553.740: Introduction to Machine Learning

Laurent Younes

Contents

CHAPTER 1

Introduction: Bias and Variance

Machine learning is an interdisciplinary subject which is often seen at the interface of computer science, applied mathematics and statistics. From statistics, and more specially nonparametric statistics, it borrows its main formalism, asymptotic results and generalization bounds. It also borrows from and extends many classical methods that have been designed for estimation and prediction. From computer science, it involves the design and implementation of efficient algorithms, which in turn often include classical and sometimes new methods from optimization theory. With the rise of massive datasets, the subject has furthermore expanded to include methods for storing, sharing and managing data, powerful computer architectures for increasingly demanding algorithms, in a new field that is (currently...) referred to as "data science." These notes, however, take a statistician viewpoint and therefore mostly focus on this side of the subject, which is, arguably, the most fundamental. This will not prevent us from detailing some of the important algorithms, and therefore introducing some notions from optimization theory, while trying to make our presentation self-contained. In contrast, we will assume sufficient knowledge in probability theory and multivariate statistics, as provided in textbooks such as [**27, 22, 34**].

In "classical" non-parametric statistics [**26, 15, 24**], one generally considers three main supervised estimation problems, namely:

- Density estimation: From a training set $(x_1, \dots, x_N)$, considered as an i.i.d. sample of a r.v. X, find a function $\hat{f}_X$ that estimate the p.d.f. of X. [1]
- Regression: given a training set $(x_1, y_1), \dots, (x_N, y_N)$, considered as a sample of a pair of random variables (X, Y) with $x_k \in \mathcal{R}$ (typically $\mathbb{R}^d$) and $y_k \in \mathbb{R}^q$, find a function $\hat{f}_n : \mathbb{R}^d \to \mathbb{R}^q$ that approximates the conditional expectation $E(Y|X)$.
- Classification: given a training set $(x_1, y_1), \dots, (x_N, y_N)$, considered as a sample of a pair of random variables (X, Y) with $x_k \in \mathcal{R}$ and y_k in a finite set $\mathcal{G}$ of classes, find a function $\hat{f}_n : \mathbb{R}^d \to \mathcal{G}$ which approximate the best prediction of the true class knowing X.

[1] i.i.d.: independent, identically distributed; i.i.d. sample: a realization of N i.i.d. random variables; p.d.f.: probability density function; r.v.: random variable.

A large part of these notes address the the last two issues. We will also discuss unsupervised "dimension reduction" and clustering methods. Density estimation will, however, not be a major focus for us, except in this chapter, in which we will use it as an example to introduce the "bias vs. variance dilemma."

1.1. Parameter Estimation

Here and in the rest of this chapter, the observation is an i.i.d. sample $(x_1, \ldots, x_N)$ of a r.v. X, taking values in $\mathbb{R}^d$, with unknown probability density function (p.d.f.) f and the goal is to estimate f. The resulting estimator therefore is a function $x \mapsto \hat{f}(x)$ from $\mathbb{R}^d$ to $[0, +\infty)$. The function $\hat{f}$ also depends on the sample, which we will make explicit if needed by writing $x \mapsto \hat{f}(x; x_1, \ldots, x_N)$.

Parameter estimation is the most common density estimation method, in which one restrict $\hat{f}$ to belong to a finite-dimensional parametric class, denoted $(f_\theta, \theta \in \Theta)$, with $\Theta \subset \mathbb{R}^p$. For example, f_θ can be a family of Gaussian distributions on $\mathbb{R}^d$. With our notation, this corresponds to estimators taking the form

$$\hat{f}(x; x_1, \ldots, x_N) = f_{\hat{\theta}(x_1,\ldots,x_N)}(x)$$

and the problem becomes the estimation of the parameter $\hat{\theta}$.

There are several, well-known methods for parameter estimation, and, since this is not the focus of the course, we only consider the most important one: maximum likelihood, which consists in computing $\hat{\theta}$ that maximizes the log-likelihood

$$C(\theta) = \frac{1}{N} \sum_{k=1}^{N} \log f_\theta(x_k)\,. \tag{1}$$

The resulting $\hat{\theta}$ (when it exists) is called the maximum likelihood estimator of θ, or m.l.e.

If the true $f = f_{\theta_*}$ belongs to the parametric class, standard results in mathematical statistics [**7, 22**] provide sufficient conditions for $\hat{\theta}$ to converge to θ_* when N tends to infinity. However, the fact that the true p.d.f. belongs to the finite dimensional class (f_θ) is an optimistic assumption that is generally false. The standard theorems in parametric statistics should be considered as a "best case scenario analysis," or a "sanity check," in which one asks whether, in the ideal situation in which f actually belongs to the parametric class, the designed estimator has a proper behavior. In non-parametric statistics, a parametric model can still be a plausible approach in order to approximate the true f, but the relevant question should then be whether $\hat{f}$ provides (asymptotically), the best approximation to f among all f_θ, $\theta \in \Theta$. The maximum likelihood estimator can be analyzed in this

setting, if one measures the difference between two density functions by the Kullback-Liebler divergence (also called differential entropy):

$$KL(f\|f_\theta) = \int_{\mathbb{R}^d} \log \frac{f(x)}{f_\theta(x)} f(x)dx$$

which is positive unless $f = f_\theta$ (and may be equal to $+\infty$). Minimizing this measure with respect to θ is equivalent to maximizing

