Lecture 5: VC bound and margin bound

Announcements

- Homework 1 due Thursday 23:59
- Moodle finally has forums to ask questions re HW or lecture (just realized yesterday)
- Project sign-ups Monday 14:00 - find your partner on moodle If you want to present a paper not on the list, please double check with us.

Feedback compilation

- Good: interactivity, intuition
- can be improved: handwriting, references to some results that are not explicitly noted in MW (adding some from SS), more intuition before proof but also more proof details

About project choice

1. Identify and motivate problem - why should I / the community care? Including literature review (done-ish)
2. "Detective hat": Intuitive (not just technical level) understanding of proof, assumptions, statement in depth
3. "Reviewer hat": Which relevant questions does it shed light on and does the paper answer/shed light on it? How significant is the addition of this paper compared to existing literature? This is a key step towards Step 4.
4. "Researcher hat": What are **interesting, impactful** follow-up questions they did not answer and would be interesting and perhaps feasible to pursue?
5. Break down the identified follow-up problem into feasible chunks (e.g. lemmas, experiments) and optionally show your attempts to tackle the first few steps.

Outline for today

- VC bound and proof
- Rademacher contraction
- Interactive: Proof using the ramp loss and contraction (students)

Recap: Massart's lemma

Note: in this lecture, we often write $z^n := z_1^n$ and the same for x.

Last time, we bounded the Rademacher for function classes $\mathcal{F}$ that induce a finite set $\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \ldots, \ell(Z_n; f)) : f \in \mathcal{F}\}$ using Massart's lemma

Lemma (Massart, SS Lemma 26.8)

For n points $Z^n := \{Z_1, \ldots, Z_n\}$, let all $h : \mathcal{Z} \to \{0,1\}$ and $\mathcal{H}(Z^n) := \{(h(Z_1), \ldots, h(Z_n)) : h \in \mathcal{H}\}$ with cardinality $|\mathcal{H}(Z^n)|$.

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) \leq \sqrt{\frac{2 \log |\mathcal{H}(Z^n)|}{n}}$$

- $|\mathcal{H}(Z^n)|$ corresponds to # labelings for Z^n induced by $\mathcal{H}$
- if $|\mathcal{H}(Z^n)|$ grows exponentially $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) = O(1)$

VC bound

We now use Massart to upper bound the generalization gap $R(f) - R_n(f)$ for function classes of finite VC dimension, where $|\mathcal{H}(Z^n)|$ does not grow exponentially in n for any Z^n.

Recap **definition VC dimension** for binary classification:

Definition (VC dimension)

Biggest $n \in \mathbb{N}$ s.t. there exists $Z^n \in \mathcal{Z}^n$ with $\mathcal{H}(Z^n) = \{0,1\}^n$

Function classes $\mathcal{F}$ with finite VC dimension can make $\mathcal{H}$ Glivenko-Cantelli, i.e. $\mathcal{R}_n(\mathcal{H}) = o(1)$. More specifically:

Theorem (uniform VC bound)

If $\mathcal{H}$ has VC dimension d_{VC}, w/ prob $\geq 1 - \delta$ for any estimator $f \in \mathcal{F}$

$$\mathbb{P}(yf(X) < 0) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i f(x_i) < 0} + 4\sqrt{\frac{d_{VC} \log(n+1)}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Proof of VC bound 1

Now we first prove a high-probability upper bound for the population 0-1 loss $\ell((x,y);f) = \mathbb{1}_{yf(x)<0}$ for finite function classes $\mathcal{F}$.

Plugging in the definition of the loss, using the uniform law, we get

$$\mathbb{P}(Yf(X) < 0) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i)<0} + 2\mathcal{R}_n(\mathcal{H}) + c\sqrt{\frac{\log(1/\delta)}{n}} \qquad (1)$$

for some universal constant c. The proof uses the uniform law (U.L.)

$$\begin{aligned} R(f) - R_n(f) &= \mathbb{E}\ell((x,y);f) - \frac{1}{n}\sum_{i=1}^{n} \ell((x,y);f) \\ &= \mathbb{P}(yf(x) < 0) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i)<0} \\ &\leq \sup_{f\in\mathcal{F}} R(f) - R_n(f) \overset{U.L.}{\leq} 2\mathcal{R}_n(\mathcal{H}) + c\sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

Proof of VC bound 2

- Note that $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H}) \leq \sup_{Z^n} \tilde{\mathcal{R}}_n(\mathcal{H})$ (this is crude!)
- Further by Massart, $\sup_{Z^n} \tilde{\mathcal{R}}_n(\mathcal{H}) \leq \sup_{Z^n} \sqrt{\frac{2\log|\mathcal{H}(Z^n)|}{n}}$ yielding

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log \sup_{Z^n} |\mathcal{H}(Z^n)|}{n}} \qquad (2)$$

(loose since distribution independent!)

