classifier

The empirical risk minimizer $\widehat{h}_{\mathcal{H}}$ of Section 3 was minimizing the empirical version of the misclassification probability $\mathbb{P}(Y \neq h(X))$ over some set $\mathcal{H}$ of classifiers. The function $\widehat{f}_{\mathcal{F}}$ minimizes instead the empirical version of $\mathbb{E}\left[\ell(Yf(X))\right]$ over some functional set $\mathcal{F}$. The classifier $\widehat{h}_{\mathcal{H}}$ can then be viewed as an empirical version of the bayes classifier h_* which minimizes $\mathbb{P}(Y \neq h(X))$ over the set of measurable functions $h : \mathcal{X} \to \{-1, +1\}$, whereas the function $\widehat{f}_{\mathcal{F}}$ is an empirical version of the function f_*^ℓ which minimizes $\mathbb{E}\left[\ell(Yf(X))\right]$ over the set of measurable functions $f : \mathcal{X} \to \mathbb{R}$. A first point is to understand the link between the bayes classifier h_* and the sign of the function f_*^ℓ. It turns out that under very weak assumptions on ℓ, the sign of f_*^ℓ exactly coincides with the bayes classifier h_*, so $\mathrm{sign}(f_*^\ell)$ minimizes the misclassification probability $\mathbb{P}(Y \neq h(X))$. Let us check this point.

Conditioning on X we have

$$\begin{aligned}\mathbb{E}\left[\ell(Yf(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\ell(Yf(X)|X\right]\right] \\ &= \mathbb{E}\left[\ell(f(X))\mathbb{P}(Y=1|X) + \ell(-f(X))(1-\mathbb{P}(Y=1|X))\right].\end{aligned}$$

Assume that ℓ is decreasing, differentiable and strictly convex (e.g. exponential or logit loss). Minimizing the above expression gives that $f_*^\ell(X)$ is the solution of

$$\frac{\ell'(-f(X))}{\ell'(f(X))} = \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\,.$$

Since ℓ is strictly convex, we have $f(X) > 0$ if and only if $\ell'(-f(X))/\ell'(f(X)) > 1$, so

$$f_*^\ell(X) > 0 \iff \mathbb{P}(Y=1|X) > 1/2 \iff \mathbb{E}[Y|X] = 2\mathbb{P}(Y=1|X) - 1 > 0\,.$$

Since $h_*(X) = \mathrm{sign}(\mathbb{E}\left[Y|X\right])$ (see Section 2), we obtain $\mathrm{sign}(f_*^\ell) = h_*$. This equality also holds true for the hinge loss ℓ (check it!).

To sum up the above discussion, the target function f_*^ℓ approximated by $\widehat{f}_{\mathcal{F}}$ does perfectly make sense for the classification problem since its sign coincides with the best possible classifier h_*.

Upper-bound on the misclassification probability

We focus now on the misclassification probability $L(\widehat{h}_{\mathcal{F}})$ of the classifier $\widehat{h}_{\mathcal{F}} = \mathrm{sign}(\widehat{f}_{\mathcal{F}})$ given by (15). In practice, it is important to have an upper-bound of the misclassification probability $L(\widehat{h}_{\mathcal{F}})$ which can be computed from the data. Next theorem provides such an upper-bound for some typical examples of set $\mathcal{F}$.

Theorem 3 Confidence bound on $L(\widehat{h}_{\mathcal{F}})$

For any $R > 0$, we set $\Delta\ell(R) = |\ell(R) - \ell(-R)|$. We assume here that the loss-function ℓ is convex, non-increasing, non-negative, α-Lipschitz on $[-R, R]$ and fulfills $\ell(z) \geq \mathbf{1}_{z<0}$ for all z in $\mathbb{R}$. We consider the classifier $\widehat{h}_{\mathcal{F}}$ given by (15).

(a) *When $\mathcal{F}$ is of the form (16) with $\mathcal{C} = \{\beta \in \mathbb{R}^p : |\beta|_1 \leq R\}$, we have with probability at least $1 - e^{-t}$*

$$L(\widehat{h}_{\mathcal{F}}) \;\leq\; \widehat{L}_n^{\ell}(\widehat{f}_{\mathcal{F}}) + 2\alpha R\sqrt{\frac{2\log(2p)}{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,. \tag{20}$$

(b) *Let $\mathcal{F}$ be the ball of radius R of a RKHS with kernel k fulfilling $k(x,x) \leq 1$ for all $x \in \mathcal{X}$. Then, we have with probability at least $1 - e^{-t}$*

$$L(\widehat{h}_{\mathcal{F}}) \;\leq\; \widehat{L}_n^{\ell}(\widehat{f}_{\mathcal{F}}) + \frac{2\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,. \tag{21}$$

Proof. We first prove a general upper-bound for $L(\widehat{h}_{\mathcal{F}})$, similar to Theorem 1.

