Machine Learning — Statistical Methods for Machine Learning

Statistical Learning

Instructor: *Nicolò Cesa-Bianchi* version of March 11, 2019

In order to analyze a learning algorithm, we must define a mathematical model of how examples $(\boldsymbol{x}, y)$ are generated. In the statistical learning framework, we assume that every example $(\boldsymbol{x}, y)$ is obtained through an independent draw from a fixed but unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. We often write $(\boldsymbol{X}, Y)$ to highlight that $\boldsymbol{x}$ and y are random variables. The assumption that not all data points $\boldsymbol{x}$ are equally likely is quite natural (for example, when data points are images, only a small fraction of all possible pixel configurations correspond to real-world images). Similarly, as we previously argued labels are typically noisy, and thus a probability distribution is a good mathematical abstraction to explain the variety of labels which a same data point may be associated with.

As we just explained, in statistical learning every example $(\boldsymbol{x}, y)$ is the realization of an independent random draw from the same joint probability distribution $\mathcal{D}$. This implies that every dataset (e.g., a training set or a test set) is a **random sample**, of the kind we study in statistics. Note that the independence assumption is actually violated in many practical domains. Consider for example the problem of categorizing news stories. The newsfeed is clearly far from being an independent process, as the evolution of news reflects developing and related stories. Although not very realistic, the independence assumption is nevertheless convenient from the viewpoint of the analytical tractability of the problem.

In statistical learning, a problem is fully specified by a pair $(\mathcal{D}, \ell)$, where $\mathcal{D}$ is the data distribution and ℓ is a loss function. The performance of a predictor $h : \mathcal{X} \to \mathcal{Y}$ with respect to $(\mathcal{D}, \ell)$ is evaluated via the **statistical risk**, defined by

$$\ell_{\mathcal{D}}(h) = \mathbb{E}\big[\ell(Y, h(\boldsymbol{X}))\big]$$

This is the expected value of the loss function on a random example $(\boldsymbol{X}, Y)$ drawn from $\mathcal{D}$. The best possible predictor $f^* : \mathcal{X} \to \mathcal{Y}$ given $\mathcal{D}$ is known as **Bayes optimal predictor**, and is defined by

$$f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$$

Note that $f^*(x)$ is the prediction minimizing the conditional risk, which is the expected loss of the prediction with respect to the distribution of the label Y conditioned on $\boldsymbol{x}$. By definition of f^*, we have that

$$\mathbb{E}\Big[\ell\big(Y, f^*(x)\big) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] \leq \mathbb{E}\Big[\ell\big(Y, h(x)\big) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big]$$

for every predictor $h : \mathcal{X} \to \mathcal{Y}$. Because the above inequality holds for every $\boldsymbol{x} \in \mathcal{X}$, it also holds in expectation with respect to the random draw of $\boldsymbol{X}$. But since the average of the conditional risk is the risk, we have that $\ell_{\mathcal{D}}(f^*) \leq \ell_{\mathcal{D}}(h)$ for every predictor h. The risk $\ell_{\mathcal{D}}(f^*)$ of the Bayes optimal predictor is called **Bayes error**. Typically, the Bayes error is larger than zero because labels are stochastic.

We now compute the Bayes optimal predictor for the quadratic loss function $\ell(y, \widehat{y}) = (y - \widehat{y})^2$ when $\mathcal{Y} \equiv \mathbb{R}$,

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \mathbb{E}\big[(Y - \widehat{y})^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \Big(\mathbb{E}\big[Y^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] + \widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \Big) \\
&= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \Big(\widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \Big) \qquad \text{(ignoring the term that does not depend on } \widehat{y}) \\
&= \mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \qquad \text{(minimizing the function } F(\widehat{y}) = \widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]\text{)}
\end{aligned}$$

Thus, the Bayes optimal prediction for the quadratic loss function is the expected value of the label conditioned on the instance.

Substituting in the conditional risk formula $\mathbb{E}\big[\big(Y - f^*(\boldsymbol{X})\big)^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$ the Bayes optimal predictor $f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ we obtain

$$\mathbb{E}\Big[\big(Y - f^*(\boldsymbol{X})\big)^2 \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid \boldsymbol{x}]\big)^2 \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = \operatorname{Var}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \ .$$

In words, the conditional risk of the Bayes optimal predictor for the quadratic loss is the variance of the label conditioned on the instance.

We now focus on binary classification, where $\mathcal{Y} = \{-1, +1\}$. In this case we denote the joint distribution of $(\boldsymbol{X}, Y)$ with the pair (D_X, η), where D_X is the marginal of $\mathcal{D}$ on $\mathcal{X}$ and η is the distribution of $Y \in \{-1, +1\}$ conditioned on $\mathcal{X}$. We view the distribution η as a function $\eta : \mathcal{X} \to [0, 1]$, where $\eta(\boldsymbol{x}) = \mathbb{P}\big(Y = +1 \mid \boldsymbol{X} = \boldsymbol{x}\big)$ is the probability that $\boldsymbol{x}$ occurs with label $+1$.