Furthermore, we have the following upper bound on the size of $\mathcal{H}(Z^n)$

Lemma (Sauer-Shelah, MW Prop 4.18.)

If $\mathcal{F}$ has VC dimension d_{VC}, then for any $Z^n = Z_1, \ldots, Z_n$ we have growth function $N_{\mathcal{H}}(n) := \sup_{Z^n \in \mathcal{Z}^n} |\mathcal{H}(Z^n)| \leq (n+1)^{d_{VC}}$ for all $n \geq d_{VC}$.

Plugging Sauer-Shelah into eq. 2, and that into eq. 1 in the uniform law to yield result

Empirical Rademacher complexity - notation

In the following, we will slightly abuse notation and write the more general empirical Rademacher complexity for $\mathbb{T} \subset \mathbb{R}^n$ as

$$\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^{n} \epsilon_i \theta_i.$$

Note that hence we can write $\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$ for $\tilde{\mathcal{R}}_n(\mathcal{H})$.

The following lemma can connect the empirical Rademacher comp. of a function class $\widetilde{\mathcal{F}}$ to the empirical Rademacher comp. of a specific loss $\ell : \mathbb{R} \to \mathbb{R}$ acting on a function class, specifically when $\mathcal{H} = \ell \circ \widetilde{\mathcal{F}}$

First note that for $\mathbb{T} = \widetilde{\mathcal{F}}(Z^n)$ we can write the empirical Rademacher complexity in two ways (abusing notation)

$$\tilde{\mathcal{R}}_n(\ell \circ \widetilde{\mathcal{F}}) = \mathbb{E} \sup_{\tilde{f} \in \widetilde{\mathcal{F}}} \sum_{i=1}^{n} \epsilon_i \ell(\tilde{f}(Z_i)) \text{ same as}$$

$$\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^{n} \epsilon_i \ell(\theta_i)$$

Rademacher contraction

In the case of classification, we often have a loss of the form (again, slightly abusing notation) $\ell(Z_i, f) = \ell(Y_i f(X_i))$ and can define $\tilde{f}(Z_i) = Y_i f(X_i)$.

The following lemma holds for general losses $\boldsymbol{\ell} : \mathbb{R}^n \to \mathbb{R}^n$ (again, abuse of notation) where the loss may differ for each element, with $\boldsymbol{\ell}(\theta) = (\ell_1(\theta_1), \ldots, \ell_n(\theta_n))$ with $L-$Lipschitz $\ell_j : \mathbb{R} \to \mathbb{R}$, i.e.

$$|\ell_j(a) - \ell_j(b)| \leq L|a - b| \text{ for all } a, b \in \mathbb{R}.$$

Lemma (Rademacher contraction, SS Lemma 26.9)

For any $\mathbb{T} \subset \mathbb{R}^n$ and $\boldsymbol{\ell} : \mathbb{R}^n \to \mathbb{R}^n$ with univariate L-Lipschitz functions it holds that

$$\tilde{\mathcal{R}}_n(\boldsymbol{\ell} \circ \mathbb{T}) \leq L \tilde{\mathcal{R}}_n(\mathbb{T})$$

In the following when $\ell_i = \ell$ for all i, then $\tilde{\mathcal{R}}_n(\boldsymbol{\ell} \circ \mathbb{T}) = \tilde{\mathcal{R}}_n(\ell \circ \mathbb{T})$ as in the previous slide.

Skipped during lecture: Proof ingredients

Let ϵ be the vector of n i.i.d. Rademacher r.v. and define the shorthand $\epsilon_{2:n} = (\epsilon_2, \ldots, \epsilon_n)$ and same for θ.

