Lecture 3: Azuma-Hoeffding and uniform law

Announcements

- HW due next Thursday 23:59, write it up entirely independently yourself
- You can check paper suggestions for project work already on the project website (link to a googlesheet)
- Lec2 slides updated regarding boundedness of martingale difference sequence & Azuma-Hoeffding (will explain again today)
- Goal of in-class lecture: cannot deliver the details of each proof completely, but primarily intuition - expect to fully understand and digest after reading the book & doing homework

Plan today

- Review of proof of uniform tail bound
- Warm-up exercise: using Azuma-Hoeffding for online learning "excess risk"
- Proof of Azuma-Hoeffding
- Uniform law with Rademacher complexity
- Intuition of Rademacher complexity

Recap: Main tail bound

- $\{Z_i\}_{i=1}^n$ are training points $\overset{iid}{\sim} \mathbb{P}$, estimator $\widehat{f}_n \in \mathcal{F}$ trained on them
- We use Z both for the collection $Z = \{Z_i\}_{i=1}^n$ and a single random vector $Z \sim \mathbb{P}$ which should be clear from the context
- Goal: want to prove that $R(\widehat{f}_n) - R_n(\widehat{f}_n) \leq \sup_{f\in\mathcal{F}} \mathbb{E}\ell(Z;f) - \frac{1}{n}\sum_{i=1}^n \ell(Z_i;f) =: g_n(Z)$ small with probability at least $1-\delta$

Theorem (Uniform tail bound)

For b-unif. bounded ℓ, it holds that

$$\mathbb{P}(\sup_{f\in\mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

Recap: What we can do with the tail bound

Using the short-term $\text{Res}(n, \mathcal{F}) := \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)]$ We immediately obtain

$$\mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \leq \text{Res}(n, \mathcal{F}) + t) \geq 1 - e^{-\frac{nt^2}{2b^2}}$$

This is a "high probability" bound in the sense that with probability at least $1 - \delta$ we have

$$\sup_{f \in \mathcal{F}} R(f) - R_n(f) \leq b\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \text{Res}(n, \mathcal{F})$$

Recap: Proof of tail bound (w/o martingale speak)

Approach: Upper bound $\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t)$ by following

1. If loss ℓ b-uniformly bounded, then $g_n = \sup_{f \in \mathcal{F}} \mathbb{E}\ell(Z, f) - \frac{1}{n}\sum_{i=1}^n \ell(Z_i, f)$ satisfies bounded difference property with $\sigma_i = \frac{2b}{n}$ for all i
2. For any g_n, we can decompose $g_n(Z) - \mathbb{E}g_n(Z) = \sum_{i=1}^n D_i$ $D_i = \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_i] - \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_{i-1}]$
3. Then, D_i satisfies that for any z_1^{i-1} there are some a_i, b_i with $b_i - a_i \leq \sigma_i$ such that $D_i|Z_1^{i-1} = z_1^{i-1} \in [a_i, b_i]$.
4. show how for such D_i (bounded martingale diff sequence) we have $\sum_{i=1}^n D_i$ concentrates around its expectation $\mathbb{E}D_i = 0$, i.e.

   $\mathbb{P}(\sum_{i=1}^n D_i > t) \leq= e^{-\frac{2t^2}{\sum_{i=1}^n \sigma_i^2}} \leq e^{-\frac{nt^2}{2b^2}}$ [Azuma Hoeffding]

Note: 2-4 proves McDiarmid using Azuma-Hoeffding, 2-3 prove that assumptions for Azuma-Hoeffding hold.

Not shown, will show today: Azuma-Hoeffding

Recap: Azuma-Hoeffding

- Hoeffding: Simple concentration for average of n independent sub-Gaussian (e.g bounded) Z_i

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z > t) \leq \mathrm{e}^{-\frac{nt^2}{2\sigma^2}}$$

- Azuma-Hoeffding: "Advanced" concentration for average of a martingale difference sequence $\{D_i\}_{i=1}^n$ bounded in intervals of length $\sigma = \frac{c}{n}$

$$\mathbb{P}(\sum_{i=1}^{n} D_i > t) \leq \mathrm{e}^{-\frac{2t^2}{n\sigma^2}} = \mathrm{e}^{-\frac{2nt^2}{c^2}}$$