Let $\mathbb{I}\{A\} \in \{0, 1\}$ be the indicator function of an event A; that is, $\mathbb{I}\{A\} = 1$ if and only if A occurs. The statistical risk with respect to the zero-one loss $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ is therefore defined by

$$\ell_{\mathcal{D}}(h) = \mathbb{E}\big[\ell(Y, h(\boldsymbol{X}))\big] = \mathbb{E}\big[\mathbb{I}\{h(\boldsymbol{X}) \neq Y\}\big] = \mathbb{P}\big(h(\boldsymbol{X}) \neq Y\big) \ .$$

In words, the risk of h is the probability that h misclassifies a random $\boldsymbol{X}$ drawn from the distribution D_X and having a random label Y drawn from the distribution $\{1 - \eta(\boldsymbol{X}), \eta(\boldsymbol{X})\}$ on $\{-1, +1\}$.

The Bayes optimal predictor $f^* : \mathcal{X} \to \{-1, +1\}$ for binary classification is derived as follows

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \mathbb{E}\big[\mathbb{I}\{Y = +1\}\mathbb{I}\{\widehat{y} = -1\} + \mathbb{I}\{Y = -1\}\mathbb{I}\{\widehat{y} = +1\} \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \Big(\mathbb{P}(Y = +1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{I}\{\widehat{y} = -1\} + \mathbb{P}(Y = -1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{I}\{\widehat{y} = +1\} \Big) \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \Big(\eta(\boldsymbol{x})\mathbb{I}\{\widehat{y} = -1\} + \big(1 - \eta(\boldsymbol{x})\big)\mathbb{I}\{\widehat{y} = +1\} \Big) \\
&= \begin{cases} -1 & \text{se } \eta(\boldsymbol{x}) < 1/2, \\ +1 & \text{se } \eta(\boldsymbol{x}) \geq 1/2. \end{cases}
\end{aligned}$$

Hence, the Bayes optimal classifier predicts the label whose probability is the highest when conditioned on the instance. Finally, it is easy to verify that the Bayes error in this case is $\ell_{\mathcal{D}}(f^*) = \mathbb{E}\big[\min\{\eta(\boldsymbol{X}),\, 1 - \eta(\boldsymbol{X})\}\big]$.

Bounding the risk. Next, we study the problem of bounding the risk of a predictor. From now on, we assume $\ell(y, \widehat{y}) \in [0, 1]$. However, keep in mind that our analysis continues to hold also when $\ell(y, \widehat{y}) \in [0, M]$ for any $M > 0$.

It should be clear that, given an arbitrary predictor h, we cannot directly compute its risk $\ell_{\mathcal{D}}(h)$ with respect to $\mathcal{D}$ because $\mathcal{D}$ is typically unknown (if we knew $\mathcal{D}$, we could directly construct the Bayes optimal predictor). We thus consider the problem of estimating the risk of a given predictor h. In order to compute this estimate, we use a so-called **test set** $S' = \big\{(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_n, y'_n)\big\}$ of examples. Then, we estimate $\ell_{\mathcal{D}}(h)$ with the **test error**, which is the average loss of h on the test set,

$$\widehat{\ell}_{S'}(h) = \frac{1}{n} \sum_{t=1}^{n} \ell\big(y'_t, h(\boldsymbol{x}'_t)\big) \ .$$

Under the assumption that the test set is generated through independent draws from $\mathcal{D}$, the test error corresponds to the sample mean of the risk. Indeed, for each $t = 1, \ldots, n$ the example $(\boldsymbol{X}'_t, Y'_t)$ is an independent draw from $\mathcal{D}$. Therefore,

$$\mathbb{E}\Big[\ell\big(Y'_t, h(\boldsymbol{X}'_t)\big)\Big] = \ell_{\mathcal{D}}(h) \ .$$

In order to compute a confidence interval for the risk using the test error, we can use the following result about the law of large numbers.

