Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 9: Learning theory III: Rademacher complexities.

Rewind to the analysis of VC bounds

- We have seen previously that if ℓ is a loss function such that $|\ell(.)| \leq B$, and $\mathcal{F}$ a class of functions of interest (classifiers, regression functions. . .) the following holds with probability at least $1 - \delta$ over the draw of the training sample S:

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

$$\leq \frac{1}{n}\mathbb{E}_{S,S'}\mathbb{E}_{\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i(\ell(f, Z_i) - \ell(f, Z_i'))\right] + B\sqrt{\frac{2\log \delta^{-1}}{n}}$$

$$\leq \frac{2}{n}\mathbb{E}_{S}\mathbb{E}_{\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \ell(f, Z_i)\right] + B\sqrt{2\frac{\log \delta^{-1}}{n}}$$

- Similarly, with probability at least $1 - \delta$ over the draw of the training sample S:

$$\sup_{f \in \mathcal{F}} \left| \mathcal{E}(\ell, f) - \widehat{\mathcal{E}}(\ell, f, S) \right| \leq \frac{2}{n} \mathbb{E}_S \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i \ell(f, Z_i) \right| \right] + B \sqrt{2 \frac{\log \delta^{-1}}{n}} .$$

- In VC theory we upper bounded the above quantity in the case of binary classifiers, but this symmetrized quantity has interesting properties of its own and can be used in much more general cases.

Definition

- We denote

$$\mathcal{R}(\mathcal{F}) = \frac{1}{n}\mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i g(Z_i)\right]$$

  the Rademacher complexity of class $\mathcal{G}$. (Note that above we applied it to $\mathcal{G} = \{\ell(f,.); f \in \mathcal{F}\}$.)

- We also denote

$$\mathcal{R}_{||}(\mathcal{F}) = \frac{1}{n}\mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i g(Z_i)\right|\right]$$

- Note: in general the supremum of an arbitrary family of measurable functions might not be measurable. Here we assume that the function spaces we consider are such that the above suprema can always be restricted to a countable subfamily.

Elementary properties of Rademacher complexities

Rademacher complexities satisfy the following composition inequalities:

- For any $a \in \mathbb{R}$, $\mathcal{R}(a\mathcal{F}) = |a|\mathcal{R}(\mathcal{F})$;
- $\mathcal{R}(\mathcal{F}+\mathcal{G}) \leq \mathcal{R}(\mathcal{F}) + \mathcal{R}(\mathcal{G})$;
- $\mathcal{R}(\{\sup(f,g); f \in \mathcal{F}, g \in \mathcal{G}\}) \leq (\mathcal{R}(\mathcal{F}) + \mathcal{R}(\mathcal{G}))$
- For any function h, $\mathcal{R}(\{h\}) = 0$; $\mathcal{R}_{||}(\{h\}) \leq \sqrt{\frac{\mathbb{E}[h^2]}{n}}$.
- Similar inequalities hold for $R_{||}$.

A cornerstone of Rademacher analysis

Theorem (Comparison principle)

Let γ be a Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz constant L. Then

$$\mathcal{R}\left(\{\gamma(f(.)), f \in \mathcal{F}\}\right) \leq L\mathcal{R}(\mathcal{F}).$$

If furthermore $\gamma(0) = 0$, then

$$\mathcal{R}_{||}\left(\{\gamma(f(.)), f \in \mathcal{F}\}\right) \leq 2L\mathcal{R}_{||}(\mathcal{F}).$$

Main lemma for the comparison principle

Lemma

Let $(A_i, B_i)_{i \in I}$ be a countable family of constants, and let X be a symmetric random variable. Then if γ is a L-Lipschitz function,

$$\mathbb{E}\left[\sup_i \left(A_i + X\gamma(B_i)\right)\right] \leq \mathbb{E}\left[\sup_i \left(A_i + LXB_i\right)\right] .$$

Consequence: bounded LS regression

- Consider the squared error loss function for regression $\ell(f,x,y) = (f(x)-y)^2$, and assume that we know that both the random variable Y and the functions $f \in \mathcal{F}$ are bounded by B. Then

$$\mathcal{R}\left(\left\{(x,y) \mapsto (f(x)-y)^2; f \in \mathcal{F}\right\}\right) \leq 4B\mathcal{R}(\mathcal{F}).$$

- Hence in this situation, with probability $1-\delta$, for all $f \in \mathcal{F}$,

$$\mathcal{E}(f(X)-Y)^2 \leq \frac{1}{n}\sum_i (f(X_i)-Y_i)^2 + 4B\mathcal{R}(\mathcal{F}) + 2B^2\sqrt{\frac{2\log\delta^{-1}}{n}}.$$

Consequence: margin-based loss functions

- Consider a loss function ℓ which is a function of the edge/margin: $\ell(f,x,y) = \ell_0(f(x)y)$, where ℓ_0 is L-Lipschitz. Then

$$\mathcal{R}(\{(x,y) \mapsto \ell(f,x,y); f \in \mathcal{F}\}) \leq L\mathcal{R}(\mathcal{F}).$$

- This applies for example to the hinge loss function $h(u) = (1-u)_+$, or the "deviance" $d(u) = -\log(1+\exp(-2u))$, if we assume the function class to be bounded (for Mcdiarmid).
- By considering the "thresholded + rescaled" hinge loss $\ell_\theta(u) = \min(h(\theta^{-1}u), 1)$, we obtain that with probability $1-\delta$, for all $f \in \mathcal{F}$,

$$\mathcal{E}(f) = \mathbb{P}[f(X)Y \leq 0] \leq \frac{1}{n}\sum_i \mathbb{1}\{f(X_i)Y_i \leq \theta\} + \theta^{-1}\mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log \delta^{-1}}{2n}}.$$

Important: for this, we don't need to assume boundedness of functions in $\mathcal{F}$ since the loss function itself is bounded by construction!

