

Lecture 6: Margin bounds and structural risk minimization

1 / 16

Announcements

- HW due on Wednesday, send to tobias dot wegel@inf.ethz.ch
- Project sign-up October 14th 2pm next week

Plan today

- Concept map team presentation
- Margin bound team presentations
- Structural risk minimization

2 / 16

Recap: Uniform law

Recall $\mathcal{H} = \ell \circ \mathcal{F}$

Theorem (Uniform law for the risk)

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

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

Our task was then to bound $\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n))$

$$\mathcal{R}_n(\mathcal{H}) := \mathbb{E}_z \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \epsilon_i h(z_i) =: \mathcal{R}_n(\mathcal{H})$$

Here, we write $\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$ (where we stress dependence on samples) for $\tilde{\mathcal{R}}_n(\mathcal{H})$ with a slight abuse of notation. More generally, for any set $\mathbb{T} \subset \mathbb{R}^n$ we define

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

3 / 16

Recap: Margin bound for linear classifiers

- Define 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 $R^\gamma(f) = \mathbb{E}_{X, Y} \mathbb{1}_{Y f(X) \leq \gamma}$. Note: If you're worried about support vectors for which $y_i f(x_i) = 1$, you can define these indicator using $< \gamma$ instead of $\leq \gamma$ and the arguments still go through

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 $c > 0$.

- The theorem holds for all fixed B, γ !

4 / 16

Interpretation of the theorem

- Recall the max-margin solution $\arg \max_{\|w\| \leq B} \min_i y_i \langle x, w_i \rangle$ for linearly separable data, where $\min_i y_i \langle x, w_i \rangle$ is called the margin of w - it minimizes $R_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \gamma}$ among all linear classifiers with $\|w\| \leq B$ for any fixed γ .
- Generally, it's not clear that the max margin solution would be "the best", but if there's a natural margin in the data distribution, then plugging it in the margin bound yields a small upper bound: If the true distribution is separated by a large margin, then choosing to use the bound with a large γ would give me a good meaningful bound for the max-margin solution since I can obtain small R_n^γ (see next slide)
- Note this probability bound holds for any fixed γ *separately*. In particular the high probability event is not uniformly over γ but this bounds the event for each γ . That is, you can't directly plug in the margin that you achieve with your algorithm (i.e. data-dependent) as γ , to the right hand side (see later slides today)

5 / 16

Interpretation: Using the uniform bound for SVM solution

- Let's recall the hard-margin SVM solution

$$w_{\text{SVM}} = \arg \min_w \|w\|_2 \text{ s.t. } y_i \langle w, x_i \rangle \geq 1 \quad \forall i$$

that is a scaled version of the max-margin solution

- What are generalization guarantees for $f_{\text{SVM}}(\cdot) = \langle w_{\text{SVM}}, \cdot \rangle$?

Assumption (B): In \mathbb{P} , classes are linearly separable (hence, non-noisy), i.e. $\exists w^*$ with smallest $\|w^*\|_2$ s.t. $\mathbb{P}(y \langle w^*, x \rangle \geq 1) = 1$

Corollary (non-uniform margin bound for hard-SVM)

Let assumptions (A) and (B) hold. Then w/ prob. at least $1 - \delta$, we have for $f_{\text{SVM}}(\cdot) = \langle w_{\text{SVM}}, \cdot \rangle$

$$\mathbb{P}[y \neq \text{sign}(f_{\text{SVM}}(x))] \leq \frac{2D \|w^*\|_2}{\sqrt{n}} + c \sqrt{\frac{\log(1/\delta)}{n}}$$

Proof: We use the margin bound theorem with $\gamma = 1$. Due to linear separability of \mathbb{P} , we know $f_{\text{SVM}} \in \mathcal{F}_{\|w^*\|_2}$ a.s. and by definition of f_{SVM} we have $R_n^1(f_{\text{SVM}}) = 0$.

6 / 16

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. \square

7 / 16

Solution: Proof of uniform law with margin loss

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

1. In particular, given $z_i = (x_i, y_i)$, define $\tilde{\mathcal{F}}(z_1^n)$ by $\tilde{f}(z_i) = y_i f(x_i)$ for $f \in \mathcal{F}$. Because \mathcal{F} is symmetric, we have $\tilde{\mathcal{F}}(z_1^n) = \mathcal{F}(x_1^n)$

2. Defining $\mathcal{H}(z_1^n) = \{l_\gamma(\cdot, f) : f \in \mathcal{F}\}$ the Rademacher complexity reads

$$\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) = \tilde{\mathcal{R}}_n(l_\gamma \circ \mathcal{F}(x_1^n)).$$