Lemma 3.1

Assume that $\sup_{f\in\mathcal{F}} |f(x)| \leq R < +\infty$. For any loss ℓ fulfilling the hypotheses of Theorem 3, we have with probability at least $1 - e^{-t}$

$$L(\widehat{h}_{\mathcal{F}}) \leq \widehat{L}_n^{\ell}(\widehat{f}_{\mathcal{F}}) + \frac{2\alpha}{n}\max_{x\in\mathcal{X}^n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}} \tag{22}$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. random variables with distribution $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$.

Proof of Lemma 3.1. The proof of this lemma relies on the same arguments as the proof of Theorem 1. The first point is to notice that since $\ell(z) \geq \mathbf{1}_{z<0}$ for all real z, we have

$$\begin{array}{rcll} L(\widehat{h}_{\mathcal{F}}) = \mathbb{P}(Y\widehat{f}_{\mathcal{F}}(X) < 0) & \leq & L^{\ell}(f) & \text{with } L^{\ell}(f) = \mathbb{E}\left[\ell(Yf(X))\right] \\ & \leq & \widehat{L}_n^{\ell}(f) + \widehat{\Delta}_n^{\ell}(\mathcal{F}) & \text{where } \widehat{\Delta}_n^{\ell}(\mathcal{F}) = \sup\limits_{f\in\mathcal{F}} \left|\widehat{L}_n^{\ell}(f) - L^{\ell}(f)\right|. \end{array}$$

As in Lemma 1.2, the McDiarmid concentration inequality (Theorem 4 in the Appendix B) ensures that with probability at least $1 - e^{-t}$ we have

$$\widehat{\Delta}_n^{\ell}(\mathcal{F}) \leq \mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathcal{F})\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,.$$

To conclude the proof of the lemma, it only remains to prove that

$$\mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathcal{F})\right] \leq \frac{2\alpha}{n}\max_{x\in\mathcal{X}^n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right]. \tag{23}$$

Following exactly the same lines as in the proof of Lemma 1.3 (replacing $\mathbf{1}_{Y_i \neq h(X_i)}$ by $\ell(Y_i f(X_i))$) we obtain

$$\mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathcal{F})\right] \leq \frac{2}{n} \max_{y \in \{-1,+1\}^n} \max_{x \in \mathcal{X}^n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i \ell(y_i f(x_i)) \right| \right].$$

We finally use the α-Lipschitz property of ℓ to conclude : according to the Contraction principle (see Proposition 5 in Appendix B) we have

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i \ell(y_i f(x_i)) \right| \right] \leq \alpha\, \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i y_i f(x_i) \right| \right] = \alpha\, \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right].$$

Combining the last two bounds gives (23) and the proof of Lemma 3.1 is complete. □

(a) Let us prove now the bound (20). The map $\beta \to \sum_{i=1}^n \sigma_i \sum_{j=1}^p \beta_j h_j(x_i)$ is linear, so it reaches its maximum and minimum on the ℓ^1-ball $\mathcal{C}$ at one of the vertices of $\mathcal{C}$. Therefore, we have

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right] = R\, \mathbb{E}_\sigma \left[\max_{j=1,\dots,p} \left| \sum_{i=1}^n \sigma_i h_j(x_i) \right| \right].$$

It remains to apply the inequality (6) with $\mathcal{V} = \{(h_j(x_1), \dots, h_j(x_n)) : j = 1, \dots, p\}$ whose cardinality is at most p in order to obtain

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right] \leq R\, \sqrt{2n \log(2p)}.$$

The bound (20) follows from Lemma 3.1.

(b) We now turn to the second bound (21) and write $\|.\|_{\mathcal{F}}$ for the norm in the RKHS. According to the reproducing formula and the Cauchy Schwartz inequality, we have

$$\left| \sum_{i=1}^n \sigma_i f(x_i) \right| = \left| \left\langle f, \sum_{i=1}^n \sigma_i k(x_i, .) \right\rangle_{\mathcal{F}} \right| \leq \|f\|_{\mathcal{F}} \left\| \sum_{i=1}^n \sigma_i k(x_i, .) \right\|_{\mathcal{F}}.$$

Applying again Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\mathbb{E}_\sigma \left[\sup_{\|f\|_{\mathcal{F}} \leq R} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right] &\leq R\, \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i k(x_i, .) \right\|_{\mathcal{F}} \right] \\
&\leq R \sqrt{\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i k(x_i, .) \right\|_{\mathcal{F}}^2 \right]} \\
&\leq R \sqrt{\sum_{i=1}^n k(x_i, x_i) \mathbb{E}_\sigma[\sigma_i^2]} \;\leq\; R\sqrt{n},
\end{aligned}$$

where we have used $k(x, x) \leq 1$ in the last inequality and $\mathbb{E}[\sigma_i \sigma_j] = 0$ for $i \neq j$ in the previous one. Combining again the reproducing property with the Cauchy-Schwartz inequality, we obtain

$$|f(x)| = |\langle f, k(x, .) \rangle_{\mathcal{F}}| \leq R \sqrt{k(x, x)} \leq R.$$

The Lemma 3.1 then gives

$$L(\widehat{h}_{\mathcal{F}}) \leq \widehat{L}_n^{\ell}(\widehat{f}_{\mathcal{F}}) + \frac{2\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\ ,$$

and the proof of Theorem 3 is complete. □

It is possible do derive risk bounds similar to (3) for $L(\widehat{h}_{\mathcal{H}})$, we refer to Bousquet, Boucheron and Lugosi [1] for a review of such results. In the remaining of these notes, we will describe two very popular classification algorithms : the so-called Support Vector Machine and AdaBoost.

