PART ONE

MACHINE LEARNING

Executive Summary

What exactly constitutes Data Science is not universally agreed upon, but it is certainly inseparable from Machine Learning. Some would even consider Data Science as a subfield of Machine Learning—its intersection with application domains. This book adopts a different viewpoint: Data Science is seen as incorporating the field of Machine Learning.

A necessarily incomplete outline of this vast field is presented in the first of the book's five parts. It starts by considering the scenario of supervised learning, in which a to-be-learned function f is available only through point values $y_i = f(x^{(i)})$ at datapoints $x^{(1)}, \ldots, x^{(m)}$. In Statistical Learning Theory, these datapoints are assumed to be realizations of some hidden random variable. Chapter 1 introduces the main notions attached to this theory, in particular the PAC-learning framework. Chapter 2 scrutinizes the concept of VC-dimension, in anticipation of its connection to the problem of binary classification, where the labels y_i take only two values. Chapter 3, of a technical nature, makes this connection precise by establishing the fundamental theorem of PAC-learning. Chapter 4 continues to probe the problem of binary classification but drops the statistical setting. It proposes some tools—in particular, support vector machines—to separate datapoints and it also acquaints the readers with kernel methods. Chapter 5 takes a careful look at the associated reproducing kernel Hilbert spaces. Chapter 6 concludes the tour of supervised learning by way of a few peeks at the regression problem, featuring real-valued labels y_i. Chapter 7 turns to the scenario of unsupervised learning, in which the labels are absent: the task examined there consists in exploiting similarity information about the datapoints to cluster them in a meaningful way. Finally, Chapter 8 presents common techniques to deal with the hindering high-dimensionality of datapoints.

Readers in search of a more detailed exposition to Machine Learning are referred to the books by Shalev-Shwartz and Ben-David (2014) and Mohri et al. (2018). For more targeted reading, they can also consult the books by Hastie et al. (2009), Scholkopf and Smola (2001), and Vershynin (2018).

1
Rudiments of Statistical Learning Theory

In the scenario considered in the next few chapters, data reach a learner in the form

$$y_i = f(x^{(i)}), \qquad i \in [1:m].$$

Both the *instances* $x^{(i)} \in \mathcal{X}$ and the *targets* $y_i \in \mathcal{Y}$ are known to the learner. It is often the case that $\mathcal{X} \subseteq \mathbb{R}^d$ is made of vectors containing d *features*, overlooking here how these features are created, and that $\mathcal{Y}$ is a discrete set whose elements represent certain classes, in which case the y_i are called *labels*. The postulate of statistical learning theory is that $x^{(1)}, \ldots, x^{(m)}$ come as independent realizations of a single random variable—whose distribution is not available to the learner. The implicit assumption that the targets y_i depend deterministically on the instances $x^{(i)}$ via $y_i = f(x^{(i)})$ for some function $f\colon \mathcal{X} \to \mathcal{Y}$ could be relaxed. It is indeed usual, although not examined in this book, to consider the couples $(x^{(i)}, y_i) \in \mathcal{X} \times \mathcal{Y}$ as independent realizations of a random variable (x, y) with a distribution on $\mathcal{X} \times \mathcal{Y}$ for which $\mathbb{E}[y|x] = f(x)$.

1.1 True and Empirical Risks

The learner's objective is to exploit the data given through the *training sample* $\mathcal{S} = ((x^{(1)}, y_1), \ldots, (x^{(m)}, y_m))$ and to produce a function $h_{\mathcal{S}}\colon \mathcal{X} \to \mathcal{Y}$, called a *predictor*, as a substitute for the unknown function $f\colon \mathcal{X} \to \mathcal{Y}$. The map $\Delta\colon \mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^m \mapsto h_{\mathcal{S}} \in F(\mathcal{X}, \mathcal{Y})$ does not need to be computationally feasible at this point, so Δ is referred to as a *learning map* rather than a learning algorithm. The performance of a given predictor $h \in F(\mathcal{X}, \mathcal{Y})$ is assessed by how small its *risk* is. The latter, also called the *generalization error*, is defined relative to a loss function by

$$\mathrm{Risk}_f(h) := \mathbb{E}[\mathrm{Loss}(h(x), f(x))], \tag{1.1}$$

where the expectation is taken over a random variable x whose distribution is the one that generated the $x^{(i)}$. The *loss function*, defined on $\mathcal{Y} \times \mathcal{Y}$ and taking values in $[0, \infty)$, should be small when its two inputs are close and large when they are far. For *binary classification*, i.e., the situation where $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, +1\}$, a popular choice is the *0/1-loss*, given by

$$\mathrm{Loss}_{0/1}(y, y') = \mathbb{1}_{\{y \neq y'\}} = \begin{cases} 1 & \text{if } y \neq y', \\ 0 & \text{if } y = y'. \end{cases}$$