The following holds for all n

- Key 1: de-symmetrize using the tower property: For any g we have $\mathbb{E}_\epsilon g(\epsilon) = \mathbb{E}_{\epsilon_1}[\mathbb{E}[g(\epsilon)|\epsilon_1]] = \frac{1}{2}\mathbb{E}[g(\epsilon)|\epsilon_1 = 1] + \frac{1}{2}\mathbb{E}g(\epsilon)|\epsilon = -1]$
- Key 2: Lipschitz property $\ell_i(\theta_i) - \ell_i(\tilde{\theta}_i) \leq L|\theta_i - \tilde{\theta}_i|$ for all i
- Key 3: For each ϵ we can define $h(\theta_{2:n}) = \sum_{i=2}^n \epsilon_i \ell_i(\theta_i)$. One can prove via contradiction that

$$\sup_{\theta,\tilde{\theta}\in\mathbb{T}} |\theta_1 - \tilde{\theta}_1| + h(\theta_{2:n}) + h(\tilde{\theta}_{2:n}) = \sup_{\substack{\theta,\tilde{\theta}\in\mathbb{T}\\ \theta_1\geq\tilde{\theta}_1}} \theta_1 - \tilde{\theta}_1 + h(\theta_{2:n}) + h(\tilde{\theta}_{2:n})$$

Skipped during lecture: R.C. contraction proof

$$n\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) = \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} \sum_{i=1}^n \epsilon_i \ell_i(\theta_i)$$

$$\overset{1.}{=} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta\in\mathbb{T}} \ell_1(\theta_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta}\in\mathbb{T}} -\ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$= \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta,\tilde{\theta}\in\mathbb{T}} \ell_1(\theta_1) - \ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{2.}{\leq} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta,\tilde{\theta}\in\mathbb{T}} L|\theta_1 - \tilde{\theta}_1| + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{3.}{=} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta\in\mathbb{T}} L\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta}\in\mathbb{T}}(-L\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{1.}{=} \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} L\epsilon_1\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i)$$

Use the same argument for the RHS inductively on each coordinate. □

Mimicking proof-based research in collaboration

- Learning objectives: Both for actual guarantees and presentation, collaboration
  1. Get intuition why a problem / conjecture should be true
  2. Break down a proof to parts
  3. Prove individual parts
- Matching questions in the interactive session today
  1. Intuitively why should enforcing a large margin yield better generalization? Show graphically (no right or wrong)
  2. Given contraction inequality, ramp loss and Rademacher complexity for linear functions, prove the margin bound
  3. Prove Rademacher complexity for linear function class

Instructions

- Groups:
  - We will divide the class into three groups of ≈ 4 people each.
  - Each group will solve one of the three questions jointly.
  - Once you know your group, choose a representative to present later
- Group work:
  - 15 minutes of discussion to solve the question - if done early, feel free to solve another groups' question
  - Another 5 minutes to prepare the representative's blackboard presentation
- Final presentation
  - 30 minutes of 3 short presentations (7 min presentation, 3 min Q&A)
  - Introduce yourself and group members by names
  - Present your results.

Primer on margins for linear classifiers

- Class of linear classifiers $\mathcal{F} = \{f : f(x) = w^\top x \; w \in \mathbb{R}^d\}$
- Intuition in introductory lectures for linearly separable data: large minimum distance to the boundary is good that can be computed as

$$d_{\min} = \min_i y_i \frac{w^\top x_i}{\|w\|_2}$$

  where $\min_i y_i \langle w, x_i \rangle$ is called the margin
- Can obtain set of maximizing directions by solving

$$\max_{\gamma, w} \gamma \text{ s.t. } y_i \langle \frac{w}{\|w\|_2}, x_i \rangle \geq \gamma$$

  which for bounded $\|w\|_2 \leq B$ is the same as solving

$$\max_{\gamma', \|w\|_2 \leq B} \gamma' \text{ s.t. } y_i \langle w, x_i \rangle \geq \gamma'$$

- We will look the generalization performance of feasible w with $\|w\|_2 \leq B$ which achieve a margin of at least some γ

Margin bound for binary classification

Key ingredient of proof (in interactive session)

Definition (ramp loss)

The ramp loss ℓ_γ is defined as

$$\ell_\gamma(u) = \begin{cases} 1 & u \in (-\infty, 0) \\ 1 - \frac{u}{\gamma} & u \in [0, \gamma] \\ 0 & u \in (\gamma, \infty) \end{cases}$$

and $\frac{1}{\gamma}$-Lipschitz.