$$E_f(\log f_\theta) = \int_{\mathbb{R}^d} \log f_\theta(x) f(x)dx\,,$$

and an empirical evaluation of this expectation is $\frac{1}{N}\sum_{k=1}^N \log f_\theta(x_k)$, which provides the maximum likelihood method. Seen in this context, consistency of the maximum likelihood estimator states that this estimator almost surely converges to a best approximator of the true f in the class $(f_\theta, \theta \in \Theta)$. More precisely, if one assumes that the function $\theta \mapsto \log f_\theta(x)$ is continuous in θ for almost all x (upper-semi continuous is actually enough) and that, for all $\theta \in \Theta$, there exists a small enough $\delta > 0$ such that

$$E_f\Big(\sup_{|\theta'-\theta|<\delta} \log f_\theta\Big) < \infty$$

then, letting Θ_* denote the set of maximizers of $E_f(\log f_\theta)$, and assuming that it is not empty, the maximum likelihood estimator $\hat\theta_N$ is such that, for all $\varepsilon > 0$ and all compact subsets $K \subset \Theta$,

$$\lim_{N\to\infty} P(d(\hat\theta_N, \Theta_*) > \varepsilon \text{ and } \hat\theta_N \in K) \to 0.$$

The interested reader can refer to [**34**], Theorem 5.14, for a proof of this statement. Note that this assertion does not exclude that $\hat\theta_N$ goes to infinity (i.e., steps out of ever compact subset K in Θ), and the boundedness of the m.l.e. is either asserted from additional properties of the likelihood, or by simply restricting Θ to be a compact set.

If $\Theta_* = \{\theta_*\}$ and the m.l.e. is consistent, the speed of convergence can also be quantified by a central limit theorem (see [**34**], Theorem 5.23) ensuring that, in standard cases $\sqrt{N}(\hat\theta_N - \theta_*)$ converges to a normal distribution.

Even though these results relate our present subject to classical parametric statistics, they are not that interesting, because, when $f \neq f_{\theta_*}$, the convergence of the m.l.e. to the best approximator in Θ still leaves a gap regarding the estimation of f. This gap is often called the bias of the class $(f_\theta, \theta \in \Theta)$. One can reduce it by considering larger classes (e.g., with more dimensions), but the larger the class, the less accurate the estimation of the best approximator becomes for a fixed sample size (the estimator has a larger *variance*). This issue is known as the "bias vs. variance dilemma," and to address it, it is necessary to adjust the class Θ to the sample size in

order to optimally balance the two types of error (and all estimation methods have at least one mechanism that allows for this). When the "tuning parameter" is the dimension of Θ, the overall approach is often referred to as the "Method of Sieves" [**20, 18**], in which the dimension of Θ is increased as a function of N in a suitable way.

Gaussian mixture models provide one of the most popular choices with the method of sieves. In the 1D case ($d = 1$), this corresponds to taking

$$\Theta_N = \left\{ f : f(x) = \frac{1}{m_N} \sum_{j=1}^{m_N} \frac{e^{-(x-y_i)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}, y_1, \ldots, y_{m_N} \in \mathbb{R}, \sigma > 0 \right\}.$$

The integer m_N allows one to tune the dimension of Θ_N and therefore controls the bias-variance trade-off. For such a set, there exists a simple algorithm for computing the m.l.e., called the EM algorithm, which will be described at the end of this chapter.

The method of sieves is not the only approach to estimate densities. Among the most widely used methods are "kernel density estimators," [**23, 29, 30**] which compute

$$\hat{f}_\sigma(x) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sigma^d} K((x - x_k)/\sigma)$$

where K is a given positive function such that $\int_{\mathbb{R}^d} K(x)dx = 1$. This condition implies that $\hat{f}_\sigma$ is a density function since

$$\begin{aligned} \int_{\mathbb{R}^d} \hat{f}_\sigma(x) &= \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sigma^d} \int_{\mathbb{R}^d} K((x - x_k)/\sigma)dx \\ &= \frac{1}{N} \sum_{k=1}^{N} \int_{\mathbb{R}^d} K(y)dy = 1 \end{aligned}$$

(the change of variable $y = (x - x_k)/\sigma$ yields $dy = dx/\sigma^d$). The bias-variance trade-off is now decided by the value of σ.

A typical choice for K is a Gaussian kernel, $K(y) = e^{-|y|^2/2}/(2\pi)^{d/2}$. In this case, the estimated density is a sum of bumps centered at the data-points $x_1, \ldots, x_N$. The width of the bumps is controlled by the parameter σ. Small σ yields less rigidity in the model, which will therefore be more affected by changes in the data: the estimated density will have a larger variance. The converse is true for large σ, at the cost of being less able to adapt to variations in the true density: the model has a larger bias.

The correct approach is to let $\sigma = \sigma_N$ depend on the size of the training set. In particular, it can be shown that, as soon as $N\sigma_N^d$ tends to infinity and σ_N tends to 0, the pointwise estimation $\hat{f}_N(x)$ converges to $f(x)$. Let

us justify this with an informal computation. We have

$$\begin{aligned} E(\hat{f}_\sigma(x)) &= \frac{1}{N\sigma^d}\sum_{k=1}^{N} E\big(K((x-X_k)/\sigma)\big) \\ &= \frac{1}{\sigma^d}\int_{\mathbb{R}^d} K((y-x)/\sigma)f(y)dy \\ &= \int_{\mathbb{R}^d} K(z)f(x+\sigma z)dz \end{aligned}$$

The bias of the estimator, i.e., the average difference between $\hat{f}(x)$ and $f(x)$ is therefore given by

$$E(\hat{f}_\sigma(x)) - f(x) = \int_{\mathbb{R}^d} K(z)(f(x+\sigma z) - f(x))dz.$$