For *regression*, i.e., the situation where $\mathcal{Y} = \mathbb{R}$, a popular choice is the *square loss*, given by

$$\mathrm{Loss}_{\mathrm{sq}}(y, y') = (y - y')^2.$$

Notice that the learner does not have access to the true risk defined in (1.1), since the distribution generating $x^{(1)}, \ldots, x^{(m)}$ is not available. But the training sample $\mathcal{S} = ((x^{(1)}, y_1), \ldots, (x^{(m)}, y_m))$ supplies an ersatz known as the *empirical risk*, which is defined by

$$\widehat{\mathrm{Risk}}_{\mathcal{S}}(h) := \frac{1}{m} \sum_{i=1}^{m} \mathrm{Loss}(h(x^{(i)}), y_i).$$

Without constraint on $h \in F(\mathcal{X}, \mathcal{Y})$, minimizing the empirical risk is easy: one can create a predictor $h_{\mathcal{S}}$ yielding $\widehat{\mathrm{Risk}}_{\mathcal{S}}(h_{\mathcal{S}}) = 0$ by forcing $h_{\mathcal{S}}(x^{(i)}) = y_i$ for each $i \in [1 : m]$ and choosing $h_{\mathcal{S}}(x)$ arbitrarily for $x \notin \{x^{(1)}, \ldots, x^{(m)}\}$, e.g. as a constant there. However, such a predictor will not generalize well, in the sense that the true risk (aka generalization error) will not be small.

This phenomenon is attenuated by calling upon a prior belief that realistic predictors are close to functions from a certain *hypothesis class* $\mathcal{H} \subseteq F(\mathcal{X}, \mathcal{Y})$. Thus, with the constraint that h belongs to $\mathcal{H}$, the *empirical risk minimization* strategy offers the natural learning map defined by

$$\Delta_{\mathcal{H}}^{\mathrm{erm}} : \mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^m \mapsto \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{\mathrm{Risk}}_{\mathcal{S}}(h) \in \mathcal{H}.$$

The risk of this empirical risk minimizer decomposes as

$$\mathrm{Risk}_f(\Delta_{\mathcal{H}}^{\mathrm{erm}}(\mathcal{S})) = \varepsilon_{\mathrm{app}} + \varepsilon_{\mathrm{est}},$$

i.e., as the sum of the *approximation error* $\varepsilon_{\mathrm{app}} \geq 0$ and the *estimation error* $\varepsilon_{\mathrm{est}} \geq 0$, respectively given by

$$\begin{aligned} \varepsilon_{\mathrm{app}} &:= \inf_{h \in \mathcal{H}} \mathrm{Risk}_f(h), \\ \varepsilon_{\mathrm{est}} &:= \mathrm{Risk}_f(\Delta_{\mathcal{H}}^{\mathrm{erm}}(\mathcal{S})) - \inf_{h \in \mathcal{H}} \mathrm{Risk}_f(h). \end{aligned}$$

The approximation error $\varepsilon_{\mathrm{app}}$ is independent of the sample $\mathcal{S}$ and reflects how

well f can be approximated by elements from the given hypothesis class. The estimation error ε_{est} is the object of the considerations that follow.

1.2 PAC-Learnability

In the *probably approximately correct* (PAC for short) framework, one attempts to make ε_{est} smaller than a prescribed accuracy $\varepsilon \in (0,1)$ with a prescribed confidence $\delta \in (0,1)$. It is sometimes required to do so via an efficient learning algorithm, i.e., an algorithm whose runtime is polynomial in ε^{-1}, δ^{-1}, and the sizes of the problem. This is not enforced in the formal definition below, in which the probability is taken over $x^{(1)}, \dots, x^{(m)}$, understood as independent random variables.