4.3 Support Vector Machine

The Support Vector Machine (SVM) algorithm corresponds to the estimator (17) with the hinge loss $\ell(z) = (1-z)_+$. The final classification is performed according to $\widehat{h}_{\mathcal{F}}(x) = \text{sign}(\widehat{f}_{\mathcal{F}}(x))$. It turns out that there is a very nice geometrical interpretation of the solution $\widehat{f}_{\mathcal{F}}$, from which originates the name "Support Vector Machine".

Proposition 2 Support Vectors

The solution of (17) is of the form $\widehat{f}_{\mathcal{F}}(x) = \sum_{i=1}^n \widehat{\beta}_i\, k(X_i, x)$ *with*

$$\begin{cases} \widehat{\beta}_i = 0 & \text{if} \quad Y_i\widehat{f}_{\mathcal{F}}(X_i) > 1 \\ \widehat{\beta}_i = Y_i/(2\lambda n) & \text{if} \quad Y_i\widehat{f}_{\mathcal{F}}(X_i) < 1 \\ 0 \leq Y_i\widehat{\beta}_i \leq 1/(2\lambda n) & \text{if} \quad Y_i\widehat{f}_{\mathcal{F}}(X_i) = 1\,. \end{cases}$$

The vectors X_i *with index* i *such that* $\widehat{\beta}_i \neq 0$ *are called support vectors.*

Proof. Writing K for the matrix $[k(X_i, X_j)]_{i,j=1,\dots,n}$, we know from (19) that the solution of (17) is of the form $\widehat{f}_{\mathcal{F}} = \sum_{j=1}^n \widehat{\beta}_j\, k(X_j, .)$ with

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^n}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \big(1 - Y_i[K\beta]_i\big)_+ + \lambda\beta^T K\beta \right\}.$$

The above minimization problem is not smooth, so we introduce some slack variables $\widehat{\xi}_i = (1 - Y_i[K\widehat{\beta}]_i)_+$ and rewrite the minimization problem as

$$(\widehat{\beta}, \widehat{\xi}) = \underset{\substack{\beta,\,\xi\,\in\,\mathbb{R}^n \text{ such that} \\ \xi_i \geq 1 - Y_i[K\beta]_i \\ \xi_i \geq 0}}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda\beta^T K\beta \right\}. \tag{24}$$

This problem is now smooth and convex and the Karush-Kuhn-Tucker conditions for the Lagrangian dual problem

$$(\widehat{\beta}, \widehat{\xi}) = \underset{\beta,\xi\in\mathbb{R}^n}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda\beta^T K\beta - \sum_{i=1}^n \big(\alpha_i(\xi_i - 1 + Y_i[K\beta]_i) + \gamma_i\xi_i\big) \right\}.$$

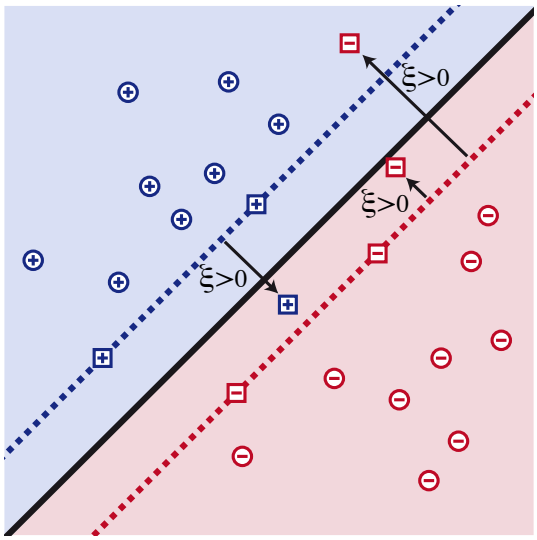

Figure 3: Classification with a linear SVM: the separating hyperplane $\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\}$ is represented in black, the margin-hyperplanes $\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = +1\}$ and $\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = -1\}$ are represented in dotted blue and red respectively. The support vectors are represented by squares.

gives the formulas

$$\text{first-order conditions} \quad : \quad 2\lambda[K\widehat{\beta}]_j = \sum_{i=1}^{n} K_{ij}\alpha_i Y_i \quad \text{and} \quad \alpha_j + \gamma_j = \frac{1}{n},$$

$$\text{slackness conditions} \quad : \quad \min(\alpha_i, \widehat{\xi}_i - 1 + Y_i[K\widehat{\beta}]_i) = 0 \quad \text{and} \quad \min(\gamma_i, \widehat{\xi}_i) = 0.$$