The variance of $\hat{f}_\sigma(x)$ is given by

$$\mathrm{var}(\hat{f}_\sigma(x)) = \frac{1}{N\sigma^{2d}}\mathrm{var}(K((x-X)/\sigma))$$

with

$$\begin{aligned} \frac{1}{N\sigma^{2d}}\mathrm{var}(K((x-X)/\sigma)) &= \frac{1}{N\sigma^{2d}}\int_{\mathbb{R}}^{d} K((x-y)/\sigma)^2 f(y)dy \\ &\quad -\frac{1}{N\sigma^{2d}}\left(\int_{\mathbb{R}^d} K((y-x)/\sigma)f(y)dy\right)^2 \\ &= \frac{1}{N\sigma^d}\int_{\mathbb{R}^d} K(z)^2 f(x+\sigma z)dz - \frac{1}{N}\left(\int_{\mathbb{R}^d} K(z)f(x+\sigma z)dz\right)^2 \end{aligned}$$

The total mean square error of the estimator is

$$E((\hat{f}_\sigma(x) - f(x))^2) = \mathrm{var}(\hat{f}_\sigma(x)) + (E(\hat{f}_\sigma(x)) - f(x))^2.$$

From the expressions above, we can already notice that, for fixed N, when $\sigma \to 0$, the variance of the estimator goes to infinity, whereas the bias goes to 0. When $\sigma \to \infty$, it is the opposite: the variance becomes small and the bias increases. This is a perfect illustration of the dilemma: there exists an intermediate value of σ that is optimal. Now, if σ is allowed to depend on N, the combination of $\sigma_N \to 0$ and $N\sigma_N^d \to \infty$ clearly implies that both bias and variance tend to 0.

In order to infer an optimal way to select σ as a function of N, we make a Taylor expansion of both bias and variance. Assume that f has at least two derivatives. One can write

$$f(x+\sigma z) = f(x) + \sigma z^T \nabla f(x) + \frac{\sigma^2}{2} z^T d^2 f(x) z + o(\sigma^2 z^2).$$

We now use the fact K is even, which ensures that $\int zK(z)dz = 0$, to write

$$E(\hat{f}_\sigma(x)) - f(x) = \frac{\sigma^2}{2} M_f(x) + o(\sigma^2)$$

with $M_f = \int K(z) z^T d^2 f(x) z \, dz$. Similarly, letting $S = \int K^2(z) \, dz$,

$$\text{var}(\hat{f}_\sigma(x)) = \frac{1}{N\sigma^d}(Sf(x) + o(\sigma^d + \sigma^2)).$$

We can obtain an asymptotically optimal value for σ by minimizing the leading terms of the mean square error, namely

$$\frac{\sigma^4}{4} M_f^2 + \frac{1}{N\sigma^d} Sf(x)$$

which yields $\sigma = O(N^{-1/(d+4)})$ and

$$E((\hat{f}_\sigma(x) - f(x))^2) = O(N^{-4/(d+4)}).$$

If f has r derivatives, one can, with a proper selection of K (not the Gaussian) and of σ_N, ensure that the mean square error

$$\int_{\mathbb{R}^d} |f(x) - \hat{f}_N(x)|^2 dx$$

has order $N^{-\frac{2r}{2r+d}}$. This can be shown to be "optimal", in the following sense, called min-max: for any other estimator, there exists a function f for which the convergence speed is at least as bad as this one.

This result is "bad news" in large dimension, since it says that, to obtain a given accuracy ε in the worst case scenario, N should be chosen of order $(1/\varepsilon)^{1+(d/2r)}$ which grows exponentially fast with the dimension. This is the *curse of dimensionality* which essentially states that the issue of density estimation may be intractable in large dimensions. The same statement is true also for regression and classification. Since machine learning essentially deals with high-dimensional data, this issue can be problematic.

Obviously, because the min-max theory is a worst-case analysis, not all situations will be intractable for a given estimator, and some cases that are challenging for one of them may be quite simple for others: even though all estimators are "cursed," the way each of them is cursed differs. Moreover, while many estimators are optimal in the min-max sense, this theory does not give any information on "how often" an estimator performs better than its worst case, or how it will perform on a given class of problems.

Another important point with this curse of dimensionality is that data may very often appear to be high dimensional while it has a simple, low-dimensional structure, maybe because many dimensions are irrelevant to the problem (they contain, for example, just random noise), or because the data supported by a non-linear low-dimensional space, such as a curve or a surface. This information is, of course, not available to the analysis, but can sometimes be inferred using some of the dimension reduction methods that will be discussed later in these notes. Sometimes, and this is also important, information on the data structure can be provided by domain knowledge, that is, by elements provided by experts on how the data has been generated

(such as underlying equations) and reasonable hypotheses that are made in the field. This source of information should never be neglected in practice.

1.2. The EM algorithm

We conclude this chapter with a description of the EM algorithm, which allows one to estimate densities that are modeled by linear combinations of, say, Gaussian densities (Gaussian mixtures). Since the EM has a much broader range of applications, we first describe it in a general setting.