Definition 1.1 A hypothesis class $\mathcal{H} \subseteq F(\mathcal{X}, \mathcal{Y})$ is called *PAC-learnable* with respect to a loss function $\text{Loss}\colon \mathcal{Y} \times \mathcal{Y} \to [0,\infty)$ if there exists a learning map $\Delta\colon \mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^m \mapsto h_{\mathcal{S}} \in \mathcal{H}$ such that, for all $f\colon \mathcal{X} \to \mathcal{Y}$ and all $\varepsilon, \delta \in (0,1)$,

$$\mathbb{P}\Big[\text{Risk}_f(h_{\mathcal{S}}) - \inf_{h \in \mathcal{H}} \text{Risk}_f(h) \le \varepsilon \Big] \ge 1 - \delta,$$

independently of the probability distribution on $\mathcal{X}$, provided that

$$m \ge m_{\mathcal{H}}(\varepsilon, \delta)$$

for some $m_{\mathcal{H}}\colon (0,1)^2 \to \mathbb{N}^*$ growing at most polynomially in ε^{-1} and δ^{-1}.

The smallest possible function $m_{\mathcal{H}}$ appearing in this definition is referred to as the *sample complexity*. For binary classification with the 0/1-loss, it would have been equivalent to state the definition with Δ specifically taken to be the empirical risk minimization map. This will be revealed by the fundamental theorem of PAC-learning in Chapter 3. As a prelude to this theorem, the next result shows that a class of boolean functions that is finite is automatically PAC-learnable for the 0/1-loss. This is an example of a distribution-free result, since no assumption on the underlying probability distribution is made.

Proposition 1.2 *Given a finite set $\mathcal{H} \subseteq F(\mathcal{X}, \{0,1\})$ and a loss function with values in $[0,1]$, the empirical risk minimization strategy provides a learning map $\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^m \mapsto h_{\mathcal{S}} \in \mathcal{H}$ such that, for all boolean functions $f\colon \mathcal{X} \to \{0,1\}$ and all $\varepsilon, \delta \in (0,1)$,*

$$\mathbb{P}\Big[\text{Risk}_f(h_{\mathcal{S}}) - \inf_{h \in \mathcal{H}} \text{Risk}_f(h) \le \varepsilon \Big] \ge 1 - \delta \tag{1.2}$$

provided that

$$m \geq \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2}. \tag{1.3}$$

Proof The inequality (1.2) shall be established in the equivalent form

$$\mathbb{P} := \mathbb{P}[\,\mathrm{Risk}_f(h_S) - \mathrm{Risk}_f(h_*) > \varepsilon] \leq \delta, \tag{1.4}$$

where $h_* \in \mathcal{H}$ is chosen so that $\mathrm{Risk}_f(h_*)$ is equal to $\inf_{h\in\mathcal{H}} \mathrm{Risk}_f(h)$ (or is arbitrarily close to it in case the infimum is not achieved). From the definition of empirical risk minimization, one observes that $\widehat{\mathrm{Risk}}_S(h_S) \leq \widehat{\mathrm{Risk}}_S(h_*)$ and, in turn, that

$$\begin{aligned}\mathrm{Risk}_f(h_S) - \mathrm{Risk}_f(h_*) &= (\,\mathrm{Risk}_f(h_S) - \widehat{\mathrm{Risk}}_S(h_S)) + (\widehat{\mathrm{Risk}}_S(h_S) - \mathrm{Risk}_f(h_*))\\ &\leq (\,\mathrm{Risk}_f(h_S) - \widehat{\mathrm{Risk}}_S(h_S)) + (\widehat{\mathrm{Risk}}_S(h_*) - \mathrm{Risk}_f(h_*))\\ &\leq 2\sup_{h\in\mathcal{H}} |\widehat{\mathrm{Risk}}_S(h) - \mathrm{Risk}_f(h)|.\end{aligned}$$

As a consequence, one has

$$\begin{aligned}\mathbb{P} &\leq \mathbb{P}\Big[\sup_{h\in\mathcal{H}} |\widehat{\mathrm{Risk}}_S(h) - \mathrm{Risk}_f(h)| > \frac{\varepsilon}{2}\Big]\\ &= \mathbb{P}\Big[|\widehat{\mathrm{Risk}}_S(h) - \mathrm{Risk}_f(h)| > \frac{\varepsilon}{2} \text{ for some } h \in \mathcal{H}\Big].\end{aligned} \tag{1.5}$$