Theorem (Azuma-Hoeffding inequality, MW Cor. 2.20)

If for martingale difference sequence $\{(D_i, \mathcal{F}_i)\}_{i=1}^n$ it holds that $D_i|\mathcal{F}_{i-1}$ almost surely lies in an interval of length L_i for all i, then

$$\mathbb{P}(\sum_{i=1}^{n} D_i \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}}$$

Next, we gain some more intuition on Azuma-Hoeffding by applying it to a different problem related to online learning

Exercise Context I: Online learning setting

- $Z_1, \ldots, Z_n$ come in one at a time.
- At each point in time i you would like to output an estimator $\widehat{f}_{i-1}$ to predict on the next sample Z_i with small loss
- As a data scientists, we naturally consider functions that are trained using the previous examples $Z_1, \ldots, Z_{i-1}$. More formally, we assume $\widehat{f}_{i-1}$ is a *deterministic function* of the previous samples $Z_1, \ldots, Z_{i-1}$ (e.g. ERM but *does not have to be*!), i.e. measurable with respect to $\sigma(Z_1, \ldots, Z_{i-1}) = \mathcal{F}_{i-1}$.
- $\widehat{f}_0$ can be any data-independent arbitrary estimator, e.g. a randomly initialized model.
- Assume the minimizer $\widehat{f}_n := \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1} \ell(Z_i; f)$ exists

Exercise Context II: Online to batch conversion

- A standard quantity people want to keep small in online learning is the regret Reg_n, the average incurred loss of the sequence $\{\widehat{f}_i\}_{i=1}^n$ with the loss of $\widehat{f}_n$

$$\mathrm{Reg}_n = \sum_{i=1}^n \ell(Z_i; \widehat{f}_{i-1}) - \sum_{i=1}^n \ell(Z_i; \widehat{f}_n)$$

- Note: Bounding the actual Reg_n is a whole area of research and in many cases, good online learning algorithms exist
- Online-to-batch conversion exploits online learning algorithms with small regret to get estimator based on batch $Z_1, \ldots, Z_n$ with good generalization. For example, one can consider a random estimator that samples from the sequence of online estimators $\{\widehat{f}_i\}_{i=0}^{n-1}$ which
  - conditioned on the data are deterministic
  - has an average (over the sampling) a risk of $\frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1})$
- We will now prove a high probability bound on the "average" excess risk $\frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1}) - R(f^\star)$

Exercise: Bound on the average excess risk

With your neighbor, prove that with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \le \frac{1}{n}\mathrm{Reg}_n + \sqrt{\frac{8\log(1/\delta)}{n}} \qquad (1)$$

with $R(f) = \mathbb{E}\ell(Z; f)$ for $\ell \in [0,1]$ using the following steps

1. Step: Prove that $D_i = [\mathbb{E}_Z \ell(Z; \widehat{f}_{i-1}) - \ell(Z_i; \widehat{f}_{i-1})] + [\ell(Z_i; f^\star) - \mathbb{E}_Z \ell(Z; f^\star)]$ is a bounded martingale difference sequence
2. Step: Decompose the risk (by including terms with $\widehat{f}$ and using its optimality) and prove

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \le \frac{1}{n}\mathrm{Reg}_n + \frac{1}{n}\sum_{i=1}^n D_i$$

3. Step: Use Step 1 and Azuma-Hoeffding to prove the bound eq. 1

Solution: Proof of average excess risk bound

We use the following shorthands for simplicity:

- $R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n}\sum_{i=1}^n \ell(Z_i; \widehat{f}_{i-1})$
- $R(\{\widehat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\ell(Z; \widehat{f}_{i-1})$

1. Risk decomposition:

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \leq R(\{\widehat{f}_i\}_{i=0}^{n-1}) - R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) + \underbrace{R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) - R_n(\widehat{f}_n)}_{=\mathrm{Reg}_n}$$

$$+ \underbrace{R_n(\widehat{f}_n) - R_n(f^\star)}_{\leq 0 \text{ by optimality of } \widehat{f}} + R_n(f^\star) - R(f^\star)$$

2. D_i is a martingale difference sequence because

$$\mathbb{E} D_i | \mathcal{F}_{i-1} = 0$$

as Z_i is independent of $\widehat{f}_{i-1}$ and bounded a.s. by 4.