Here is the context. Let $U = (V, H)$ be a pair of random variables taking values in a set M. Let μ_0 be a measure on M, and (π_θ) a family of densities with respect to a product measure $\mu = \mu_V \otimes \mu_H$ (if you are unfamiliar with measure theory, think of $d\mu = dv\, dh$). Assume that a sample $(v_1, \ldots, v_N)$ of V (the visible part of U; H is the hidden, also called latent, part) is observed. Let ψ_θ be the marginal distribution of V when the distribution of U is π_θ. We want to compute the maximum likelihood estimator, $\hat{\theta}$, which maximizes

$$\sum_{k=1}^{N} \log \psi_\theta(v_k).$$

It is generally the case that the computation of the MLE for complete observations, ie. the maximization of

$$\sum_{k=1}^{N} \log \pi_\theta(v_k, h_k).$$

is easy, whereas the problem with the marginal, which is

$$\sum_{k=1}^{N} \log \int_H \pi_\theta(v_k, h) d\mu_H$$

is hard. The EM algorithm iteratively increases the latter using a sequence of optimization problems involving the former.

The key formula is the following: we have

$$\log \psi_\theta(v) = \max_\pi E_\pi \log \left(\frac{\pi_\theta(v, H)}{\pi(H)} \right)$$

the maximum being over all densities π with respect to μ_H. To prove this, introduce the conditional density $\pi_\theta(h|v) = \pi_\theta(v, h)/\psi_\theta(v)$. We have

$$\begin{aligned} (2) \qquad E_\pi \left(\log \frac{\pi_\theta(v, H)}{\pi(H)} \right) &= \int_H \log \frac{\pi_\theta(v, h)}{\pi(h)} \pi(h) d\mu_H(h) \\ &= \int_H \log \frac{\pi_\theta(h|v)\psi_\theta(v)}{\pi(h)} \pi(h) d\mu_H(h) \\ &= \int_H \log \frac{\pi_\theta(h|v)}{\pi(h)} \pi(h) d\mu_H(h) + \log \psi_\theta(v) \\ &= -KL(\pi \| \pi_\theta(.|v)) + \log \psi_\theta(v) \end{aligned}$$

where KL is the Kullback-Leibler divergence. Because it is known that $KL(f\|g)$ is always positive and vanishes only for $f = g$, the result is proved, with the additional fact that the maximum is attained for $\pi = \pi_\theta(.|v)$.

As a consequence, the log-likelihood can be written

$$\sum_{k=1}^{N} \log \psi_\theta(v_k) = \sum_{k=1}^{N} \max_{\pi_k} E_{\pi_k}\left(\log \frac{\pi_\theta(v_k, H)}{\pi_k(H)}\right)$$

and the maximum likelihood is equivalent to computing

$$\max_{\theta, \pi_1, \ldots, \pi_N} \sum_{k=1}^{N} E_{\pi_k} \log\left(\frac{\pi_\theta(v_k, H)}{\pi_k(H)}\right)$$

This can be done by looping over two steps which are: maximize with respect to θ with fixed $\pi_1, \ldots, , \pi_N$, then maximize over the π_k's with fixed θ. The solution of the last problem is already known, since we must have $\pi_k = \pi_\theta(.|v_k)$. Therefore, the EM algorithm is, letting θ_n be the current parameter after loop n:

Compute θ_{n+1} by maximizing $\theta \mapsto \sum_{k=1}^{N} E_{\theta_n}\left(\log \pi_\theta(v_k, H)|V = v_k\right)$
(We have used the fact that $\log \pi_k(H)$ does not depend on θ.)

Let us see now how this framework specializes to mixtures of Gaussian distributions. Such mixtures are densities ψ_θ of the form

$$\psi_\theta(x) = \sum_{j=1}^{p} \frac{\alpha_j}{(2\pi)^{d/2}\sqrt{\det \Sigma_j}} e^{-(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}$$

where θ contains all the parameters: the weights, $\alpha_1, \ldots, \alpha_p$, which are positive and sum to 1, the means, $\mu_1, \ldots, \mu_p$ and the covariance matrices $\Sigma_1, \ldots, \Sigma_p$. Introduce a class variable with values in $H = \{1, \ldots, p\}$ and the joint density function

$$\pi_\theta(x, h) = \frac{\alpha_h}{(2\pi)^{d/2}\sqrt{\det \Sigma_h}} e^{-(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)}$$

For given θ and θ', let

$$U_x(\theta, \theta') = E_\theta\left(\log \pi_{\theta'}(x, H)|X = x\right) + \frac{d}{2}\log 2\pi.$$

We have

$$U_x(\theta, \theta') = \sum_{h=1}^{p} \left(\log \alpha'_h - \frac{1}{2}\log\det \Sigma'_h - (x - \mu'_h)^T {\Sigma'_h}^{-1}(x - \mu'_h)\right) \pi_\theta(h|x)$$

with

$$\pi_\theta(h|x) = \frac{\frac{\alpha_h}{\sqrt{\det \Sigma_h}} e^{-(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)}}{\sum_{j=1}^{p} \frac{\alpha_j}{\sqrt{\det \Sigma_j}} e^{-(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}}$$

If θ_n is the current parameter in the EM, the next one must maximize $\sum_{k=1}^{N} U_{x_k}(\theta_n, \theta')$. This can be solved in close form. For the α'_hs, one must maximize

$$\sum_{k=1}^{N}\sum_{h=1}^{p}(\log \alpha'_h)\pi_{\theta_n}(h|x_k)$$

under the constraint that $\sum_h \alpha'_h = 1$, which yields

$$\alpha'_h = \sum_k \pi(h|x_k) / \sum_{j,k} \pi_{\theta_n}(j|x_k) = Z_h/N$$

with $Z_h = \sum_{k=1}^{N} \pi_{\theta_n}(h|x_k)$.