We deduce from the first first-order conditions that $\widehat{\beta}_i = \alpha_i Y_i/(2\lambda)$. Since $\widehat{f}_{\mathcal{F}}(X_i) = [K\widehat{\beta}]_i$, the first slackness condition enforces that $\widehat{\beta}_i = 0$ if $Y_i \widehat{f}_{\mathcal{F}}(X_i) > 1$. The second slackness condition together with the second first-order optimality condition enforces that $\widehat{\beta}_i = Y_i/(2\lambda n)$ if $\widehat{\xi}_i > 0$ and $0 \leq Y_i\widehat{\beta}_i \leq 1/(2\lambda n)$ otherwise. To conclude the proof of the proposition, we notice that when $\widehat{\xi}_i > 0$ we have $\widehat{\beta}_i$ and α_i non-zero, and therefore $Y_i \widehat{f}_{\mathcal{F}}(X_i) = 1 - \widehat{\xi}_i < 1$ according to the first slackness condition. □

Let us now interpret geometrically Proposition 2.

Geometrical interpretation : linear kernel

We start with the simplest kernel $k(x, y) = \langle x, y\rangle$ for all $x, y \in \mathbb{R}^d$. The associated RKHS is the space of linear forms $\mathcal{F} = \{\langle w, .\rangle : w \in \mathbb{R}^d\}$. In this case

$$\widehat{f}_{\mathcal{F}}(x) = \sum_{i=1}^{n} \widehat{\beta}_i \langle X_i, x\rangle = \langle \widehat{w}, x\rangle \quad \text{with} \quad \widehat{w} = \sum_{i=1}^{n} \widehat{\beta}_i X_i \,,$$

so the classifier $\widehat{h}_{\mathcal{F}}(x) = \text{sign}(\langle \widehat{w}, x\rangle)$ assigns labels to points according to their position relative to the hyperplane $\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\}$. The normal to the hyperplane $\widehat{w}$ is a linear

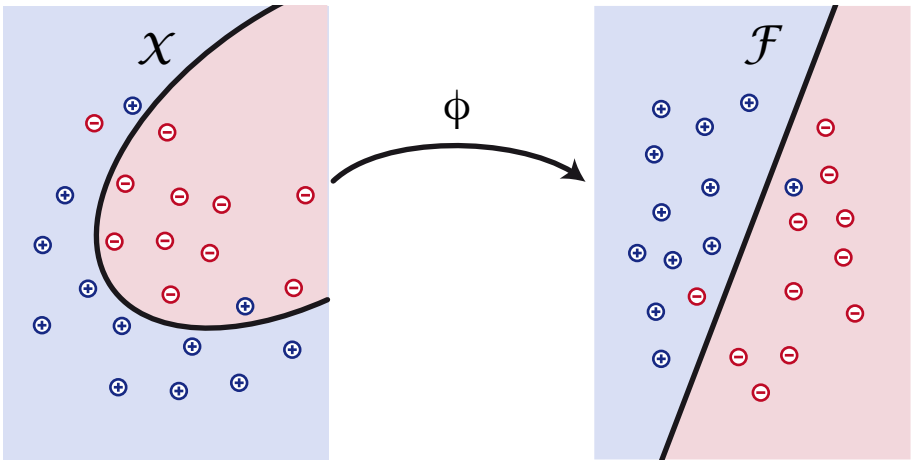

Figure 4: Classification with a non-linear kernel: the linear classification in $\mathcal{F}$ produces a non-linear classification in $\mathcal{X}$ via the reciprocal image of ϕ.

combination of the support vectors, which are the data points X_i such that $Y_i\langle \widehat{w}, X_i\rangle \leq 1$. They are represented by squares in the Figure 3. The hyperplanes $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = +1\right\}$ and $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = -1\right\}$ are usually called margin-hyperplanes. We notice the following important property of the SVM. If we add to the learning dataset a point X_{n+1} which fulfills $Y_{n+1}\langle \widehat{w}, X_{n+1}\rangle > 1$, then the vector $\widehat{w}$ and the classifier $\widehat{h}_{\mathcal{F}}$ do not change. In other words, only data points that are wrongly classified or classified with not enough margin (i.e. $Y_i\langle \widehat{w}, X_i\rangle \leq 1$) do influence the separating hyperplane $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\right\}$.

Geometrical interpretation : arbitrary positive definite kernels

Let us denote by $\phi : \mathcal{X} \to \mathcal{F}$ the map $\phi(x) = k(x, .)$. According to the reproducing property and Proposition 2, we have

$$\widehat{f}_{\mathcal{F}}(x) = \langle \widehat{f}_{\mathcal{F}}, \phi(x)\rangle_{\mathcal{F}} = \left\langle \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i), \phi(x) \right\rangle_{\mathcal{F}}.$$

A point $x \in \mathcal{X}$ is classified according to the sign of the above scalar product. Therefore the points $\phi(x) \in \mathcal{F}$ are classified according to the linear classifier on $\mathcal{F}$

$$f \mapsto \text{sign}\left(\langle \widehat{w}_\phi, f\rangle_{\mathcal{F}}\right) \quad \text{where } \widehat{w}_\phi = \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i).$$

The separating frontier $\left\{x \in \mathcal{X} : \widehat{f}_{\mathcal{F}}(x) = 0\right\}$ of the classifier $\widehat{h}_{\mathcal{F}}$ is therefore the reciprocal image by ϕ of the hyperplane $\{f \in \mathcal{F} : \langle \widehat{w}_\phi, f\rangle_{\mathcal{F}} = 0\}$, as represented in Figure 4. We observe, that the kernel k delinearalizes the SVM, in the sense that it produces a non-linear classifier $\widehat{h}_{\mathcal{F}}$ with the same computational cost as a linear one in $\mathbb{R}^n$.