For a fixed $h \in \mathcal{H}$, the *Hoeffding inequality* (see Theorem B.6) yields

$$\begin{aligned}&\mathbb{P}\Big[|\widehat{\mathrm{Risk}}_S(h) - \mathrm{Risk}_f(h)| > \frac{\varepsilon}{2}\Big]\\ &\quad= \mathbb{P}\Big[\Big|\frac{1}{m}\sum_{i=1}^m \mathrm{Loss}(h(x^{(i)}), f(x^{(i)})) - \mathbb{E}[\mathrm{Loss}(f(x), h(x))]\Big| > \frac{\varepsilon}{2}\Big]\\ &\quad\leq 2\exp\Big(-\frac{\varepsilon^2 m}{2}\Big),\end{aligned}$$

having used the fact that the random variables $\mathrm{Loss}(h(x^{(i)}), f(x^{(i)}))$ take values in $[0, 1]$. A union bound in (1.5) now implies that

$$\mathbb{P} \leq 2|\mathcal{H}|\exp\Big(-\frac{\varepsilon^2 m}{2}\Big).$$

This is bounded above by δ exactly when $m \geq 2\ln(2|\mathcal{H}|/\delta)/\varepsilon^2$, i.e., when Condition (1.3) is fulfilled. □

1.3 Validation

With m and δ being fixed, it is apparent from (1.3) that enlarging the class $\mathcal{H}$ has the effect of increasing (a bound on) the estimation error ε_{est}. At the same time, enlarging the class $\mathcal{H}$ has the effect of decreasing the approximation error ε_{app}. Thus, in order to keep the total error $\varepsilon_{\text{app}} + \varepsilon_{\text{est}}$ low, a compromise is to be found for the size of $\mathcal{H}$. This observation exemplifies the *bias-complexity tradeoff*. In more general situations, it remains intuitive that a small hypothesis class is not flexible enough to perform well on the sample (this phenomenon is called *underfitting*), while a large hypothesis class can match the sample perfectly but perform poorly on other datapoints (this phenomenon is called *overfitting*); see Figure 1.1 for an illustration.

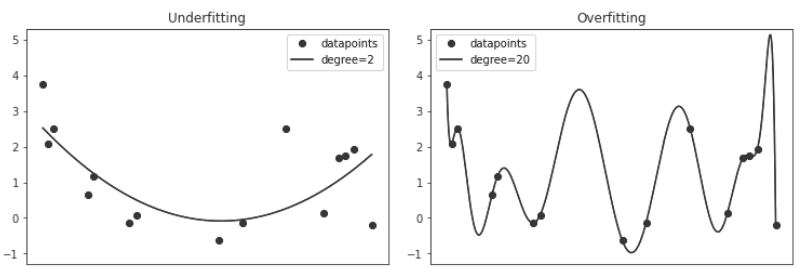

Figure 1.1 Data fitting with univariate polynomials results in underfitting when the degree is low (left) and in overfitting when the degree is high (right).

Even after having decided on a hypothesis class $\mathcal{H}$ and a learning map Δ, the learner will still find it difficult to evaluate the true risk of the predictor $h = \Delta(\mathcal{S})$, as the definition (1.1) involves two unknown entities: the function f and the distribution over which the expectation is taken. The natural ersatz $\widehat{\text{Risk}}_{\mathcal{S}}$ is not a reliable substitute for $\text{Risk}_f(h)$ because the learning map Δ is designed to make this empirical risk small, yet its performance on unseen datapoints remains uncertain. A heuristic workaround consists in partitioning the sample $\mathcal{S}$ into a *training set* $\mathcal{T}$ and a *validation set* $\mathcal{V}$. The training set $\mathcal{T}$ is used to produce the predictor $h = \Delta(\mathcal{T})$, whose performance is then assessed by the empirical risk $\widehat{\text{Risk}}_{\mathcal{V}}(h)$ relative to the validation set $\mathcal{V}$. *Cross-validation* actually consists in partitioning $\mathcal{S}$ into K groups $\mathcal{U}_1, \dots, \mathcal{U}_K$ of roughly equal size and to repeat, for each $k \in [1:K]$, the above procedure with $\mathcal{S} \setminus \mathcal{U}_k$ and $\mathcal{U}_k$ as training and validation sets, respectively.

Exercises

1.1 Given $h \in F(\mathcal{X}, \mathcal{Y})$, verify that the expectation of the empirical risk over the independent random variables $x^{(1)}, \ldots, x^{(m)}$ agrees with the true risk, i.e., that $\mathbb{E}[\widehat{\text{Risk}}_{\mathcal{S}}(h)] = \mathbb{E}[\text{Loss}(h(x), f(x))]$. Verify also that its variance satisfies $\mathbb{V}[\widehat{\text{Risk}}_{\mathcal{S}}(h)] = \mathbb{V}[\text{Loss}(h(x), f(x))]/m$.

