Machine Learning 10: Regularization and Stability

Master 2 Computer Science

Aurélien Garivier

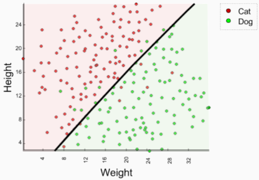
2019-2020

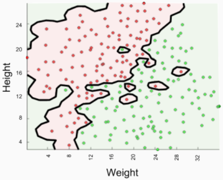
Table of contents

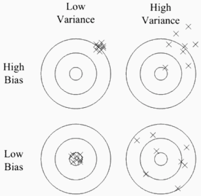
Regularization and Structural Risk Minimization

Overfitting

Example: linear classification with polynomial features







Src: `http://mlwiki.org`

$\rightarrow$ how to get the best from several hypothesis classes?

Nonuniform Learnability

Definition

A hypothesis class $\mathcal{H}$ is *nonuniformy learnable* if there exists a learning algorithm A and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then with probability at least $1 - \delta$ over the sample $S \sim D^{\otimes m}$,

$$L_D\big(A(S)\big) \leq L_D(h) + \epsilon \ .$$

Theorem

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.

Structural Risk Minimization 1/2

Proof of sufficiency: Let $\mathcal{H} = \cup_{d \in \mathbb{N}} \mathcal{H}_d$, where each hypothesis class $\mathcal{H}_d$ is PAC learnable with uniform convergence rate $m^{UC}_{\mathcal{H}_d}$, and let $\epsilon_d : \mathbb{N} \times (0,1) \to (0,1)$ be defined as

$$\epsilon_d(m,\delta) = \min\left\{\epsilon \in (0,1) : m^{UC}_{\mathcal{H}_d}(\epsilon,\delta) \leq m\right\} \ .$$

For every $h \in \mathcal{H}$ let $d(h) = \min\left\{d : h \in \mathcal{H}_d\right\}$. Let also $w : \mathbb{N} \to [0,1]$ be such that $\sum_{d=0}^{\infty} w(d) \leq 1$.

Lemma

For every $\delta \in (0,1)$ and for every distribution D, with probability at least $1-\delta$ over the sample $S \sim D^{\otimes m}$,

$$\forall h \in \mathcal{H}, \ \ L_D(h) \leq L_S(h) + \epsilon_{d(h)}\Big(m, w\big(d(h)\big)\delta\Big) \ .$$

Structural Risk Minimization 2/2

Structural Risk Minimization (SRM)

$$A(S) \in \underset{h \in \mathcal{H}}{\arg\min}\, L_S(h) + \epsilon_{d(h)}\Big(m, w\big(d(h)\big)\delta\Big) .$$

Typical choice: $w(d) = \frac{6}{\pi^2(d+1)^2}$ gives for SRM the nonuniform learning rate

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq m_{\mathcal{H}_{d(h)}}^{UC}\left(\frac{\epsilon}{2}, \frac{6\delta}{\pi^2 d(h)^2}\right) .$$

If $\mathrm{VCdim}(\mathcal{H}_d) = d$, $m_{\mathcal{H}_d}^{UC}(\epsilon/2, \delta) = C\,\frac{d+\log(1/\delta)}{\epsilon^2}$ and hence

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) - m_{\mathcal{H}_d}^{UC}(\epsilon/2, \delta) \leq \frac{8C\log(2d)}{\epsilon^2} .$$

Remark: other strategy = *aggregation*, cf PAC-Bayes learning.

Minimum Description Length and Occam's razor

Entiae non sunt multiplicanda praeter necessitatem
(Entities are not to be multiplied without necessity)
Here: A short explanation tends to be more valid
(generalize better) than a long explanation

Suggests a choice for $w(d)$: should penalize complexity.

More precisely: if $|h|$ is the length of a prefix-free binary code for the hypothesis h, set

$$w(h) = 2^{-|h|} .$$

By Hoeffding's inequality, this typically yields the

Minimum Description Length (MDL) estimator:

$$A(S) \in \operatorname*{arg\,min}_{h \in \mathcal{H}} L_S(h) + \sqrt{\frac{|h| + \log \frac{2}{\delta}}{2m}} .$$

This heuristic needs to be justified statistically (often possible).

Regularization and Stability

Stable Rules do not overfit

Theorem

Let D be a distribution on $\mathcal{X} \times \{\pm 1\}$, $S = (z_1, \ldots, z_m)$ be an iid sequence of examples, z' be another independent sample of D, and let I be an independent sample of the uniform distribution on $\{1, \ldots, m\}$. For all $1 \leq i \leq m$, let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. Then, for any learning alogrithm A,

$$\mathbb{E}_S\left[L_D\big(A(S)\big) - L_S\big(A(S)\big)\right] = \mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big) - \ell\big(A(S), z_I\big)\right] .$$

Indeed, $\mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big)\right] = \mathbb{E}_S\left[L_D\big(A(S)\big)\right]$, and $\mathbb{E}_{S,I}\left[\ell\big(A(S), z_I\big)\right] = \mathbb{E}_S\left[L_S\big(A(S)\big)\right]$.

Definition

Algorithm A is said to be *on-average-replace-one*-stable with rate $\epsilon : \mathbb{N} \to \mathbb{R}$ if for every distribution D and every sample size $m \in \mathbb{N}$,

$$\mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big) - \ell\big(A(S), z_I\big)\right] \leq \epsilon_m .$$

Tikhonov Regularization as a Stabilizer

We consider a class $\mathcal{H} = \left\{ h_w : w \in \bigcup_{d \geq 0} \mathbb{R}^d \right\}$.

Definition

Tikhonov's Regularized Loss Minimizer is defined as

$$A(S) \in \underset{h_w \in \mathcal{H}}{\arg\min}\, L_S(h) + \lambda \|w\|^2 \ ,$$

where $\lambda > 0$ is a parameter.

