Machine Learning — Statistical Methods for Machine Learning

Stability and risk control for SVM

Instructor: *Nicolò Cesa-Bianchi* version of June 1, 2022

ERM addresses underfitting by choosing the predictor minimizing the training error. Overfitting, instead, is controlled by restricting the class of predictors that ERM can choose from. Note that ERM must be sensitive to small changes in the training set, as these may change the predictor minimizing the training error. We now investigate stability, a technique that controls overfitting by preventing the learning algorithm to react to small changes in the training set.

Fix a training set S of examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ and use $\boldsymbol{z}_t$ to denote the t-th example $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. Given a loss function ℓ and a predictor h, let $\ell(h, \boldsymbol{z}_t) = \ell\big(h(\boldsymbol{x}_t), y_t\big)$ and use

$$\ell_S(h) = \frac{1}{m} \sum_{t=1}^{m} \ell(h, \boldsymbol{z}_t)$$

to denote the training error of h. From now on, we assume S is a sample of examples $Z_t = (\boldsymbol{X}_t, Y_t)$ independently drawn from a distribution $\mathcal{D}$. We use $S^{(t)}$ to denote S where the t-th example $(\boldsymbol{X}_t, Y_t)$ is replaced by $\boldsymbol{Z}'_t = (\boldsymbol{X}'_t, Y'_t)$, also drawn from $\mathcal{D}$ independently of S.

Fix a learning algorithm A and let $h_S = A(S)$ and $h_{S^{(t)}} = A\big(S^{(t)}\big)$. We say that A is ε-stable if, for each $t = 1, \ldots, m$

$$\mathbb{E}\big[\ell(h_{S^{(t)}}, \boldsymbol{Z}_t) - \ell(h_S, \boldsymbol{Z}_t)\big] \leq \varepsilon$$

where the expected value is computed with respect to the random draw of S and $\boldsymbol{Z}'_t$. Note that, in general, $\ell(h_{S^{(t)}}, \boldsymbol{Z}_t) > \ell(h_S, \boldsymbol{Z}_t)$ because $\boldsymbol{Z}_t$ is not in $S^{(t)}$. Stability demands that $\ell(h_{S^{(t)}}, \boldsymbol{Z}_t)$ be not much bigger than $\ell(h_S, \boldsymbol{Z}_t)$ in expectation with respect to S and $\boldsymbol{Z}'_t$. In other words, an algorithm is stable if replacing a single example in the training set does not significantly increase the loss on that example. In our analysis, we use the following equivalent definition of stability:

$$\mathbb{E}\big[\ell(h_S, \boldsymbol{Z}'_t) - \ell(h_{S^{(t)}}, \boldsymbol{Z}'_t)\big] \leq \varepsilon$$

The next two results show that a stable algorithm does not overfit.

Theorem 1. *If A is ε-stable, then $\mathbb{E}\big[\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)\big] \leq \varepsilon$.*

Proof. Let $S = \big\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\big\}$ and $S' = \big\{(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_m, y'_m)\big\}$ be drawn i.i.d. from $\mathcal{D}$. Then

$$\mathbb{E}\big[\ell_S(h_S)\big] = \mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m} \ell(h_S, \boldsymbol{z}_t)\right] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_S, \boldsymbol{z}_t)\big] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_{S^{(t)}}, \boldsymbol{z}'_t)\big] \ .$$

Moreover,

$$\ell_{\mathcal{D}}(h_S) = \mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) \mid S\big] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) \mid S\big]$$

averaging with respect to the random draw of S, this implies

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] = \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t)\big] \ .$$

Therefore,

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)\big] = \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) - \ell(h_{S^{(t)}}, \boldsymbol{z}'_t)\big] \leq \varepsilon$$

due to the stability assumption. □

Because stability and minimization of training error work against each other, ERM is typically not stable. On the other hand, a stable learning algorithm that outputs predictors with small empirical risk must have a small variance error.

Theorem 2. *If A is ε-stable and also approximately minimizes the empirical risk in a given class of predictors, that is*

$$\ell_S(h_S) \leq \inf_{h\in\mathcal{H}} \ell_S(h) + \gamma$$

for some $\gamma > 0$, then

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \leq \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) + \varepsilon + \gamma \ .$$

Proof. Let h^* be the predictor with smallest risk in $\mathcal{H}$. Then

$$\begin{aligned}\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] &= \mathbb{E}\Big[\underbrace{\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)}_{\leq\varepsilon \text{ (stability)}}\Big] + \mathbb{E}\Big[\underbrace{\ell_S(h_S) - \ell_S(h^*)}_{\leq\gamma \text{ (ERM approximation)}}\Big] + \mathbb{E}\big[\ell_S(h^*)\big] \\ &\leq \varepsilon + \gamma + \mathbb{E}\big[\ell_S(h^*)\big] \ .\end{aligned}$$