Check: The average excess risk over $\{\widehat{f}\}_{i=1}^n$ is similar in terms of rate for large n as long as $R(\widehat{f}_n)$ is bounded

Proof of Azuma-Hoeffding

1. First of all, we have for all sequences z_1^{i-1} that for some $b_i - a_i \leq L_i$

$$\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1} = z_1^{i-1}] \leq e^{\lambda^2 (b_i - a_i)^2/8} \leq e^{\lambda^2 L_i^2/8}$$

by the fact that R.V. bounded in an interval of length L_i are $L_i/2$ subgaussian (for the right constant check MW Exercise 2.4., for an easier proof for the wrong constant check MW Example 2.4.) and hence a.s. the random variable $\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1}] \leq e^{\lambda^2 L_i^2/8}$

2. If D_i are independent, we have $\mathbb{E} e^{\lambda \sum_{i=1}^n D_i} = \prod_{i=1}^n \mathbb{E} e^{\lambda D_i}$

3. Note that since D_i are $\mathcal{F}_i$-measurable by definition of martingale difference sequence, we have $\mathbb{E}[e^{\lambda D_i} | G] = e^{\lambda D_i}$ for all $G \in \mathcal{F}_i$

4. Now using the tower property (TP) of conditional expectations iteratively, we see that $\sum_{i=1}^n D_i$ is $\sqrt{\sum_{i=1}^n \frac{L_i^2}{4}}$-subgaussian:

$$\mathbb{E} e^{\lambda \sum_{i=1}^n D_i} \overset{(TP)}{=} \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]]$$

$$\overset{(3.)}{=} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}[e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]] \leq e^{\lambda^2 L_i^2/8} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i}] = e^{\lambda^2 \sum_{i=1}^n L_i^2/8}$$

Bounding Res$(n, \mathcal{F})$, Rademacher complexity

Today we use shorthand $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$ and write the uniform tail bound this way. Then we have

$$\sup_{f\in\mathcal{F}} \mathbb{E}\ell(Z, f) - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, f) = \sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

and it follows that

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq \text{Res}(n, \mathcal{F}) + t) \leq e^{-\frac{nt^2}{2b^2}} \qquad (2)$$

The next four sessions will be about how to bound Res$(n, \mathcal{F})$!

Step I (this week): we first use eq. 2 & that Res$(n, \mathcal{F})$ is bounded by

Definition (Rademacher complexity)

Given a function class $\mathcal{H}$ and distribution $\mathbb{P}$ on its domain $\mathcal{Z}$, for i.i.d. Rademacher R.V. ϵ_i, we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon,Z} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(Z_i)$$

Step II (next 2 weeks): We'll discuss how to bound $\mathcal{R}_n(\mathcal{H})$ as a function of $n, \mathcal{H}$

Step I: Uniform law with Rademacher complexity

Theorem (Uniform law for the risk, MW Thm 4.10.)

For b-unif. bounded $\mathcal{H}$, with prob. over the training data

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq e^{-\frac{nt^2}{2b^2}}$$

- By using $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$ we get

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon,Z} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(Z_i) = \mathbb{E}_{\epsilon,Z} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \ell(Z_i, f)$$

  and after showing Res$(n, \mathcal{F}) \leq 2\mathcal{R}_n(\mathcal{H})$, directly obtain our desired bound on $\sup_{f\in\mathcal{F}} R(f) - R_n(f)$

- Note if $\mathcal{R}_n(\mathcal{H}) = o(1)$, then $\sup_{f\in\mathcal{F}} R(f) - R_n(f) \overset{a.s.}{\to} 0$.
- Before the proof, we aim to gain some intuition for the quantity $\mathcal{R}_n(\mathcal{H})$ and how it may behave with different n and $\mathcal{H}$

References

Azuma-Hoeffding

- MW Chapter 2

Online to batch conversion with Azuma-Hoeffding

- https://home.ttic.edu/~tewari/lectures/lecture13.pdf

Uniform law and Rademacher complexity

- MW Chapter 4