Machine Learning — Statistical Methods for Machine Learning

Statistical Learning

Instructor: *Nicolò Cesa-Bianchi* version of March 25, 2024

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 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. This corresponds to assuming that labels of any given datapoint are stochastic.

Assuming that every example $(\boldsymbol{x}, y)$ is the realization of an independent random draw from the same joint probability distribution $\mathcal{D}$ implies that every dataset (e.g., a training set or a test set) is a statistical sample. 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, and works suprisingly well in practice.

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

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

The quantity $\mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$ is the conditional risk, which is the expected loss of the prediction with respect to the distribution of the label Y conditioned on $\boldsymbol{x}$. Hence $f^*(x)$ is the prediction minimizing the conditional risk. By definition of f^*, we have that

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

for every predictor $h : \mathcal{X} \to \mathcal{Y}$ and for any $\boldsymbol{x} \in \mathcal{X}$. 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, for any predictor h,

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

we have that $\ell_{\mathcal{D}}(f^*) \le \ell_{\mathcal{D}}(h)$ for every predictor h. The risk $\ell_{\mathcal{D}}(f^*)$ of the Bayes optimal predictor is called **Bayes risk**. Typically, the Bayes risk 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 $\mathcal{Y} \equiv \mathbb{R}$,

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

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. By averaging over $\boldsymbol{X}$ we obtain $\ell_{\mathcal{D}}(f^*) = \mathbb{E}\big[\operatorname{Var}[Y \mid \boldsymbol{X}]\big]$. Namely, the Bayes risk for the quadratic loss is the expected conditional variance of the label. Note that $\mathbb{E}\big[\operatorname{Var}[Y \mid \boldsymbol{X}]\big]$ is generally different from $\operatorname{Var}[Y]$. Indeed, the law of total variance says that $\operatorname{Var}[Y] - \mathbb{E}\big[\operatorname{Var}[Y \mid \boldsymbol{X}]\big] = \operatorname{Var}\big[\mathbb{E}[Y \mid \boldsymbol{X}]\big]$.

We now focus on binary classification, where $\mathcal{Y} = \{-1,1\}$. Let $\eta(\boldsymbol{x})$ be the probability of $Y=1$ conditioned on $\boldsymbol{X}=\boldsymbol{x}$. We view $\eta(\boldsymbol{x}) = \mathbb{P}\big(Y=+1 \mid \boldsymbol{X}=\boldsymbol{x}\big)$ as the value on $\boldsymbol{x}$ of a function $\eta : \mathcal{X} \to [0,1]$.

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

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

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

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

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 risk 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 can use the **test set** $S' = \big\{(\boldsymbol{x}_1', y_1'), \dots, (\boldsymbol{x}_n', y_n')\big\}$. We can then estimate $\ell_{\mathcal{D}}(h)$ with the **test error**, which is the average loss of h on the test set,

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

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 $t = 1, \dots, n$ the example $(\boldsymbol{X}_t', Y_t')$ is an independent draw from $\mathcal{D}$. Therefore,

$$\mathbb{E}\Big[\ell\big(Y_t', h(\boldsymbol{X}_t')\big)\Big] = \ell_{\mathcal{D}}(h) \qquad t = 1, \dots, n$$

Note that the above equalities rely on the assumption that h does not depend on the test set. If it did, then the above equalities would not be necessarily true. This fact is important in the analysis of learning algorithms.

In order to compute how good is the test error as an estimate for the risk, we can use the following result about the law of large numbers.

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

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

In the rest of this course, we repeatedly use the following facts:

- For any two events A and B, if $A \Rightarrow B$, then $\mathbb{P}(A) \le \mathbb{P}(B)$
- (Union bound) For any collection $A_1, \dots, A_n$ of (not necessarily disjoint) events,

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

  If the events $A_1, \dots, A_n$ are pairwise disjoint, then the union bound holds with equality.