For μ_h, one minimizes $\sum_{k=1}^{N}(x_k - \mu'_h)^T {\Sigma'_h}^{-1}(x_k - \mu'_h)\pi_{\theta_n}(h|x_k)$ which yields

$$\mu'_h = \frac{1}{Z_h}\sum_{k=1}^{N} x_k \pi(h|x_k)$$

Finally, we get

$$\Sigma'_h = \frac{1}{Z_h}\sum_{k=1}^{N}(x_k - \mu_h)(x_k - \mu_h)^T \pi(h|x_k).$$

CHAPTER 2

The General Regression and Classification Problems

2.1. Regression

Let (Ω, P) denote a probability space. The goal of regression is to find the best prediction of a random variable $Y : \Omega \to \mathbb{R}^q$ (the output) by another random variable $X : \Omega \to \mathcal{R}$ (the input). In these notes, we will most of the time assume that $q = 1$, which will simplify the notation. In this framework, a predictor is a function $f : \mathcal{R}^d \to \mathbb{R}$.

To measure the quality of a predictor, one uses a cost function

$$\begin{array}{rccc} r: & \mathbb{R}\times\mathbb{R} & \to & [0,+\infty) \\ & (y,y') & \mapsto & r(y,y') \end{array}$$

that evaluates the difference between y and y'. The cost, or risk, associated to f is

$$R(f) = E(r(Y, f(X))).$$

It is quite easy to describe the optimal predictor in this context. We will need for this to use conditional expectations and probabilities and proceed first to some reminders of their definitions and properties.

Let $\xi : \Omega \to \mathcal{X}$ and $\eta : \Omega \to \mathcal{Y}$ be two random variables from the probability space Ω to (measurable...) spaces $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y}$, with ξ integrable. The conditional expectation $E(\xi|\eta)$ is a random variable from Ω to $\mathcal{X}$ such that

(i) For any $A \subset \mathcal{X}$, the set $\{\omega : E(\xi|\eta)(\omega) \in A\}$ can be expressed in the form $\{\omega : \eta(\omega) \in B\}$ for some $B \subset \mathcal{Y}$. (Here A and B must both be measurable.)

(ii) For any measurable function $g : \mathcal{Y} \to \mathbb{R}$, one has

$$E[\xi g(\eta)] = E\Big[E(\xi|eta)g(\eta)\Big].$$

Condition (i) implies that the values of $E(\xi|\eta)$ can be expressed by the values of ξ alone. Under mild assumptions, always true in our case, this means that there exists a function $h : \mathcal{Y} \to \mathcal{X}$ such that $E(\xi|\eta) = h(\eta)$ and one often denotes $h(y) = E(\xi|\eta = y)$. It is quite easy to deduce from the second condition the well-known identity

$$E(E(\xi|\eta)) = E(\xi).$$

Moreover, if ξ is square integrable, then $E(\xi|\eta)$ minimizes $E[(\xi-\zeta)^2]$ among all functions $\zeta : \Omega \to \mathcal{X}$ that satisfy (i).

If $A \subset \mathcal{Y}$ and $P(\eta \in A) > 0$, the conditional probability $P(\xi \in B|\eta \in A) = P(\xi \in B \text{ and } \eta \in A)/P(\eta \in A)$ is well defined, as is

$$E(\xi|\eta \in A) = \frac{1}{P(\eta \in A)} E(\xi \chi_{\eta \in A}).$$

If $y \in \mathcal{Y}$, the conditional probability $P(\xi \in B|\eta = y)$ is not so easy to define, because, very often, $P(\eta = y) = 0$ for all y. It is a rather technical result in probability theory that, under mild assumptions on the considered spaces, such a y-indexed family of probability distributions can be defined with the property that, for any function $f : \mathcal{X} \to \mathbb{R}$, $E(f(\xi)|\eta)(\omega)$ coincides with the "usual" integral of $f(\xi)$ for $B \mapsto P(\xi \in B|\eta = \eta(\omega))$. In these notes, we will almost always be in one of the following cases:

(1) $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^k$ with joint p.d.f. f, in which case $B \mapsto P(\xi \in B|\eta = \eta(\omega))$ has p.d.f.

$$f(x|y) = \frac{f(x,y)}{\int_{\mathbb{R}^d} f(x',y)dx'}.$$

(2) $\mathcal{Y}$ is finite, in which case the conditional distribution is easy to define.

Because one can write $R(f) = E[E(r(Y, f(X))|X)]$, R can be minimized by separately minimizing $E(r(Y, f(x))|X = x)$ for each x with respect to the scalar value $f(x)$. The obtained function $x \mapsto \hat{f}(x)$ is called the *Bayes estimator* associated to R. The most common case is for $r(y, y') = |y - y'|^2$, in which case $\hat{f}(x)$ minimizes $t \mapsto E(|y - t|^2|X = x)$ and therefore coincides with the conditional expectation $\hat{f}(x) = E(Y|X = x)$.

Unfortunately, this estimator is not directly available, because the true distribution of (X, Y), or that of Y given X, is unknown. It must be inferred from the available training set: $((x_1, y_1), \ldots, (x_N, y_N))$. This is where a distinction between two different ways to address the problem can be made. The first approach, called model-based, or generative, first tries to estimate (or approximate) the joint density of (X, Y), or the conditional density of X given Y, using techniques of parametric or nonparametric density estimation, and then compute the Bayes estimator with the obtained density. The second approach, called task-oriented, or discriminative, directly targets the risk $R(f)$ by replacing it by its empirical estimate:

$$\mathcal{E}(f) = \frac{1}{N} \sum_{k=1}^{N} r(y_k, f(x_k))$$

and minimizing it with respect to f, with the assumption that f belongs to some parametric class of functions. In this way, the modeling effort is switched from the p.d.f. to the regression function.

