Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 5: statistical learning theory II: Vapnik-Chervonenkis theory.

The story so far

- To bound the generalization error of a single function (e.g. based on a test set error), a variety of methods are available (Binomial inversion for the 0-1 loss; Chernoff's method and Chernoff's, Bernstein's, Hoeffding's inequalities for bounded losses).
- To bound the generalization error of a function $\widehat{f}$ chosen from a countable pool $\mathcal{F}$, based on the training sample S, we can bound the generalization error of $\widehat{f}$ provided we have a uniform control over $\mathcal{F}$ of the form: with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \qquad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f, S) + \varepsilon(\delta, f, n).$$

  (We need this because both $\widehat{f}$ and $\widehat{\mathcal{E}}(., S)$ are random quantities involving the training set S, hence they are dependent).
- To do this, we proposed to use the union bound (Bonferroni's correction) if $\mathcal{F}$ is finite, or the union-bound-with-a-prior ("Occam's razor"), where the prior can be seen as a repartition of confidence, or a prior belief about "complexity".

There are at least two reasons why the union bound will not be satisfactory in many cases:

- A lot of interesting function classes are uncountable!
- Even for a countable class, if two (classification) functions f, f' are very "close" to each other, we expect that they will tend to have a similar behavior on the same training set.
- Hence, if the confidence interval for the first function f is valid for a sample S, then it is "likely" that it is also the case for the second function f'.
- In the union bound, we always consider the worst case where the CIs for two distinct functions will fail on different set of samples. This is probably very overpessimistic.

A detour via bounded regression

- Assume we want to estimate $\eta(Y|X)$, consider the squared error function $\ell(f, X, Y) = (f(X) - Y)^2$, and want to pick an estimator in some fixed class $\mathcal{F}$ of bounded by 1 and continuous functions.
- We want to control uniformly the deviation between true and empirical error

$$\mathcal{E}(\ell, f) - \widehat{\mathcal{E}}(\ell, f, S) = (P - P_n)(\ell(f, X, Y)),$$

where we introduce the notation $P(.)$ for expectation wrt. the drawing probability P, and $P_n(.)$, the empirical expectation.

- Denote $\mathcal{G}$ the loss class based on $\mathcal{F}$

$$\mathcal{G} = \{(X, Y) \mapsto \ell(f, X, Y) | f \in \mathcal{F}\} .$$

- If d is a distance, introduce the notion of covering number $\mathcal{N}(\mathcal{G}, d, \epsilon)$ the smallest cardinality M of a set $\mathcal{M} = \{g_1, \ldots, g_M\}$ such that the ε-balls $\mathcal{B}_d(g_i, \epsilon)$ centered on elements of $\mathcal{M}$ "cover" $\mathcal{G}$.
- Let's apply this with the supremum norm distance, and the union bound. It comes: with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}$$

$$\mathcal{E}(\ell, f) - \widehat{\mathcal{E}}(\ell, f, \mathcal{S}) \leq 2\varepsilon + \sqrt{\frac{2(\log \mathcal{N}(\mathcal{G}, \|\|_\infty, \varepsilon) + \log \delta^{-1})}{n}}.$$

- (This can be optimized in ε).
- Unfortunately, this approach cannot be directly applied to classification functions: why?

The plan of what is to come

- The goal: for a set of classification functions $\mathcal{F}$, obtain a uniform CI of the form: with probability $1-\delta$, we have

$$\forall f \in \mathcal{F}, \qquad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f, S) + \varepsilon(\delta, n)\,.$$

  (here we consider the case where ε does not depend on f: comparable to the “Bonferroni” union bound in the finite case).
- This is equivalent to showing that, with probability $1-\delta$,

$$\sup_{f \in \mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f, S) \leq \varepsilon(\delta, n)\,.$$

- Step 1: show that the random variable

$$\sup_{f \in \mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f, S)$$

  “concentrates” around its expectation (i.e. is close to it with high probability)
- Step 2: upper bound this expectation in some way.

Concentration and stability

The following theorem is very important:

Theorem (Azuma, McDiarmid)

Let $(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$ *be a measurable function such that*

$$\forall 1 \leq i \leq n \, , \forall (x_1, \ldots, x_n) \text{ and } x_i' , \\ \left| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n) \right| \leq c_i \, . \quad (1)$$

Then, if $(X_1, \ldots, X_n)$ *are independent (not necessary i.d.), it holds that*

$$\mathbb{P}\left[f(X_1, \ldots, X_n) - \mathbb{E}\left[f \right] \geq \varepsilon \right] \leq \exp - \frac{2\varepsilon^2}{\sum_{1 \leq i \leq n} c_i^2} \, .$$

Proof: apply Hoeffding's inequality conditionally and repeatedly.

Application to bounded losses

- Consider the functional:

$$S = ((X_i, Y_i)_{1\leq i\leq n}) \mapsto f(S) = \sup_{f\in\mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f, S)\,.$$

- It is "$\frac{B}{n}$-stable" in the sense of the previous theorem.
- Hence, with probability $1-\delta$ over the draw of S, we have

$$\begin{aligned}\mathcal{E}(\widehat{f}) - \widehat{\mathcal{E}}(\widehat{f}, S) &\leq \sup_{f\in\mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f, S)\\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f, S)\right] + B\sqrt{2\frac{\log \delta^{-1}}{n}}\,.\end{aligned}$$

Dealing with the expectation 1: symmetrization

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathcal{E}(f) - \widehat{\mathcal{E}}(f,S)\right] \leq \mathbb{E}_S \mathbb{E}_{S'} \mathbb{E}_{(\sigma_i)_{1\leq i\leq n}} \left[\sup_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\ell(f,X_i,Y_i) - \ell(f,X_i',Y_i')\right)\right] ,$$

where:

- S' is a "phantom" sample, draw exactly like S but independently;
- $(\sigma_i)_{1\leq i\leq n}$ is a family of "random signs" (a.k.a. Rademacher variables), that is, $\sigma_i = 2B_i - 1$ where B_i is Bernoulli$\left(\frac{1}{2}\right)$.