The proof is concluded by observing that $\mathbb{E}\big[\ell_S(h^*)\big] = \ell_{\mathcal{D}}(h^*)$, namely the expected value of the empirical risk is the risk. □

In case of predictors parameterized by a vector $\boldsymbol{w} \in \mathbb{R}^d$ (like linear predictors), ERM can be made stable by adding to the empirical loss a so-called **regularization term**. We also need an additional condition, namely that the loss function ℓ is such that $\ell(\cdot, \boldsymbol{z})$ be convex and Lipschitz, where $\ell(\boldsymbol{w}, \boldsymbol{z})$ is the error of $\boldsymbol{w}$ on the example $\boldsymbol{z}$. Recall that Lipschitz means that there exists a constant L such that $\big|\ell(\boldsymbol{w}, \boldsymbol{z}) - \ell(\boldsymbol{w}', \boldsymbol{z})\big| \leq L\,\|\boldsymbol{w} - \boldsymbol{w}'\|$ for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ and for all $\boldsymbol{z} = (\boldsymbol{x}, y)$. No other assumptions on ℓ are required.

Theorem 3. *Let ℓ be a loss function such that $\ell(\cdot, \boldsymbol{z})$ is convex, differentiable*[1] *and Lipschitz with constant $L > 0$. Then the learning algorithm A such that*

$$A(S) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\,\|\boldsymbol{w}\|^2\right)$$

is $(2L)^2/(\lambda m)$-stable for every $\lambda > 0$.

[1]Some loss functions, notably the hinge loss, are not everywhere differentiable. However, they are everywhere subdifferentiable, which is sufficient for this theorem to hold.

PROOF. Introduce

$$F_S(\boldsymbol{w}) = \ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

and also

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} F_S(\boldsymbol{w}) \qquad \text{and} \qquad \boldsymbol{w}_{S^{(t)}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} F_{S^{(t)}}(\boldsymbol{w}) \ .$$

In order to prove stability, we need to upper bound $\mathbb{E}\big[\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t)\big]$. We actually prove a stronger result by bounding the quantity $\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t)$ for all S and $\boldsymbol{z}'_t$. As a first step, we use the Lipschitz condition to write

$$\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t) \le L\,\|\boldsymbol{w}_S - \boldsymbol{w}_{S^{(t)}}\| \ . \tag{1}$$

Next, we upper bound $\|\boldsymbol{w}_S - \boldsymbol{w}_{S^{(t)}}\|$. Introduce the abbreviations $\boldsymbol{w} = \boldsymbol{w}_S$ e $\boldsymbol{w}' = \boldsymbol{w}_{S^{(t)}}$. Then

$$\begin{aligned}
F_S(\boldsymbol{w}') - F_S(\boldsymbol{w}) &= \ell_S(\boldsymbol{w}') - \ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\big(\|\boldsymbol{w}'\|^2 - \|\boldsymbol{w}\|^2\big) \\
&= \ell_{S^{(t)}}(\boldsymbol{w}') - \ell_{S^{(t)}}(\boldsymbol{w}) + \frac{\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)}{m} - \frac{\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)}{m} + \frac{\lambda}{2}\big(\|\boldsymbol{w}'\|^2 - \|\boldsymbol{w}\|^2\big) \\
&= F_{S^{(t)}}(\boldsymbol{w}') - F_{S^{(t)}}(\boldsymbol{w}) + \frac{\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)}{m} - \frac{\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)}{m} \\
&\le \frac{\big|\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)\big|}{m} + \frac{\big|\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)\big|}{m} \\
&\le \frac{2L}{m}\,\|\boldsymbol{w} - \boldsymbol{w}'\|
\end{aligned}$$

where the first inequality holds because $\boldsymbol{w}' = \boldsymbol{w}_{S^{(t)}}$ minimizes $F_{S^{(t)}}$ and the second inequality holds because $\ell(\cdot, \boldsymbol{z})$ is Lipschitz.

We proceed by noting that the function F_S is λ-strongly convex: indeed, $\ell(\cdot, \boldsymbol{z})$ is convex, $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ is λ-strongly convex, which implies that their sum is also λ-strongly convex. Therefore, by definition of strongly convex function,

$$F_S(\boldsymbol{w}') \ge F_S(\boldsymbol{w}) + \nabla F_S(\boldsymbol{w})^\top(\boldsymbol{w}' - \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 \ .$$

Because $\boldsymbol{w}$ is the minimizer of F_S, $\nabla F_S(\boldsymbol{w}) = \boldsymbol{0}$ and so

$$F_S(\boldsymbol{w}') - F_S(\boldsymbol{w}) = \Big(\ell_S(\boldsymbol{w}') + \frac{\lambda}{2}\|\boldsymbol{w}'\|^2\Big) - \Big(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\Big) \ge \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2$$