Lemma 1 (Chernoff-Hoeffding)**.** *Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables with expectation μ and such that $0 \leq Z_t \leq 1$ for each $t = 1, \ldots, n$. Then, for any given $\varepsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t > \mu + \varepsilon\right) \leq e^{-2\varepsilon^2 n} \qquad \textit{and} \qquad \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t < \mu - \varepsilon\right) \leq e^{-2\varepsilon^2 n} \ .$$

Using the Chernoff-Hoeffding bound with $Z_t = \ell(y_t, h(x_t)) \in [0, 1]$ we can compute a confidence interval for the risk as follows

$$\mathbb{P}\Big(\big|\ell_{\mathcal{D}}(h) - \widehat{\ell}(h)\big| > \varepsilon\Big) = \mathbb{P}\Big(\ell_{\mathcal{D}}(h) - \widehat{\ell}(h) > \varepsilon\Big) + \mathbb{P}\Big(\widehat{\ell}(h) - \ell_{\mathcal{D}}(h) > \varepsilon\Big) \leq 2\, e^{-2\varepsilon^2 n} \tag{1}$$

where the probability is computed with respect to random draw of the test set. This inequality shows that the probability that a test set gives a test error $\widehat{\ell}_{S'}(h)$ differing from the true risk $\ell_{\mathcal{D}}(h)$ for more than ε quickly decreases with the size n of the test set.

More specifically, by setting to δ the right-hand side of (1), for any given $\delta \in (0, 1)$, we get that

$$\big|\ell_{\mathcal{D}}(h) - \widehat{\ell}_{S'}(h)\big| \leq \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$$

holds with probability al least $1 - \delta$ with respect to the random draw of the test set.

The inequality (1) is telling us how to use a test set to estimate the risk of a classifier. More precisely, the inequality shows that the test set, which is how we measure in practice the performance of a classifier on unseen data, is close to the statistical risk with high probability.

In order to understand the occurrence of overfitting, we now study a specific learning algorithm in the statistical framework we just defined. As usual, we assume the algorithm obtains a training set $S = \big\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\big\}$, in input, where $(\boldsymbol{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$. The algorithm generates a predictor $h : \mathcal{X} \to \mathcal{Y}$ belonging to a given model space $\mathcal{H}$.

If the algorithm is forced to pick a predictor from $\mathcal{H}$, then the best thing the algorithm can do is to choose the best possible predictor $h^* \in \mathcal{H}$, that is the one satisfying

$$\ell_{\mathcal{D}}(h^*) = \min_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) \ .$$

Thanks to the law of large numbers, we know that the training error $\widehat{\ell}_S(h^*)$ is close to $\ell_{\mathcal{D}}(h^*)$ with high probability with respect to the random draw of the training set on which $\widehat{\ell}_S(h^*)$ is computed.

Consider the algorithm choosing $\widehat{h} \in \mathcal{H}$ minimizing the training error,

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{\ell}_S(h) \ .$$

This algorithm, which is very natural in many applications, is known as Empirical Risk Minimizer (ERM).

Unfortunately, we cannot directly apply the Chernoff-Hoeffding bound to $\widehat{h}$ to show that $\ell_{\mathcal{D}}(\widehat{h})$ is close to $\widehat{\ell}_S(\widehat{h})$. The reason is that $\widehat{h}$ is a function of the training set, and thus a random variable. Chernoff-Hoeffding ensures that $\widehat{\ell}_S(h)$ is close to $\ell_{\mathcal{D}}(h)$ for any fixed h, whereas $\widehat{h}$ is not fixed as it depends on the sample.

In order to analyze the risk of $\widehat{h}$, we then proceeds as follows:

$$\begin{aligned} \ell_{\mathcal{D}}(\widehat{h}) = {} & \ell_{\mathcal{D}}(\widehat{h}) - \ell_{\mathcal{D}}(h^*) && \text{variance error} \\ & + \ell_{\mathcal{D}}(h^*) - \ell_{\mathcal{D}}(f^*) && \text{bias error} \\ & + \ell_{\mathcal{D}}(f^*) && \text{Bayes error} \end{aligned}$$

where f^* is the Bayes optimal predictor for $\mathcal{D}$. The Bayes error is not controllable, because it only depends on $\mathcal{D}$ and the loss ℓ. The bias error arises because $\mathcal{H}$ does not necessarily contain the Bayes optimal predictor. The variance error arises because the risk of $\widehat{h}$ is generally different from the training error of $\widehat{h}$ (which is the smallest in $\mathcal{H}$ by design of the algorithm).