You can observe SVM in action with the following recreative applet:
http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml

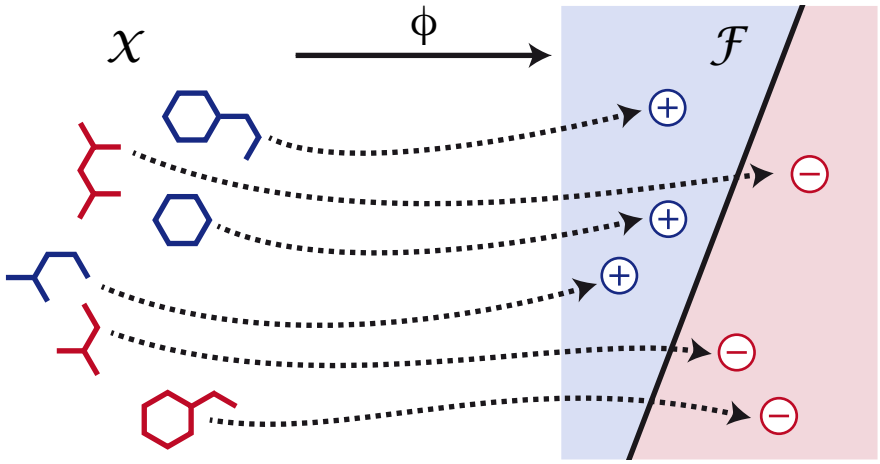

Figure 5: Classification of molecules with a SVM.

Why are RKHS useful?

There are mainly two main reasons for using RKHS. The first reason is that using RKHS allows to delinearlize some algorithm by mapping $\mathcal{X}$ in $\mathcal{F}$ with $\phi : x \to k(x, .)$, as represented in Figure 4. It then provides non-linear algorithms with almost the same computational complexity as a linear one.

The second reason is that it allows to apply to any set $\mathcal{X}$ some algorithms that are defined for vectors. Assume for example that we want to classify some proteins or molecules according to their therapeutic properties. Let $\mathcal{X}$ represents our set of molecules. For any $x, y \in \mathcal{X}$, let us represent by $k(x, y)$ how similar they are for us. If our kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, then we can directly apply the SVM algorithm in order to classify them, see Figure 5. Of course, the key point in this case is to design the kernel k. Usually, the kernel $k(x, y)$ is designed according to some properties of x, y which are known to be relevant for the classification problem. For example, the number of common short sequences is a useful index of similarity between two proteins. The computational complexity for evaluating $k(x, y)$ is also an issue which is crucial in many applications with complex data. We refer to Jean-Philippe Vert's slides on computational biology for many promising applications in biology and medicine:
http://cbio.ensmp.fr/~jvert/talks/120302ensae/ensae.pdf

4.4 AdaBoost

AdaBoost is an algorithm which computes an approximation of the estimator (15) with the exponential loss $\ell(z) = e^{-z}$ and the functional space $\mathcal{F} = \text{span}\{h_1, \ldots, h_p\}$ where $h_1, \ldots, h_p$ are p arbitrary classifiers.

The principle of the AdaBoost algorithm is to perform a greedy minimization of (15). More precisely it computes a sequence of functions $\widehat{f}_m$ for $m = 0, \ldots, M$ by starting from $\widehat{f}_0 = 0$

and then solving for $m = 1, \dots, M$

$$\widehat{f}_m = \widehat{f}_{m-1} + \beta_m h_{j_m} \quad \text{where} \quad (\beta_m, j_m) = \underset{\substack{j=1,\dots,p \\ \beta \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \exp\Big(-Y_i\big(\widehat{f}_{m-1}(X_i) + \beta h_j(X_i)\big)\Big).$$

The final classification is performed by $\widehat{h}_M(x) = \operatorname{sign}(\widehat{f}_M(x))$ which is an approximation of $\widehat{h}_{\mathcal{H}}$ defined by (15).