Assuming that the input space is Euclidean ($\mathcal{R} = \mathbb{R}^d$), a simple model-based estimate can be based on kernel density estimators (we will later see kernel methods for regression which will address the issue in a completely different way). To estimate the joint p.d.f. of (X, Y), the kernel must take the form $(x, y) \mapsto K(x, y)$. We will here assume that this kernel takes the simpler form $K(x, y) = K_1(x)K_2(y)$, and that the kernel for y is symmetric: $K_1(y) = K_2(-y)$, which is typical. The resulting density estimate at width σ is, in this case:

$$\hat{\varphi}(x, y) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sigma^{d+1}} K_1\left(\frac{x - x_k}{\sigma}\right) K_2\left(\frac{y - y_k}{\sigma}\right)$$

This implies that the conditional expectation (we assume a quadratic loss) is

$$\hat{f}(x) = \frac{\frac{1}{N}\sum_{k=1}^{N} \frac{1}{\sigma^{d+1}} \int_{\mathbb{R}} y K_1\left(\frac{x-x_k}{\sigma}\right) K_2\left(\frac{y-y_k}{\sigma}\right) dy}{\frac{1}{N}\sum_{k=1}^{N} \frac{1}{\sigma^{d+1}} \int_{\mathbb{R}} K_1\left(\frac{x-x_k}{\sigma}\right) K_2\left(\frac{y-y_k}{\sigma}\right) dy}$$

Since K_2 is symmetric, the expectation $\int_{\mathbb{R}} y K_2\left(\frac{y-y_k}{\sigma}\right) dy$ is equal to y_k, and the regression can be written

$$\hat{f}(x) = \frac{\sum_{k=1}^{N} y_k K_1\left(\frac{x-x_k}{\sigma}\right)}{\sum_{k=1}^{N} K_1\left(\frac{x-x_k}{\sigma}\right)}$$

which is the kernel-density estimator for regression.

In the discriminative approach, it is important to understand that the regression function has to be constrained in order to solve the problem. Indeed, if f had an infinite number of free parameters, the minimization of

$$\mathcal{E}(f) = \frac{1}{N} \sum_{k=1}^{N} r(y_k, f(x_k))$$

could be done by solving exactly $y_k = f(x_k)$ for all k (using Lagrange polynomials, for example), so that f would exactly adapt to the training set. Since the training data can be noisy, and provides only partial information on X and Y, this exact fit is undesirable, and likely to provide a very large risk (the obtained estimator has zero bias, but inevitably a huge variance). There are essentially two ways to add constraints to f: model it using a finite number of parameters, or enforce some smoothness properties (controlling, for example the L^2-norm of its derivatives). A typical example of the first case (we will discuss examples of the second approach later in the notes) is the linear model: f is assumed to belong to $\mathcal{F}$, with

$$\mathcal{F} = \left\{ f : f(x) = \beta_0 + \sum_{i=1}^{d} \beta_i x(i), \beta_0, \ldots, \beta_d \in \mathbb{R} \right\}$$

where $x(1), \ldots, x(d)$ are the d coordinates of x. A more complex model can involve two layers:

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p} w_j \psi\Big(\beta_{j0} + \sum_{i=1}^{d} \beta_{ji} x(i)\Big), w_j, \beta_{ji} \in \mathbb{R} \right\}$$

with a fixed function ψ, which corresponds to standard models of neural networks (these examples of course assume that $\mathcal{R} = \mathbb{R}^d$).

Let us illustrate again the bias-variance dilemma in this regression context, using a quadratic loss function. Let f_* be an optimal estimator of $E(Y|X)$ in $\mathcal{F}$, and $\hat{f}_N$ be the minimizer of the empirical risk over $\mathcal{F}$. The risk associated to $\hat{f}_N$ can be written

$$\begin{aligned} R(\hat{f}_N) &= E(|Y - \hat{f}_N(X)|^2) \\ &= E(|Y - f_*(X)|^2) + E(|\hat{f}_N(X) - f_*(X)|^2) \\ &+ 2E((Y - f_*(X))(f_*(X) - \hat{f}_N(X)) \end{aligned}$$

In many cases, the last term is zero. This is true, for example, if there exists a family of functions (f_λ), for small enough λ in a neighborhood of 0, such that $f_0 = f_*$, $f_\lambda \in \mathcal{F}$ and $g := df_\lambda/d\lambda$ exists at $\lambda = 0$ (i.e., one can find a differentiable curve in $\mathcal{F}$ containing f_*). Then $R(f_\lambda)$ is minimal at $\lambda = 0$, and its derivative, which is $-2E((Y - f_*(X))g(X))$, is equal to 0. So, we will have $E((Y - f_*(X))(f_*(X) - \hat{f}_N(X))) = 0$ as soon as a f_λ can be found with $g = f_*(X) - \hat{f}_N(X)$. This is true in particular if the class $\mathcal{F}$ is linear,[1] since in this case, one can take $f_\lambda = f_* + \lambda(\hat{f}_N - f_*)$.

Assuming this, we have

$$R(\hat{f}_N) = E(|Y - f_*(X)|^2) + E(|\hat{f}_N(X) - f_*(X)|^2)\,.$$

The first term is the error associated to the best approximation of Y by some $f(X)$ for $f \in \mathcal{F}$. This is the bias. The second is the error due to using the empirical loss to estimate $\hat{f}_*$, and this is the variance. As before, the bias is typically small for "large" function classes, but then, the variance becomes larger.