Next, we bound the variance error. For every given training set S, we have that

$$\begin{aligned} \ell_{\mathcal{D}}(\widehat{h}) - \ell_{\mathcal{D}}(h^*) &= \ell_{\mathcal{D}}(\widehat{h}) - \widehat{\ell}_S(\widehat{h}) + \widehat{\ell}_S(\widehat{h}) - \ell_{\mathcal{D}}(h^*) \\ &\leq \ell_{\mathcal{D}}(\widehat{h}) - \widehat{\ell}_S(\widehat{h}) + \widehat{\ell}_S(h^*) - \ell_{\mathcal{D}}(h^*) \\ &\leq \big|\ell_{\mathcal{D}}(\widehat{h}) - \widehat{\ell}_S(\widehat{h})\big| + \big|\widehat{\ell}_S(h^*) - \ell_{\mathcal{D}}(h^*)\big| \\ &\leq 2 \max_{h \in \mathcal{H}} \big|\widehat{\ell}_S(h) - \ell_{\mathcal{D}}(h)\big| \end{aligned}$$

where we used the assumption that $\widehat{h}$ minimizes $\widehat{\ell}_S(h)$ among all $h \in \mathcal{H}$. Therefore, for all $\varepsilon > 0$,

$$\ell_{\mathcal{D}}(\widehat{h}) - \ell_{\mathcal{D}}(h^*) > \varepsilon \quad \Rightarrow \quad \max_{h \in \mathcal{H}} \big|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\big| > \frac{\varepsilon}{2} \quad \Rightarrow \quad \exists h \in \mathcal{H} : \ \big|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\big| > \frac{\varepsilon}{2} \ .$$

Since the above chain of implications holds for any realization of the training set, we can write

$$\mathbb{P}\left(\ell_{\mathcal{D}}(\widehat{h}) - \ell_{\mathcal{D}}(h^*) > \varepsilon\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right) \ .$$

We now study the case $|\mathcal{H}| < \infty$, that is when the model space contains a finite number of predictors. Note that the event

$$\exists h \in \mathcal{H} : \ \left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}$$

is the union over $h \in \mathcal{H}$ of the (not necessarily disjoint) events

$$\left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}$$

We may then use the union bound

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

which holds for any collection $A_1, \ldots, A_n$ of events. By doing so we get

$$\begin{aligned}
\mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right) &= \mathbb{P}\left(\bigcup_{h \in \mathcal{H}} \left(\left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right)\right) \\
&\leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right) \\
&\leq |\mathcal{H}| \max_{h \in \mathcal{H}} \mathbb{P}\left(\left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right) \\
&\leq |\mathcal{H}| 2e^{-m\varepsilon^2/2} \qquad (2)
\end{aligned}$$

where in the last step we used the Chernoff-Hoeffding bound.

In conclusion, we have that

$$\mathbb{P}\left(\ell_{\mathcal{D}}(\widehat{h}) - \ell_{\mathcal{D}}(h^*) > \varepsilon\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| > \frac{\varepsilon}{2}\right) \leq 2|\mathcal{H}|e^{-m\varepsilon^2/2} \ . \qquad (3)$$

Setting the left-hand side of (3) equal to δ and solving for ε we obtain that

$$\ell_{\mathcal{D}}(\widehat{h}) \leq \ell_{\mathcal{D}}(h^*) + \sqrt{\frac{2}{m} \ln \frac{2|\mathcal{H}|}{\delta}}$$

holds with probability at least $1 - \delta$ with respect to the random draw of a training set of cardinality m.

Note that the risk of $\widehat{h}$ is bounded through a sum of two terms: $\ell_{\mathcal{D}}(h^*)$ and $\sqrt{\frac{2}{m} \ln \frac{2|\mathcal{H}|}{\delta}}$. Lacking further information on $\mathcal{D}$, and for a given cardinality m of the training set, in order to decrease $\sqrt{\frac{2}{m} \ln \frac{2|\mathcal{H}|}{\delta}}$, which is our bound on the variance error, we must decrease $|\mathcal{H}|$. But decreasing $|\mathcal{H}|$ might cause an increase of $\ell_{\mathcal{D}}(h^*)$, which produces a corresponding increase of the bias error. In light of this statistical analysis, we may say that overfitting in the ERM algorithm is caused by an unbalance between the variance error and the bias error. In particular, overfitting is when the

variance error is much larger than the bias error, and underfitting is when the bias error is much larger than the variance error.

In the proof of the bound on the variance error, we have also shown in (2) that

$$\forall h \in \mathcal{H} \quad \left|\widehat{\ell}(h) - \ell_{\mathcal{D}}(h)\right| \leq \sqrt{\frac{1}{2m} \ln \frac{2|\mathcal{H}|}{\delta}}$$

with probability at least $1 - \delta$ with respect to the random draw of the training set. This implies that when the cardinality of the training set is sufficiently large with respect to $\ln |\mathcal{H}|$, then the training error $\widehat{\ell}_S(h)$ becomes a good estimate for the statistical risk $\ell_{\mathcal{D}}(h)$ *simultaneously* for all predictors $h \in \mathcal{H}$. In this regime, namely when the law of large numbers holds uniformly with respect to the choice of $h \in \mathcal{H}$, it is clear that any algorithm ranking the predictors in $\mathcal{H}$ according to their training error is protected from overfitting.