The exponential loss allows to compute (β_m, j_m) very efficiently. Actually, setting $w_i^{(m)} = n^{-1} \exp(-Y_i \widehat{f}_{m-1}(X_i))$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \exp\Big(-Y_i\big(\widehat{f}_{m-1}(X_i) + \beta h_j(X_i)\big)\Big) = (e^{\beta} - e^{-\beta}) \sum_{i=1}^{n} w_i^{(m)} \mathbf{1}_{h_j(X_i) \neq Y_i} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)}.$$

When the condition

$$\operatorname{err}_m(j) = \frac{\sum_{i=1}^{n} w_i^{(m)} \mathbf{1}_{h_j(X_i) \neq Y_i}}{\sum_{i=1}^{n} w_i^{(m)}} \leq \frac{1}{2} \quad \text{for all } j = 1, \dots, p,$$

is fulfilled, the minimizers (β_m, j_m) are given by

$$j_m = \underset{j=1,\dots,p}{\operatorname{argmin}} \operatorname{err}_m(j) \quad \text{and} \quad \beta_m = \frac{1}{2} \log\left(\frac{1 - \operatorname{err}_m(j_m)}{\operatorname{err}_m(j_m)}\right).$$

Noticing that $-Y_i h(X_i) = 2\mathbf{1}_{Y_i \neq h(X_i)} - 1$ we obtain the standard formulation of the AdaBoost algorithm.

AdaBoost

Init: $w_i^{(1)} = 1/n$, for $i = 1, \dots, n$

Iterate: For $m = 1, \dots, M$ do

$$\begin{aligned} j_m &= \underset{j=1,\dots,p}{\operatorname{argmin}} \operatorname{err}_m(j) \\ 2\beta_m &= \log(1 - \operatorname{err}_m(j_m)) - \log(\operatorname{err}_m(j_m)) \\ w_i^{(m+1)} &= w_i^{(m)} \exp(2\beta_m \mathbf{1}_{h_{j_m}(X_i) \neq Y_i}), \quad \text{for } i = 1, \dots, n \\ \text{STOP} \quad &\text{if} \quad \min_{j=1,\dots,p} \operatorname{err}_{m+1}(j) > 1/2 \end{aligned}$$

Output: $\widehat{f}_M(x) = \sum_{m=1}^{M} \beta_m h_{j_m}(x)$.

We notice that the AdaBoost algorithm gives more and more weight in $\operatorname{err}_m(j)$ to the data points X_i which are wrongly classified at the stage m.

You can observe AdaBoost in action (with half-plane weak-learners h_j) with the following recreative applet: http://cseweb.ucsd.edu/∼yfreund/adaboost/

5 Going beyond these lecture notes

For the reader interested to go beyond the basic concepts presented in these lecture notes, we refer to the survey by Boucheron, Bousquet and Lugosi [1] for recent developments on the topic and a comprehensive bibliography. For more practical consideration, we refer to the book by Hastie, Tibshirani and Friedman [5] where many practical algorithms are described and discussed. Finally, we point out that the concepts introduced here arise also for the ranking problem (rank at best some data, as google does), see Clémençon, Lugosi and Vayatis [3].

A Reproducing kernel Hilbert spaces

Reproducing Kernel Hilbert Spaces (RKHS) are some functional Hilbert spaces, where the smoothness of a function is driven by its norm. RKHS also fulfill a special "reproducing property" which is crucial in practice since it allows efficient numerical computations.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a positive definite kernel if it is symmetric ($k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$) and if for any $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$ and $a_1, \ldots, a_N \in \mathbb{R}$ we have

$$\sum_{i,j=1}^{N} a_i a_j k(x_i, x_j) \geq 0. \tag{25}$$

Examples of positive definite kernels in $\mathcal{X} = \mathbb{R}^d$:

- linear kernel: $k(x, y) = \langle x, y \rangle$
- Gaussian kernel: $k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$
- histogram kernel ($d = 1$): $k(x, y) = \min(x, y)$
- exponential kernel: $k(x, y) = e^{-\|x-y\|/\sigma}$

We can associate to a positive definite kernel k a special Hilbert subspace $\mathcal{F}$ of $\mathbb{R}^{\mathcal{X}}$ called reproducing kernel Hilbert space associated to k.

Proposition 3 Reproducing Kernel Hilbert Space (RKHS)

To any positive definite kernel k on $\mathcal{X}$, we can associate a (unique) Hilbert space $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ fulfilling

- *$k(x, .) \in \mathcal{F}$ for all $x \in \mathcal{X}$*
- **reproducing property:** *$f(x) = \langle f, k(x, .) \rangle_{\mathcal{F}}$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$.*

The space $\mathcal{F}$ is called the reproducing kernel Hilbert space associated to k.

Proof. Let us consider the linear space $\mathcal{F}_0$ spanned by the family $\{k(x,.) : x \in \mathcal{X}\}$

$$\mathcal{F}_0 = \left\{ f : \mathcal{X} \to \mathbb{R} : f(x) = \sum_{i=1}^{N} a_i k(x_i, x),\ N \in \mathbb{N},\ x_1, \dots, x_N \in \mathcal{X},\ a_1, \dots, a_N \in \mathbb{R} \right\}.$$

To any $f = \sum_{i=1}^{N} a_i k(x_i,.)$ and $g = \sum_{j=1}^{M} b_j k(y_j,.)$ we associate

$$\langle f, g \rangle_{\mathcal{F}_0} := \sum_{i=1}^{N} \sum_{j=1}^{M} a_i b_j k(x_i, y_j) = \sum_{i=1}^{N} a_i g(x_i) = \sum_{j=1}^{M} b_j f(y_j).$$

