

Lecture 5: VC bound and margin bound

1 / 14

Announcements

Projects released

Feedback compilation:

- Cardinality of \mathcal{H} , VC dimension intuition - will explain again in recap
- How everything discussed in class is related (incorporated into exercise session later today)
- Notation - how can this be improved?

Outline today:

- Recap: Massart's Lemma, VC bound and proof
- Exercise (interactive) session: Proof using the ramp loss and contraction (students)

2 / 14

Recap: Set $\mathcal{H}(Z^n)$

- This lecture, we often write $z^n := z_1^n$ and the same for x .
- We write $\mathcal{H}(z^n) = \{(h(z_1), \dots, h(z_n)) : h \in \mathcal{H}\}$ which is equal to $\{(\ell(z_1; f), \dots, \ell(z_n; f)) : f \in \mathcal{F}\}$ if $\mathcal{H} = \{h(\cdot) : \ell(\cdot; f) \text{ for } f \in \mathcal{F}\}$
- Both Rademacher complexity and VC dimension measure how “flexible” the function class is to fit any noise variable combination
- We now consider the case of classification where $h : \mathcal{Z} \rightarrow \{0, 1\}$, hence $\mathcal{H}(z^n) \subset \{0, 1\}^n$
- Then the cardinality of $\mathcal{H}(z^n)$, denoted by $|\mathcal{H}(z^n)|$, corresponds to number of labelings for z^n induced by \mathcal{H} . This quantity bounds the empirical Rademacher complexity via Massart’s Lemma (next slide).
- The more “flexible” \mathcal{F} , we can find z^n with $|\mathcal{H}(z^n)| = 2^n$ for larger n

3 / 14

Recap: Massart’s lemma

Lemma (Massart, SS Lemma 26.8)

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

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

- if $|\mathcal{H}(Z^n)|$ grows exponentially $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) = O(1)$
- Proof sketch:
 - Step 1: subgaussianity of $\frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i)$
 - Step 2: Use expectation of the max of N zero-mean subgaussian R.V. X_1, \dots, X_N with sub-gaussian parameter σ

$$\mathbb{E} \max_{i=1..N} X_i \leq \sqrt{2\sigma^2 \log N}$$

4 / 14

Proof of Massart's lemma (some more details)

- Step 1: subgaussianity of $\frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i)$
 - For Rademacher ϵ_i and any Z_1^n we have that $\theta_i := h(Z_i) \in \{0, 1\}$.
 - Then $\frac{1}{n} \epsilon^\top \theta$ is zero-mean and $\frac{1}{\sqrt{n}}$ sub-gaussian. This follows from the fact that $[a_i, b_i]$ bounded r.v. are $[b_i - a_i]/2$ subgaussian.
- Step 2: Use the fact from HW 1 that, for N zero-mean subgaussians X_1, \dots, X_N with sub-gaussian parameter σ

$$\mathbb{E} \max_{i=1..N} X_i \leq \sqrt{2\sigma^2 \log N}$$

Here, $N = \mathcal{H}(Z^n)$ the number of different vectors $(h(Z_1), \dots, h(Z_n))$

5 / 14

Recap: VC dimension and VC bound

Definition (VC dimension)

\mathcal{H} has VC dimension d_{VC} if d_{VC} is the largest n for which there exists $z^n \in \mathcal{Z}^n$ with $\mathcal{H}(z^n) = \{0, 1\}^n$ or in other words

$$d_{VC}(\mathcal{H}) = \max\{n : \sup_{z^n \in \mathcal{Z}^n} |\mathcal{H}(z^n)| = 2^n\}$$

- Bounded VC dimension: the *growth function* $\sup_{z^n \in \mathcal{Z}^n} |\mathcal{H}(z^n)|$ (as a function of n) is not exponential as $n \rightarrow \infty$
- Infinite VC dimension: For any n points can “achieve” all labelings - decision trees, indicators of finite subsets, all functions

Theorem (uniform VC bound)

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

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

6 / 14

Proof of VC bound part 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} .

- By definition of the loss and using the uniform law (U.L.), we get

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

for some universal constant c , since

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

7 / 14

Proof of VC bound part 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

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

(loose since distribution independent!)

- Finally, 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 $Z^n = Z_1, \dots, Z_n$ 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

8 / 14

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

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

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

9 / 14

Margin bound for binary classification

Key ingredient of proof (in interactive session)

Definition (ramp loss)

The ramp loss l_γ is defined as

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

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

10 / 14

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

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

for some constant $c > 0$.

11 / 14

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

12 / 14

Instructions

- Groups:
 - We will divide the class into four groups of ≈ 3 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 presentations
 - 40 minutes of 4 short presentations (7 min presentation, 3 min Q&A)
 - Introduce yourself and group members by names
 - Present your results.

13 / 14

References

- Massart, Rademacher for classification: Shalev-Schwartz & Ben-David Chapter 26
- Margin bound & structural risk minimization: Shalev-Schwartz, Ben-David: Chapter 7, 26

14 / 14