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 (where the test error is written as ℓ instead of $\ell_{S'}$),

$$\begin{aligned} \mathbb{P}\Big(\big|\ell_{\mathcal{D}}(h) - \ell(h)\big| > \varepsilon\Big) &= \mathbb{P}\Big(\ell_{\mathcal{D}}(h) - \ell(h) > \varepsilon \ \cup \ \ell(h) - \ell_{\mathcal{D}}(h) > \varepsilon\Big) \\ &= \mathbb{P}\Big(\ell_{\mathcal{D}}(h) - \ell(h) > \varepsilon\Big) + \mathbb{P}\Big(\ell(h) - \ell_{\mathcal{D}}(h) > \varepsilon\Big) \le 2\, e^{-2\varepsilon^2 n} \qquad (1) \end{aligned}$$

where in the last step we applied the union bound to the disjoint events $\ell_{\mathcal{D}}(h) - \ell(h) > \varepsilon$ and $\ell(h) - \ell_{\mathcal{D}}(h) > \varepsilon$. Note that the probability is computed with respect to the random draw of the test set. This inequality shows that the probability that a test set gives a test error $\ell_{S'}(h)$ 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: if we set to $\delta \in (0,1)$ the right-hand side of (1) and then solve for ε, we get that

$$\left|\ell_{\mathcal{D}}(h) - \ell_{S'}(h)\right| \le \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$$

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 error, which is how we measure in practice the performance of a classifier on unseen data, is close to the statistical risk with high probability.

Overfitting and underfitting. Fix a learning problem $(\mathcal{D}, \ell)$ and consider a generic learning algorithm A. In the following, we write $A(S)$ to denote the predictor output by A when given the training set S. Let $\mathcal{H}_m$ be the set of predictors generated by A on training sets of size m: $h \in \mathcal{H}_m$ if and only if there exists a training set S of size m such that $A(S) = h$. For example, if A is an algorithm for training a neural network, then $\mathcal{H}_m$ is the set of predictors obtained by training the neural network using training sets of fixed size m. Let h^* be any predictor with minimum risk $\ell_{\mathcal{D}}(h^*)$ in $\mathcal{H}_m$. That is,

$$h^* \in \operatorname*{argmin}_{h \in \mathcal{H}_m} \ell_{\mathcal{D}}(h)$$

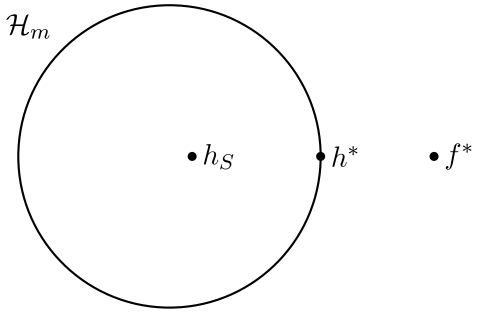

Figure 1: A visual explanation of the bias-variance decomposition.

Fix a training set S of size m and let $h_S = A(S) \in \mathcal{H}_m$. The following is called the **bias-variance** decomposition:

$$\begin{aligned} \ell_{\mathcal{D}}(h_S) = \ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*) & \qquad \text{estimation (or variance) error} \\ + \ell_{\mathcal{D}}(h^*) - \ell_{\mathcal{D}}(f^*) & \qquad \text{approximation (or bias) error} \\ + \ell_{\mathcal{D}}(f^*) & \qquad \text{Bayes risk} \end{aligned}$$

where f^* is the Bayes optimal predictor for $(\mathcal{D}, \ell)$ (see Figure 1).

Note that:

- the Bayes error does not depend on the learning algorithm A,
- the approximation error is large when $\mathcal{H}_m$ does not contain a good approximation of f^*,
- the estimation error is large when S does not contain information sufficient to identify h^*.