1.2 Let $\mathcal{H}$ be the hypothesis class of *affine functions* on $\mathbb{R}^d$, i.e., of functions of the form

$$x \in \mathbb{R}^d \mapsto a_0 + a_1 x_1 + \cdots + a_d x_d \in \mathbb{R}.$$

For the square loss, observe that the empirical risk minimization strategy reduces to the least-squares problem of minimizing $\|y - Xa\|_2^2$ over all $a \in \mathbb{R}^{d+1}$ for some matrix $X \in \mathbb{R}^{m \times (d+1)}$ to identify.

1.3 Let a sample $\mathcal{S}$ be partitioned into a training set $\mathcal{T}$ and a validation set $\mathcal{V}$. Considering the hypothesis class of affine functions and the square loss, let $h_{\mathcal{T}}$ denote the empirical risk minimizer relative to $\mathcal{T}$. Prove that the expected empirical risk of $h_{\mathcal{T}}$ is no larger on $\mathcal{T}$ than on $\mathcal{V}$, i.e., that

$$\mathbb{E}\Big[\widehat{\text{Risk}}_{\mathcal{T}}(h_{\mathcal{T}})\Big] \le \mathbb{E}\Big[\widehat{\text{Risk}}_{\mathcal{V}}(h_{\mathcal{T}})\Big],$$

with expectation taken over all the independent random variables $x^{(i)}$.

1.4 When (x, y) is a random variable over $\mathcal{X} \times \mathcal{Y}$, the *risk* of a predictor $h \colon \mathcal{X} \to \mathcal{Y}$ is defined relative to a loss function via

$$\text{Risk}(h) := \mathbb{E}[\text{Loss}(h(x), y)],$$

with expectation now taken jointly over x and y.
For regression with the square loss, defining $f(x) := \mathbb{E}[y|x]$ to be the conditional probability of y given x, establish the identity

$$\text{Risk}(h) = \text{Risk}(f) + \mathbb{E}\Big[(h(x) - f(x))^2\Big],$$

showing that f is an optimal predictor.

For classification with the 0/1-loss, prove that an optimal predictor is given by the *Bayes predictor* defined for $x \in \mathcal{X}$ by

$$f(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1|x] \ge \mathbb{P}[y = 0|x], \\ 0 & \text{otherwise.} \end{cases}$$

2

Vapnik–Chervonenkis Dimension

While the concept of dimension usually applies to sets, the concept of *Vapnik–Chervonenkis dimension*, or VC-dimension for short, applies to families of sets (as subsets of some bigger set $\mathcal{X}$). Since one can identify a subset of $\mathcal{X}$ with its indicator function via the correspondence between $\mathcal{S} \subseteq \mathcal{X}$ and $\mathbb{1}_{\mathcal{S}} \in F(\mathcal{X}, \{0,1\})$, the concept of VC-dimension applies in a similar way to families of boolean functions. This is the viewpoint taken in Machine Learning, where a family of boolean functions is thought of as a hypothesis class. The fundamental theorem of PAC-learning, to be covered in the next chapter, will reveal the importance of the concept of VC-dimension: a hypothesis class is PAC-learnable if and only if it has a finite VC-dimension.

2.1 Definitions

Here is the formal definition of *VC-dimension* that adopts the viewpoint of boolean functions.

Definition 2.1 Let $\mathcal{H}$ be a family of boolean functions defined on a set $\mathcal{X}$. A subset $\mathcal{Y}$ of $\mathcal{X}$ is said to be *shattered* by $\mathcal{H}$ if any $g\colon \mathcal{Y} \to \{0,1\}$ takes the form $g = h_{|\mathcal{Y}}$ for some $h \in \mathcal{H}$. The VC-dimension of $\mathcal{H}$ is the largest size of a subset shattered by $\mathcal{H}$. In short,

$$\mathrm{vc}(\mathcal{H}) = \sup\{m \in \mathbb{N}^* : \tau_{\mathcal{H}}(m) = 2^m\},$$

where the *shatter function* (aka *growth function*) is defined by

$$\tau_{\mathcal{H}}(m) = \max_{|\mathcal{Y}|=m} \big|\{h_{|\mathcal{Y}}, h \in \mathcal{H}\}\big|.$$

By adopting the viewpoint of sets rather than boolean functions, one can repeat the definition of VC-dimension as the equivalent statement below, which