From the last two equalities we see that $\langle f, g \rangle_{\mathcal{F}_0}$ does not depend on the choice of the expansion of f and g, so it is well-defined. Furthermore, the application $(f, g) \to \langle f, g \rangle_{\mathcal{F}_0}$ is bilinear, symmetric, positive (according to (25)) and we have the reproduction property

$$f(x) = \langle f, k(x,.) \rangle_{\mathcal{F}_0} \quad \text{for all } x \in \mathcal{X} \quad \text{and } f \in \mathcal{F}_0.$$

The Cauchy-Schwartz inequality $\langle f, g \rangle_{\mathcal{F}_0} \leq \|f\|_{\mathcal{F}_0} \|g\|_{\mathcal{F}_0}$ and the reproduction formula give

$$|f(x)| \leq \sqrt{k(x,x)}\ \|f\|_{\mathcal{F}_0}. \tag{26}$$

As a consequence $\|f\|_{\mathcal{F}_0} = 0$ implies $f = 0$ so $\langle f, g \rangle_{\mathcal{F}_0}$ is a scalar product. Therefore $\mathcal{F}_0$ is a pre-Hilbert space and to obtain $\mathcal{F}$ we only need to complete $\mathcal{F}_0$.

Let us consider two sequences $(x_i) \in \mathcal{X}^{\mathbb{N}}$ and $(a_i) \in \mathbb{R}^{\mathbb{N}}$ fulfilling $\sum_{i,j \geq 1} a_i a_j k(x_i, x_j) < +\infty$. According to (26) for any $M < N$ and $x \in \mathcal{X}$ we have

$$\Big| \sum_{i=M+1}^{N} a_i k(x_i, x) \Big| \leq \sqrt{k(x,x)} \sum_{i,j=M+1}^{N} a_i a_j k(x_i, x_j).$$

When $\sum_{i,j \geq 1} a_i a_j k(x_i, x_j)$ is finite, the right-hand side goes to 0 when M, N goes to infinity, so the partial series $\sum_{i=1}^{N} a_i k(x_i, x)$ is Cauchy and it converges when $N \to \infty$. We can therefore define the space

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} : f(x) = \sum_{i=1}^{\infty} a_i k(x_i, x),\ (x_i) \in \mathcal{X}^{\mathbb{N}},\ (a_i) \in \mathbb{R}^{\mathbb{N}},\ \sum_{i,j \geq 1} a_i a_j k(x_i, x_j) < +\infty \right\}$$

and the bilinear form

$$\langle f, g \rangle_{\mathcal{F}} := \sum_{i,j=1}^{\infty} a_i b_j k(x_i, y_j) = \sum_{i=1}^{\infty} a_i g(x_i) = \sum_{j=1}^{\infty} b_j f(y_j)$$

for $f = \sum_{i=1}^{\infty} a_i k(x_i,.)$ and $g = \sum_{j=1}^{\infty} b_j k(y_j,.)$. Exactly as before, the application $(f, g) \to \langle f, g \rangle_{\mathcal{F}}$ is a scalar product fulfilling the reproduction property

$$f(x) = \langle f, k(x,.) \rangle_{\mathcal{F}} \quad \text{for all } x \in \mathcal{X} \text{ and } f \in \mathcal{F}.$$

Finally, the space $\mathcal{F}$ endowed with $\langle .,. \rangle_{\mathcal{F}}$ is the completion in $\mathbb{R}^{\mathcal{X}}$ of $\mathcal{F}_0$ endowed with $\langle .,. \rangle_{\mathcal{F}_0}$, so it is a reproducing kernel Hilbert space. □

The norm of a function f in a RKHS $\mathcal{F}$ is strongly linked to its smoothness. This appears clearly in the inequality

$$|f(x) - f(x')| = |\langle f, k(x,.) - k(x',.)\rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \, \|k(x,.) - k(x',.)\|_{\mathcal{F}}.$$

Let us illustrate this point by describing the RKHS associated to the histogram and Gaussian kernels.

Example 1: RKHS associated to the histogram kernel.
The Sobolev space

$$\mathcal{F} = \big\{ f \in C([0,1],\mathbb{R}) : f \text{ is a.e. differentiable with } f' \in L^2([0,1]) \text{ and } f(0) = 0 \big\}$$

endowed with the scalar product $\langle f, g \rangle_{\mathcal{F}} = \int_0^1 f'g'$ is a RKHS with reproducing kernel $k(x,y) = \min(x,y)$ on $[0,1]$. Actually, $k(x,.) \in \mathcal{F}$ for all $x \in [0,1]$ and

$$f(x) = \int_0^1 f'(y) \mathbf{1}_{y \leq x} \, dy = \langle f, k(x,.) \rangle_{\mathcal{F}}, \quad \text{for all } f \in \mathcal{F} \text{ and } x \in [0,1].$$

In this case the norm $\|f\|_{\mathcal{F}}$ corresponds simply to the L^2-norm of the derivative of f. The smaller is this norm, the smoother is f.