Combining the two inequalities we get

$$\frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 \le \frac{2L}{m}\|\boldsymbol{w} - \boldsymbol{w}'\| \qquad \text{ovvero} \qquad \|\boldsymbol{w} - \boldsymbol{w}'\| \le \frac{4L}{\lambda m}$$

which, together with (1) shows the stability of $\boldsymbol{w} = \boldsymbol{w}_S$. □

We now show how the notion of stability can be used to control the risk of the SVM predictor. First, recall that the hinge loss $\ell\big(\boldsymbol{w}, (\boldsymbol{x}, y)\big) = \big[1 - y\,\boldsymbol{w}^\top\boldsymbol{x}\big]_+$ is convex in $\boldsymbol{w}$ for $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$. In

order to compute the Lipschitz constant L, observe that, using Taylor's theorem and the Cauchy-Schwartz inequality $\boldsymbol{u}^\top \boldsymbol{v} \leq \|\boldsymbol{u}\| \, \|\boldsymbol{v}\|$, we can write

$$\ell(\boldsymbol{w}', \boldsymbol{z}) - \ell(\boldsymbol{w}, \boldsymbol{z}) \leq \nabla \ell(\boldsymbol{w}, \boldsymbol{z})^\top \big(\boldsymbol{w}' - \boldsymbol{w}\big) \leq \|\nabla \ell(\boldsymbol{w}, \boldsymbol{z})\| \, \big\|\boldsymbol{w}' - \boldsymbol{w}\big\|$$

Now note that $\nabla \ell(\boldsymbol{w}, \boldsymbol{z}) = y\boldsymbol{x} \, \mathbb{I}\{y \, \boldsymbol{w}^\top \boldsymbol{x} \leq 1\}$, and thus $\|\nabla \ell(\boldsymbol{w}, \boldsymbol{z})\| \leq \|\boldsymbol{x}\|$. Assuming $\|\boldsymbol{x}_t\| \leq X$ for $t = 1, \ldots, m$ we can then set $L = X$ and show the following result.

Theorem 4. *Given a training set S, the SVM solution*

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \right) \tag{2}$$

satisfies

$$\mathbb{E}\big[\ell_{\mathcal{D}}^{0\text{-}1}(\boldsymbol{w}_S)\big] \leq \min_{\boldsymbol{u} \in \mathbb{R}^d} \Big(\ell_{\mathcal{D}}(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 \Big) + \frac{4X^2}{\lambda m} \ .$$

where $\ell_{\mathcal{D}}^{0\text{-}1}(\boldsymbol{w}_S)$ is the risk of $\boldsymbol{w}_S$ with respect to the zero-one loss.

Proof. Clearly, for each $\boldsymbol{u} \in \mathbb{R}^d$ we have

$$\ell_S(\boldsymbol{w}_S) \leq \ell_S(\boldsymbol{w}_S) + \frac{\lambda}{2} \|\boldsymbol{w}_S\|^2 \leq \ell_S(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 \ . \tag{3}$$

Therefore, because Theorem 3 implies that $\boldsymbol{w}_S$ is $(4L^2)/(\lambda m)$-stable for $L = X$,

$$\begin{aligned}
\mathbb{E}\big[\ell_{\mathcal{D}}(\boldsymbol{w}_S)\big] &\leq \mathbb{E}\big[\ell_S(\boldsymbol{w}_S)\big] + \frac{4X^2}{\lambda m} \qquad \text{by Theorem 1} \\
&\leq \mathbb{E}\left[\ell_S(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2\right] + \frac{4X^2}{\lambda m} \qquad \text{using (3)} \\
&= \ell_{\mathcal{D}}(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 + \frac{4X^2}{\lambda m}
\end{aligned}$$

The proof is concluded by noting that, for any linear predictor $\boldsymbol{w}$, $\ell_{\mathcal{D}}^{0\text{-}1}(\boldsymbol{w}) = \mathbb{P}\big(Y \, \boldsymbol{w}^\top \boldsymbol{X} \leq 0\big) \leq \ell_{\mathcal{D}}(\boldsymbol{w})$. This is an immediate consequence of the fact that the hinge loss is a convex upper bound on the zero-one loss. □

Consistency of SVM. In a kernel space $\mathcal{H}_K$ the SVM objective function (2) becomes

$$g_S = \operatorname*{argmin}_{g \in \mathcal{H}_K} \left(\ell_S(g) + \frac{\lambda}{2} \|g\|_K^2 \right)$$

If the kernel is Gaussian, then one can prove that SVM becomes a consistent learning algorithm (with respect to the zero-one loss) when the regularization parameter λ is chosen as a function λ_m of the training set size m. In particular, for $m \to \infty$, λ_m must satisfy the two following conditions: $\lambda_m = o(1)$ and $\lambda_m = \omega(m^{-1/2})$.