Rademacher complexity of kernel function classes

- We can apply the previous analysis for kernel function classes: consider the Hilbert ball B_R of radius R in the RKHS, then we have

$$\mathcal{R}(B_R) \leq \frac{1}{\sqrt{n}} R \sqrt{\mathbb{E}\left[k(X,X)\right]}\,.$$

- Thus, we obtain that with probability at least $1-\delta$, $\forall f \in B_R$:

$$\begin{aligned}
&\mathbb{P}\left[f(X)Y \leq 0\right] \\
&\leq \frac{1}{n} \sum_{1 \leq i \leq n} \mathbb{1}\{f(X_i)Y_i \leq \theta\} + \frac{1}{\sqrt{n}} \frac{R}{\theta} \sqrt{\mathbb{E}\left[k(X,X)\right]} + \sqrt{\frac{2\log \delta^{-1}}{n}}\,.
\end{aligned}$$

- Above we assumed that R, θ are fixed but by homogeneity we can take $R = 1$ w.l.o.g. and θ then plays the role of the "margin" parameter.

Rademacher complexity of ensemble methods

- The Rademacher complexity satisfies the following property:

$$\mathcal{R}(\mathrm{conv}(\mathcal{F})) = \mathcal{R}(\mathcal{F})\,.$$

- Hence, if the set of base classifiers has finite Rademacher complexity, say has VC dimension less that d , then with probability at least $1-\delta$, $\forall\alpha$ probability distribution on the base classifier set $\mathcal{F}$:

$$\mathbb{P}\left[F_\alpha(X)Y \le 0\right]$$
$$\le \frac{1}{n}\sum_{1\le i\le n} \mathbb{1}\{F_\alpha(X_i)Y_i \le \theta\} + \frac{1}{\theta}\sqrt{\frac{2(d+1)\log 2n}{n}} + \sqrt{\frac{2\log\delta^{-1}}{n}}\,,$$

  where we recall that $F_\alpha(x) = \mathbb{E}_{t\sim\alpha}\left[f_t(x)\right]$.
- It looks as if we did not increase the complexity by going from single classifiers in $\mathcal{F}$ to ensembles ?? How is that possible?

Uniform bound wrt. regularization term

- In the two previous bounds we fixed the value of the margin (or equivalently by homogeneity, the value of the "regularization parameter": Hilbert norm for the kernel case; sum of the coefficients in the ensemble case).
- In practice this parameter gets picked from the data by the algorithm, so how can we make the bound valid in this case?
- Apply once again Occam's razor over some discretized sequence of the regularization parameters, e.g. radii $R = 1, 2, \cdot$, with a "flat prior", for example $\pi(i) = \frac{6}{\pi^2} i^{-2}$.

Important difference between 1-norm and 2-norm regularization

- Consider a simplified case where we consider an ensemble methods taking classifiers in a finite base classifer space $\mathcal{F}$ (taking values in $[-1, 1]$) of cardinality D .
- If we consider this as a feature mapping and apply the "geometric large margin" approach, we obtain a bound with a factor $R\sqrt{\mathbb{E}\left[k(X, X)\right]} = R\sqrt{D}$ for the complexity term; here R is the ℓ_2 norm of the coefficient vector w .
- If we apply the "ensemble large margin" approach, we obtain a bound with a factor $R\sqrt{\log D}$, where R is the ℓ_1 norm of the coefficient vector w .
- Note that this does not mean that the ensemble approach is intrisically "less complex" since in practice one will look for ℓ_1 balls of much larger radius to approximate the data. But it illustrates that in high dimension the two measures of complexity become of an increasingly different nature.

Empirical Rademacher

- Note that we can apply MacDiarmid's inequality to the Rademacher complexity itself.
- Hence with probability at least $1 - \delta$,

$$\mathcal{R}(\mathcal{F}) \leq \widehat{\mathcal{R}}(\mathcal{F}) + B\sqrt{\frac{2\log \delta^{-1}}{n}}\,.$$

where

$$\widehat{\mathcal{R}}(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i g(Z_i)\right]$$

- Thus, at the price of tripling the trailing term in the bounds, we can replace Rademacher complexities by their empirical counterparts. (Note that by applying directly McDiamid's inequality to the sum of all terms, we avoid dividing δ by 2. . .)

Empirical Rademacher for certain classes

- In the case of kernel classes, we obtain

$$\widehat{\mathcal{R}}(B_R) \leq \frac{R}{\sqrt{n}} \left(\frac{1}{n} \sum_{1 \leq i \leq n} k(X_i, X_i) \right) ;$$

  note how the obtained bound now resembles even more closely the semi-heuristic “margin” bound that served to motivate the SVMs.

- In the case of VC sets of classifiers (hence this also applies to ensembles), remember we proved

$$\widehat{\mathcal{R}}(B_R) \leq \sqrt{\frac{2 \log H_{\mathcal{F}}(S)}{n}} ,$$

  where $H_{\mathcal{F}}(S)$ is the shattering coefficient on the sample S.