Example 2: RKHS associated to the Gaussian kernel.
Let us write $\mathbf{F}[f]$ for the Fourier transform in $\mathbb{R}^d$ with normalization

$$\mathbf{F}[f](\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-\mathbf{i}\langle \omega, t \rangle}, \quad \text{for } f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \text{ and } \omega \in \mathbb{R}^d.$$

For any $\sigma > 0$, the functional space

$$\mathcal{F}_\sigma = \left\{ f \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \text{ such that } \int_{\mathbb{R}^d} \big|\mathbf{F}[f](\omega)\big|^2 e^{\sigma|\omega|^2/2} \, d\omega < +\infty \right\},$$

endowed with the scalar product $\langle f, g \rangle_{\mathcal{F}_\sigma} = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \overline{\mathbf{F}[f](\omega)} \mathbf{F}[g](\omega) e^{\sigma|\omega|^2/2} \, d\omega$ is a RKHS associated with the Gaussian kernel $k(x,y) = \exp(-\|y-x\|^2/2\sigma^2)$. Actually, for all $x \in \mathbb{R}^d$ the function $k(x,.)$ belongs to $\mathcal{F}_\sigma$ and straightforward computations give

$$\langle k(x,.), f \rangle_{\mathcal{F}_\sigma} = \mathbf{F}^{-1}\big[\mathbf{F}[f]\big](x) = f(x) \quad \text{for all } f \in \mathcal{F} \text{ and all } x \in \mathbb{R}^d.$$

The space $\mathcal{F}_\sigma$ gathers very regular functions and the norm $\|f\|_{\mathcal{F}_\sigma}$ directly controls the smoothness of f. We note that when σ increases the space $\mathcal{F}_\sigma$ shrinks and contains smoother and smoother functions.

B Concentration inequalities

Non-asymptotic analyses in statistical learning often rely on McDiarmid concentration inequality [6].

Theorem 4 McDiarmid (1989)

Let $\mathcal{X}$ be some measurable set and let us consider $F : \mathcal{X}^n \to \mathbb{R}$ such that there exists $\delta_1, \ldots, \delta_n$ fulfilling

$$\left|F(x_1, \ldots, x_i', \ldots, x_n) - F(x_1, \ldots, x_i, \ldots, x_n)\right| \leq \delta_i, \quad \text{for all} \quad x_1, \ldots, x_n, x_i' \in \mathcal{X},$$

for any $i = 1, \ldots, n$. Then for any $t > 0$ and any independent random variables $X_1, \ldots, X_n$, we have

$$\mathbb{P}\big(F(X_1, \ldots, X_n) > \mathbb{E}\left[F(X_1, \ldots, X_n)\right] + t\big) \leq \exp\left(-\frac{2t^2}{\delta_1^2 + \ldots + \delta_n^2}\right).$$

We refer to the Chapter 9 of Devroye, Györfi and Lugosi [4] for the classical proof of this result, combining a Markov inequality with martingale arguments. A more conceptual proof based on the Entropy method is given in Boucheron, Lugosi and Massart [2], Chapter 6.

Another important concept in statistical learning theory is the Contraction Principle.

Theorem 5 Contraction Principle

Let $\mathcal{Z}$ be a bounded subset of $\mathbb{R}^n$ and $\varphi : \mathbb{R} \to \mathbb{R}$ an α-Lipschitz function fulfilling $\varphi(0) = 0$. For $\sigma_1, \ldots, \sigma_n$ i.i.d. random variables with distribution $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$, we have

$$\mathbb{E}_\sigma\left[\sup_{z \in \mathcal{Z}} \left|\sum_{i=1}^n \sigma_i \varphi(z_i)\right|\right] \leq \alpha\, \mathbb{E}_\sigma\left[\sup_{z \in \mathcal{Z}} \left|\sum_{i=1}^n \sigma_i z_i\right|\right].$$

A concise proof is available in Chapter 11 of the book by Boucheron, Lugosi and Massart [2].

References

[1] Boucheron, S., Bousquet, O. and Lugosi, G. *Theory of classification: some recent advances.* ESAIM Probability & Statistics, **9** (2005) : 323–375.

[2] Boucheron, S., Lugosi, G. and Massart, P. *Concentration Inequalities.* Oxford University Press, 2013.

[3] Clémençon, S., Lugosi, G. and Vayatis, N. *Ranking and empirical risk minimization of U-statistics.* The Annals of Statistics, **36** (2008) : 844–874.

[4] Devroye, L., Györfi, L. and Lugosi, G. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[5] Hastie, T., Tibshirani, R. and Friedman, J. *The element of statistical learning.* Springer, 2009.

[6] McDiarmid, C. *On the Method of Bounded Differences.* Surveys in Combinatorics **141** (1989) : 148–188.

Author informations

Christophe Giraud

Professor at Paris-Sud University and Ecole Polytechnique
Director of the master program "Mathematics for Life Sciences"
http://webens-ng.math.u-psud.fr/M2/MathSV/

Contact: christophe.giraud@math.u-psud.fr

Personal website: http://www.cmap.polytechnique.fr/∼giraud/