With square loss on $\mathbb{R}^d$, the resulting estimator is called *ridge regression*:

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\arg\min}\, \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \big(\langle w, x_i \rangle - y_i \big)^2 + \lambda \|w\|^2 = \big(2\lambda m I_d + X^T X \big)^{-1} X^T y \ ,$$

where $X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}$.

Tikhonov's RLM for convex loss is stable

Denote $f_S(w) = L_S(w) + \lambda\|w\|^2$. If ℓ is convex, then f is 2λ-strongly convex, and thus

$$f_S\big(A(S^{(i)}) - f_S\big(A(S)\big) \geq \lambda\big\|A(S^{(i)}) - A(S)\big\|^2 ,$$

and

$$f_S\big(A(S^{(i)})\big) - f_S\big(A(S)\big) = \underbrace{L_{S^{(i)}}\big(A(S^{(i)})\big) + \lambda\big|A(S^{(i)})\big|^2 - L_{S^{(i)}}\big(A(S)\big) - \lambda\big|A(S)\big|^2}_{\leq 0}$$

$$+ \frac{\ell\big(A(S^{(i)}), z_i\big) - \ell\big(A(S), z_i\big)}{m} + \frac{\ell\big(A(S), z'\big) - \ell\big(A(S^{(i)}), z'\big)}{m} ,$$

and hence

$$\lambda\big\|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell\big(A(S^{(i)}), z_i\big) - \ell\big(A(S), z_i\big)}{m} + \frac{\ell\big(A(S), z'\big) - \ell\big(A(S^{(i)}), z'\big)}{m}$$

Lipschitz loss

When the loss $\ell(\cdot, z)$ is ρ-Lipschitz for every z, we obtain that

$$\lambda \left\| A(S^{(i)}) - A(S) \right\|^2 \leq \frac{2\rho \left\| A(S^{(i)}) - A(S) \right\|}{m} ,$$

when entails $\left\| A(S^{(i)}) - A(S) \right\| \leq \dfrac{2\rho}{\lambda m}$.

RLM generalizes well Lispchitz Losses

When the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz for all z, Tikhonov's RLM is on-average-one-stable with rate $\frac{2\rho^2}{\lambda m}$, and hence

$$\mathbb{E}_S \left[L_D\big(A(S)\big) - L_S\big(A(S)\big) \right] \leq \frac{2\rho^2}{\lambda m} .$$

Remark: when ℓ is β-smooth and non-negative, and when $\ell(0, z) \leq C$ for all z, one can prove that for $\lambda \geq \frac{2\beta}{m}$ Tikhonov's RLM satisfies

$$\mathbb{E}_S \left[L_D\big(A(S)\big) - L_S\big(A(S)\big) \right] \leq \frac{48\beta}{\lambda m} \mathbb{E}\left[L_S\big(A(S)\big) \right] \leq \frac{48\beta C}{\lambda m} .$$

Controlling Fitting-Stability Tradeoff

Fitting-stability tradeoff:

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] = \underbrace{\mathbb{E}_S\Big[L_S\big(A(S)\big)\Big]}_{\text{fitting error}} + \underbrace{\mathbb{E}_S\Big[L_D\big(A(S)\big) - L_S\big(A(S)\big)\Big]}_{\text{generalization error} = \text{stability}} .$$

The stronger the regularization (the larger λ), the better the stability BUT the higher the bias.

But for every $h_w \in \mathcal{H}$,

$$\mathbb{E}_S\Big[L_S\big(A(S)\big)\Big] \leq \mathbb{E}_S\Big[L_S(h_w) + \lambda\|w\|^2\Big] = L_D(h_w) + \lambda\|w\|^2 .$$

Oracle inequality

If the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz for all z, Tikhonov's RLM satisfies

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \lambda\|w\|^2 + \frac{2\rho^2}{\lambda m}$$

Corollary

If $\forall h_w \in \mathcal{H}, \|w\| \leq B$ and if the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz for all z, Tikhonov's RLM with $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ satisfies:

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \rho B \sqrt{\frac{8}{m}}\,.$$

Hence, for every $\epsilon > 0$, if $m \geq \frac{8\rho^2 B^2}{\epsilon^2}$ then for every distribution D $\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \epsilon$.

The same kind of result can be obtained for β-smooth, non-negative losses: with $\lambda = \epsilon/(3B^2)$, for every $m \geq \frac{150\beta B^2}{\epsilon^2}$, whatever the distribution D, $\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \epsilon$.

In practice, λ is most often chosen by cross-validation.

Example: Ridge regression generalizes well

Theorem

Let D be a distribution over $\mathcal{X} \times [-1, 1]$, where $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \leq 1\right\}$. Let $\mathcal{H} = \left\{w \in \mathbb{R}^d : \|w\| \leq B\right\}$. For any $\epsilon \in (0, 1)$, let $m \geq m_{\mathcal{H}}(\epsilon) = 150B^2/\epsilon^2$. Then ridge regression with parameter $\lambda = \epsilon/(3B^2)$ satisfies:

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon \; .$$

Furthermore, for every $\delta \in (0, 1)$ and every $m \geq m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}(\epsilon\delta)$,

$$\mathbb{P}_S\Big(L_D\big(A(S)\big) \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon\Big) \geq 1 - \delta \; .$$

Expectation to high-probability PAC learning: the sample complexity can be reduced to $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}(\epsilon/2)\big\lceil \log_2(1/\delta)\big\rceil + \left\lceil \frac{\log(4/\delta) + \log(\lceil \log_2(1/\delta)\rceil)}{\epsilon^2}\right\rceil$ when the loss function is bounded by 1.