Dealing with the expectation 2: “Shattering” coefficients

- We restrict our attention now to the case of $0-1$ loss for classification.
- Let us look at the expectation over the Rademacher signs only, everything else being fixed.
- Consider the application

$$f \in \mathcal{F} \mapsto G_{S,S'}(f) = \left(\mathbb{1}\{f(X_1) \neq Y_1\}, \ldots, \mathbb{1}\{f(X_1') \neq Y_1'\}, \ldots\right) .$$

- When (S, S') is fixed, the supremum operation is actually only over $H_{\mathcal{F}}(S, S') = \text{card}\left(G_{S,S'}(\mathcal{F})\right)$ distinct elements!

Expectation of the supremum of sub-Gaussian variables

Lemma

Let $Z_1, \ldots, Z_M$ a family of random variables satisfying for a certain constant $\sigma > 0$:

$$\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] \leq \exp\left(\frac{\sigma^2 \lambda^2}{2}\right) \text{ for all } 1 \leq i \leq M\,.$$

(the family does not have to be independent not i.d.).
Then

$$\mathbb{E}\left[\max_{1 \leq i \leq M} Z_i\right] \leq \sigma \sqrt{2 \log M}\,.$$

Putting everything together, we have proved:

Theorem

Consider the 0-1 classification loss $\ell(f, X, Y) = \mathbb{1}\{f(X) \neq Y\}$. *Consider* *any* *learning algorithm returning* $\widehat{f} \in \mathcal{F}$ *depending on* S. *With probability* $1 - \delta$ *over the draw of* S, *the following holds:*

$$\mathcal{E}(\widehat{f}) - \widehat{\mathcal{E}}(\widehat{f}) \leq \frac{\mathbb{E}_{S,S'}\left[\sqrt{2 \log H_{\mathcal{F}}(S, S')}\right]}{\sqrt{n}} + \sqrt{\frac{\log \delta^{-1}}{2n}}.$$

Now, what about this $\log H_{\mathcal{F}}(S, S')$ quantity?

The Sauer-Vapnik lemma

Lemma

Assume for the family of classifiers $\mathcal{F}$, there exists d such that, for any sample S of size d,

$$H(S_d) < 2^d .$$

Then for all $n > d$ for all sample S of size n,

$$H_{\mathcal{F}}(S) \leq \sum_{i=0}^{d-1} \binom{n}{i}$$

The quantity $d - 1$ is then called the Vapnik-Chervonenkis dimension of class $\mathcal{F}$.

One last effort

We can then upper bound the shattering coefficient by a more tractable quantitity: if $\mathcal{F}$ has VC dimension d , then

$$(i)\forall S\,,|S| = n\,, H_{\mathcal{F}}(S) \leq (n+1)^d\;;$$

$$(ii)\forall S\,,|S| = n \geq d\,, H_{\mathcal{F}}(S) \leq \left(\frac{ne}{d}\right)^d\;;$$

The final result

Theorem

Consider the 0-1 classification loss $\ell(f, X, Y) = \mathbb{1}\{f(X) \neq Y\}$*. Let* $\mathcal{F}$ *be a set of classifiers of VC dimension* d*. Consider any learning algorithm returning* $\widehat{f} \in \mathcal{F}$ *depending on* S*. With probability* $1-\delta$ *over the draw of* S*, the following holds:*

$$\mathcal{E}(\widehat{f}) - \widehat{\mathcal{E}}(\widehat{f}) \leq \sqrt{\frac{2(d+1)\log(2n)}{n}} + \sqrt{\frac{\log \delta^{-1}}{2n}}.$$

Examples of VC dimensions

Theorem

The class of indicators of parallelepipeds in $\mathbb{R}^k$ has VC dimension $2k$.

Theorem

The class of linear separators in $\mathbb{R}^k$ has VC dimension $k + 1$.

This last theorem allows to upper bounds the VC dimension of "generalized linear separators" including indicators or spheres, ellipsoids…

Combining VC theory + Occam's Razor

- We can consider different algorithms $\widehat{f_1}, \ldots, \widehat{f_k}$ picking their classifiers in classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of increasing VC-dimensions $d_1 < \ldots < d_k$.
- We can apply a principle similar to Occam's Razor, getting a uniform bound over these different algorithms via a prior π on $\{1, \ldots, k\}$ (uniform for example) .
- In this case, with probability $1 - \delta$ it holds:

$$\forall 1 \leq i \leq k \,, \forall f \in \mathcal{F}_i,$$

$$\mathcal{E}(\widehat{f}) \leq \widehat{\mathcal{E}}(\widehat{f}) + \sqrt{\frac{2(d+1)\log(2n)}{n}} + \sqrt{\frac{\log \pi(i)^{-1}}{2n}} + \sqrt{\frac{\log \delta^{-1}}{2n}} \,.$$

- If we pick the model minimizing this bound, this leads to Vapnik's "structural risk minimization" (SRM) principle.