Margin bound for linear classifiers

Definitions

- Set of linear functions $\mathcal{F}_B = \{f(x) = \langle w, x \rangle : \|w\|_2 \leq B\}$
- Define the risk $R_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $R^\gamma(f) = \mathbb{E}_{X,Y}\mathbb{1}_{Yf(X)\leq\gamma}$

Assumption (A): Boundedness of covariates $\mathbb{P}(\|x\|_2 \leq D) = 1$

Theorem (margin bound for linear classifiers)

If the assumptions are valid for any fixed γ, w/ prob. at least $1-\delta$, for any $f \in \mathcal{F}_B$ we have

$$R^0(f) = \mathbb{P}[y \neq sign(f(x))] \leq R_n^\gamma(f) + \frac{2DB}{\gamma\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

for some constant $c > 0$.

Solution: Proof of margin bound for linear classifiers

1. First we prove the following lemma

Lemma (uniform law with margin loss)

For $\mathcal{F}$ symmetric, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} R^0(f) - R_n^\gamma(f) \geq \frac{2}{\gamma}\mathcal{R}_n(\mathcal{F}) + t\right) \leq e^{-cnt^2}$$

2. Then we note that the class of linear functions $\mathcal{F}_B$ is symmetric and

Lemma (Rademacher complexity of bounded linear function class)

For $\mathcal{F}_B$ the empirical Rademacher complexity for specific $x_1, \dots x_n$ is

$$\tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) \leq \frac{B \max_i \|x_i\|_2}{\sqrt{n}}$$

so that $\mathcal{R}_n(\mathcal{F}_B) \leq \sup_{x_1^n \in \mathcal{X}_1^n} \tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) \leq \frac{BD}{\sqrt{n}}$

3. Plugging in $t = c\sqrt{\frac{\log(1/\delta)}{n}}$ then yields the theorem. □

Solution: Proof of uniform law with margin loss

Define $R_{\ell_\gamma}(f) := \mathbb{E}_{(X,Y)}\ell_\gamma(Yf(X))$ and $R_{\ell_\gamma,n}(f)$ its empirical version. We first use the uniform law to bound $R_{\ell_\gamma}(f)$.

1. In particular, given $z_i = (x_i, y_i)$, define $\widetilde{\mathcal{F}}(z_1^n)$ by $\tilde{f}(z_i) = y_i f(x_i)$ for $f \in \mathcal{F}$. Because $\mathcal{F}$ is symmetric, we have $\widetilde{\mathcal{F}}(z_1^n) = \mathcal{F}(x_1^n)$
2. Defining $\mathcal{H}(z_1^n) = \{\ell_\gamma(\cdot, f) : f \in \mathcal{F}\}$ the Rademacher complexity reads
$$\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) = \tilde{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{F}(x_1^n)).$$
3. The contraction inequality implies $\tilde{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{F}(x_1^n)) \leq \frac{1}{\gamma}\tilde{\mathcal{R}}_n(\mathcal{F}(x_1^n))$ and the same holds when taking expectations
4. The uniform law then yields that w.p. $\geq 1 - e^{-cnt^2}$
$$\sup_{f\in\mathcal{F}} R_{\ell_\gamma}(f) - R_{\ell_\gamma,n}(f) \leq \frac{2}{\gamma}\mathcal{R}_n(\mathcal{F}) + t$$
5. The lemma follows by noting that for every $\gamma > 0$ and any f it holds that $R^0(f) \leq R_{\ell_\gamma}(f)$ and $R_{\ell_\gamma,n}(f) \leq R_n^\gamma(f)$.

Solution: Rademacher complexity for linear classes

Proof of lemma via direct calculation

We utilize the fact that $\|x\|_2 = \sqrt{\|x\|_2^2}$ and that $\sqrt{\cdot}$ is a concave function whence Jensen's inequality yields

$$n\tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) = \mathbb{E}_\epsilon \sup_w \sum_i \epsilon_i w^\top x_i \leq B\mathbb{E}_\epsilon \| \sum_i \epsilon_i x_i \|$$
$$= B\sqrt{\mathbb{E}_\epsilon \| \sum_i \epsilon_i x_i\|^2} = B\sqrt{\sum_i \|x_i\|^2} \leq B\sqrt{n} \max_i \|x_i\|_2$$

In contrast: Rade. Comp. via VC Dimension

1. VC dimension of a class of linear classifiers (without bias term!) in R^d is d ($d_{\mathrm{VC}} \geq d$ is clear, $d_{\mathrm{VC}} \leq d$ via construction using linear dependence for $d+1$ points)
2. Then, using the VC bound we would obtain a bound of the order $\sqrt{\frac{d\log(n+1)}{n}}$, which is generally much larger then the dimension independent B.

References

- Massart, Rademacher for classification: Shalev-Schwartz & Ben-David Chapter 26