2.2. Classification

The context of classification is exactly the same as for regression, except that the predicted variable, Y, is qualitative, taking values in a finite set, $\mathcal{G}$, of categories or classes. The classifier is therefore a function $f : \mathcal{R} \to \mathcal{G}$. The difference with the previous case is that loss functions used with quantitative variables are not adapted to classification (if the class labels

[1] A warning on the term linear: stating that the class $\mathcal{F}$ is linear does not mean that it consists of linear functions of x, but that it is stable by linear combination: if $f, g \in \mathcal{F}$, then $\alpha f + \beta g \in \mathcal{F}$ for all $\alpha, \beta \in \mathbb{R}$.

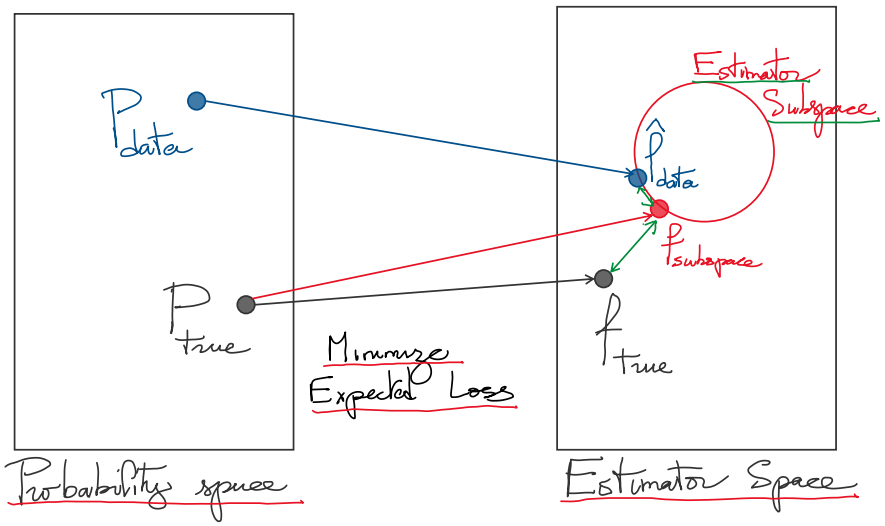

FIGURE 1. Statistical Learning: When P_{true} is the distribution of the data, the true estimator f_{true} minimizes the expected loss function. Based on data $Z_1, \ldots, Z_N$, the sample-based distribution is $(\delta_{Z_1} + \cdots + \delta_{Z_N})/N$ and the empirical loss is minimized over a subset $\mathcal{S}$ of the space of all possible estimators. The expected discrepancy between the resulting estimator and the one minimizing the true expected loss on the subspace is the "variance" of the method, and the expected discrepancy between this subspace-constrained estimator and and the optimal one is the "bias."

represent vehicles, for example, the expression $(\text{car} - \text{truck})^2$ is meaningless). The most commonly used function in this context is the 0-1 loss

$$r(y, y') = 1 \text{ if } y \neq y', \text{ and 0 otherwise.}$$

Using this loss function yields the average risk $R(f) = 1 - P(Y = f(X))$ so that optimal f have to maximize the rate of good classification: $P(Y = f(X))$. Similar to the case of regression, we can write $P(Y = f(X)) = E(P(Y = f(X)|X))$ and the optimal estimator must satisfy

$$f_*(x) = \underset{c}{argmax}\ P(Y = c|X = x)$$

This provides the maximum *a posteriori* (MAP) estimator given by the mode of the posterior distribution, which is the conditional distribution of the output given the observation. This is also called the Bayes estimator. But it only has a theoretical use, since the conditional distribution of Y given X is unknown. Similar to regression, generative and discriminative

approaches exists, as well as a bias-variance trade-off. Many related issues will be discussed in the remaining of this course.

2.3. Measuring the error

2.3.1. Prediction error. In this short section, we discuss quantitative measures of the quality of a model. These are various definitions of *errors* that can be either theoretical or empirical.

Let Y denote the target variables (either quantitative or qualitative) and X be the input. For a function $f : x \mapsto y = f(x)$, the prediction (or generalization) error is

$$\Delta(f) = E(r(Y, f(X)))$$

where r is the loss function. In this expectation, both Y and X are assumed to be random.

Of course, since the distribution of (Y, X) is unknown (otherwise the best estimator is Bayes and the problem is solved), $\Delta(f)$ cannot be computed. The whole issue of machine learning is, however, still to control this error, because this is the one that one would ultimately like to be small.

Assume f is estimated from a training set T, and let $\hat{f}_T$ denote the estimator. We will consider T as a random variable, a collection of N i.i.d. repeats of the joint distribution of (X, Y), so that $T = ((X_1, Y_1), \dots, (X_N, Y_N))$. This implies that $\Delta(\hat{f}_T)$ is also a random variable (it depends on the training set), which is often emphasized by writing

$$\Delta(\hat{f}_T) = E(r(Y, \hat{f}_T(X))|T)$$

where (Y, X) are independent of T (interpreted as a random new observation, which justifies the term of generalization error for Δ). The expectation $\bar{\Delta} = E(\Delta(\hat{f}_T))$ is the error averaged over all possible training sets.