We now establish a connection between the bias-variance decomposition and underfitting-overfitting. This connection is best understood when A is the ERM algorithm run over an arbitrary class $\mathcal{H}$ of predictors. Recall that ERM minimizes the training error in $\mathcal{H}$,

$$A(S) = h_S = \operatorname*{argmin}_{h \in \mathcal{H}} \ell_S(h)$$

Similarly to before, the best predictor in the class $\mathcal{H}$ is any predictor h^* satisfying

$$h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h)$$

Fix any such h^* and note that it does not depend on the sample. Suppose that S is at least large enough to ensure that $\ell_S(h^*)$ and $\ell_{\mathcal{D}}(h^*)$ are close to each other with high probability (by the Chernoff-Hoeffding bound). Now, if the variance error is large, then $\ell_{\mathcal{D}}(h_S) \gg \ell_{\mathcal{D}}(h^*)$ and also $\ell_S(h^*) \gg \ell_S(h_S)$ because ERM did not choose h^*. Hence $\ell_{\mathcal{D}}(h_S) \gg \ell_S(h_S)$ and we have overfitting. Vice versa, if the bias error is large and the variance error is small, then $\ell_{\mathcal{D}}(h^*)$ is large and $\ell_{\mathcal{D}}(h^*) \le \ell_{\mathcal{D}}(h_S) \le \ell_{\mathcal{D}}(h^*) + \varepsilon$ for $\varepsilon \ge 0$ small. In this case, as we see next, $\ell_S(h_S)$ is typically close to $\ell_{\mathcal{D}}(h_S)$, and so we have underfitting.

To formalize the above reasoning, we cannot directly apply the Chernoff-Hoeffding bound to h_S because $\ell_S(h_S)$ and $\ell_{\mathcal{D}}(h_S)$ are both random variables, whose expectations do not necessarily coincide. To bound the variance error of ERM we thus proceed as follows. For every given training set S of size m, we have that

$$\begin{aligned}
\ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*) &= \ell_{\mathcal{D}}(h_S) - \ell_S(h_S) + \ell_S(h_S) - \ell_{\mathcal{D}}(h^*) \\
&\le \ell_{\mathcal{D}}(h_S) - \ell_S(h_S) + \ell_S(h^*) - \ell_{\mathcal{D}}(h^*) \\
&\le \big|\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)\big| + \big|\ell_S(h^*) - \ell_{\mathcal{D}}(h^*)\big| \\
&\le 2 \max_{h \in \mathcal{H}} \big|\ell_S(h) - \ell_{\mathcal{D}}(h)\big|
\end{aligned}$$

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

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

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

$$\mathbb{P}\left(\ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*) > \varepsilon\right) \le \mathbb{P}\left(\exists h \in \mathcal{H} : \ \big|\ell_S(h) - \ell_{\mathcal{D}}(h)\big| > \frac{\varepsilon}{2}\right) \ .$$

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

$$\exists h \in \mathcal{H} : \ \big|\ell_S(h) - \ell_{\mathcal{D}}(h)\big| > \frac{\varepsilon}{2}$$

is the union over $h \in \mathcal{H}$ of the (not necessarily disjoint) events $\big|\ell_S(h) - \ell_{\mathcal{D}}(h)\big| > \frac{\varepsilon}{2}$. Using the union bound we get

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

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

In conclusion, we have that

$$\mathbb{P}\left(\ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*) > \varepsilon\right) \le 2|\mathcal{H}| e^{-m\varepsilon^2/2} \; . \qquad (3)$$

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

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

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

For a given size m of the training set, in order to decrease our bound $\sqrt{\frac{2}{m}\ln\frac{2|\mathcal{H}|}{\delta}}$ on the variance error of ERM, 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 conclude that the ERM algorithm generates predictors with high risk (compared to Bayes risk) when there is an unbalance between the variance error and the bias error. In particular, overfitting occurs when the variance error dominates the bias error, and underfitting occurs when the bias error dominates the variance error.

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

$$\forall h \in \mathcal{H} \quad \big|\ell_S(h) - \ell_{\mathcal{D}}(h)\big| \le \sqrt{\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}}$$

holds with probability at least $1-\delta$ with respect to the random draw of the training set. This implies that when the soze of the training set is sufficiently large with respect to $\ln|\mathcal{H}|$, then the training error $\ell_S(h)$ becomes a good estimate for the statistical risk $\ell_{\mathcal{D}}(h)$ *simultaneously* for all predictors $h \in \mathcal{H}$. This is sufficient to prevent overfitting, as it tells us that ranking the predictors in $\mathcal{H}$ according to their training error approximately corresponds to ranking them according to their risk.