Many learning algorithms optimize an empirical version of Δ, namely, given observations $(x_1, y_1), \dots, (x_N, y_N)$, they minimize

$$\hat{\Delta}(f) = \frac{1}{N} \sum_{k=1}^{N} r(y_k, f(x_k)) \,.$$

Whether or not an estimator $\hat{f}_T$ is designed as a minimizer of $\hat{\Delta}$, its associated *in-sample* error is defined by

$$\mathcal{E}_T = \hat{\Delta}(\hat{f}_T) = \frac{1}{N} \sum_{k=1}^{N} r(y_k, \hat{f}_T(x_k)).$$

This can be a poor estimation of the true error, Δ. Notice that $\mathcal{E}_T$ is not an average of independent terms, since in each term, the function $\hat{f}_T$ depends on the whole training set. As a consequence, the law of large numbers does not apply to it. Typically, the training set error underestimates (sometimes dramatically) the true generalization error. Unless properly corrected, it

should not be used, in general, to assess the performance of an algorithm, or the generalization error of the obtained predictor.

A more sensible way to estimate Δ is to use "test data", in which an additional test set $T' = ((Y'_1, X'_1), \ldots, (Y'_{N'}, X'_{N'}))$ is used and the estimation of the error is

$$\mathcal{E}_{T,T'} = \frac{1}{N'} \sum_{k=1}^{N'} r(y'_k, \hat{f}_T(x'_k)).$$

which in this case converges to $\Delta(\hat{f}_T)$ when N' tends to infinity. This estimator corresponds to a relatively wealthy situation in which one can spare a part of the available data for testing.

2.3.2. In-sample error. Both variables X and Y are assumed to be random in this setting, but there are often situations when one of them is "more random" than the other. Randomness on Y is associated to measurement errors, or ambiguity in the decision. Randomness in X more generally relates to the issue of sampling a dataset in a large dimensional space. In some cases, Y is not random at all: for example, in object recognition, the question "is this a pipe?" in Fig. 2 has a deterministic answer.

Sometimes, X is small dimensional and densely sampled, but Y is subject to noise. In such cases (which are unfortunately not the most interesting ones in modern machine learning), one can define the in-sample generalization error given by

$$\mathcal{E}'(T) = \frac{1}{N} \sum_{k=1}^{N} E(r(Y', \hat{f}_T(x_k))|T, X = x_k).$$

In this expectation, T (and therefore $\hat{f}_T$) is fixed, and the expectation is over Y', which follows the conditional distribution of Y given $X = x_k$. This corresponds to averaging over new training sets, when one keeps the same input variables, but generates new outputs.

Consider the case of a square error $r(y, y') = (y - y')^2$. We have

$$\begin{aligned}
\mathcal{E}'(T) &= \frac{1}{N} \sum_{k=1}^{N} E((Y'_k - \hat{f}_T(X_k))^2|T) \\
&= \frac{1}{N} \sum_{k=1}^{N} E((Y'_k - Y_k + Y_k - \hat{f}_T(X_k))^2|T) \\
&= \mathcal{E}(T) - \frac{2}{N} \sum_{k=1}^{N} E((Y'_k - Y_k)\hat{f}_T(X_k)|T) + \frac{1}{N} \sum_{k=1}^{N} E((Y'_k)^2 - Y_k^2|T)
\end{aligned}$$

FIGURE 2. The treachery of images, by Magritte. The text says "this is not a pipe". (Reproduced without permission.)

Write $T = (\mathcal{X}, \mathcal{Y})$ to separate the predictor and predicted variables in the training set. We compute the "expected optimism," given by

$$E(\mathcal{E}'(T) - \mathcal{E}(T)|\mathcal{X}) = -\frac{2}{N}\sum_{k=1}^{N} E((Y'_k - Y_k)\hat{f}_T(X_k)|\mathcal{X}) + \frac{1}{N}\sum_{k=1}^{N} E((Y'_k)^2 - Y_k^2|\mathcal{X})$$

(so that expectation is taken for the conditional distribution of $\mathcal{Y}$ given $\mathcal{X}$). The last term vanishes because Y and Y' have the same distribution given $\mathcal{X}$. We have

$$\begin{aligned} E((Y'_k - Y_k)\hat{f}_T(X_k)|\mathcal{X}) &= E(Y'_k|\mathcal{X})E(\hat{f}_T(X_k(|X) - E(Y_k\hat{f}_T(X_k)|\mathcal{X}) \\ &= -\text{cov}(Y_k, \hat{f}_T(X_k)|\mathcal{X}) \end{aligned}$$

in which we have used the fact that Y'_k and $\mathcal{Y}$ are conditional independent given $\mathcal{X}$ and that Y'_k and Y_k have the same distribution, still conditional to $\mathcal{X}$. We therefore find

$$E(\mathcal{E}'(T) - \mathcal{E}(T)|\mathcal{X}) = \frac{2}{N}\text{cov}(Y_k, \hat{f}_T(X_k)|\mathcal{X}).$$

We will see in Chapter 12 how the expected optimism can be used to penalize the training error and often improve the reliability of the estimators.

2.3.3. Cross validation. In many applications, there is not enough data available to afford sparing part of it for a test set. In such cases, cross validation is a useful alternative. The n-fold cross validation method separates the training set into n non-overlapping sets of equal sizes, and estimates n classifiers (or regressions) by leaving out one of these subsets as a test set. A generalization error is estimated from the test set and averaged over the n results.

The limit case is when $n = N$, the number of training samples: this is the leave-one-out cross validation, which can be quite computationally intensive if the classifiers are hard to train. The issue also of the leave-one-out method

is that the N classifiers are quite correlated, which implies that the variance of the average (over random training sets) can be large (certainly not scale like $1/N$). The issue with n-fold validation when n is small is that the the error is estimated from classifiers with $100/n\%$ less data, which can bias the result (typically over-estimate the error), especially when N is small. A balance has to be found, obviously, a reasonable choice being often provided by n around 10.