3. The contraction inequality implies $\tilde{\mathcal{R}}_n(l_\gamma \circ \mathcal{F}(x_1^n)) \leq \frac{1}{\gamma} \tilde{\mathcal{R}}_n(\mathcal{F}(x_1^n))$ and the same holds when taking expectations

4. The uniform law then yields that w.p. $\geq 1 - e^{-cnt^2}$

$$\sup_{f \in \mathcal{F}} R_{l_\gamma}(f) - R_{l_\gamma,n}(f) \leq \frac{2}{\gamma} \mathcal{R}_n(\mathcal{F}) + t$$

5. The lemma follows by noting that for every $\gamma > 0$ and any f it holds that $R^0(f) \leq R_{l_\gamma}(f)$ and $R_{l_\gamma,n}(f) \leq R_n^\gamma(f)$.

8 / 16

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

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

In contrast: Rade. Comp. via VC Dimension

1. VC dimension of a class of linear classifiers (without bias term!) in R^d is d ($d_{VC} \geq d$ is clear, $d_{VC} \leq d$ 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 than the dimension independent B .

9 / 16

Going back to interpreting the result

Recall the following:

- a linear classifier w has (data) margin $\text{margin}(w) = \min_i y_i \langle w, x_i \rangle$
- and distance of its support vectors to the boundary is $d_{\min}(w) = \text{margin}(w) / \|w\|_2$.
- the margin bound reads

$$\mathbb{P}[y \neq \text{sign}(\langle w, x \rangle)] \leq R_n^\gamma(w) + \frac{2DB}{\gamma\sqrt{n}} + c \sqrt{\frac{\log(1/\delta)}{n}}$$

If we could plug in $\gamma = \text{margin}(w)$ and $B = \|w\|_2$ for any data-dependent w within $\|w\|_2 \leq B$, then

$$\mathbb{P}[y \neq \text{sign}(\langle w, x \rangle)] \leq \frac{2D}{d_{\min}(w)\sqrt{n}} + c \sqrt{\frac{\log(1/\delta)}{n}}$$

where possibly $d_{\min}(w) \gg \frac{\gamma}{B}$.

10 / 16

Suboptimality of margin bound for hard-SVM

- For example recall the corollary: under assumptions (A), (B) with prob. $\geq 1 - \delta$ we have

$$\mathbb{P}[y \neq \text{sign}(\langle w_{\text{SVM}}, x \rangle)] \leq \frac{2D \|w^*\|_2}{\sqrt{n}} + c \sqrt{\frac{\log(1/\delta)}{n}}$$

$\|w^*\|_2$ on the RHS is data (but not distribution) independent!

- In particular, observe that

$$\|w^*\|_2 \geq \frac{\|w^*\|_2}{\text{margin}(w^*)} = \frac{1}{d_{\min}(w^*)} \geq \frac{1}{d_{\min}(w_{\text{SVM}})} = \|w_{\text{SVM}}\|_2,$$

i.e. $d_{\min}(w) \gg \frac{\gamma}{B}$.

Is there a bound where we can plug in $d_{\min}(w_{\text{SVM}})$?

11 / 16

Data-dependent hard-margin SVM bound

Yes, if we pay a little price in the probability (and another constant):

Theorem (data-dependent hard-margin SVM bound)

Under assumptions (A), (B), w/ prob. at least $1 - \delta$,

$$\mathbb{P}[y f_{\text{SVM}}(x) < 0] \leq \frac{2eD \|w_{\text{SVM}}\|_2}{\sqrt{n}} + c \sqrt{\frac{\log(2/\delta) + 2 \log(2 \log \|w_{\text{SVM}}\|_2)}{n}}$$

- The upper bound depends on the actual outcome of the algorithm that isn't constrained to a function class \mathcal{F}_B with a pre-fixed B !
- Proof relies on the concept of structural risk minimization! So we'll have a short interlude on that before going back to the proof

12 / 16

Interlude: Structural risk minimization

- Say we have a nested family of function spaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots$
- For each \mathcal{H}_k define the event

$$E_{k,h} := \left\{ R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_k)}{n}} + 2\mathcal{R}_n(\mathcal{H}_k) \right\}$$

- Now what can be say about the union $\mathcal{H} = \cup_{k \in \mathbb{N}} \mathcal{H}_k$?
- Define $k(h) = \min\{k : f \in \mathcal{H}_k\}$ which for each h finds the minimum set \mathcal{H}_k it belongs to

Corollary (of uniform law: structural risk minimization)

If $\mathbb{P}(\cap_{h \in \mathcal{H}_k} E_{k,h}) \geq 1 - \delta_k$ for each k and if $\sum_k \delta_k \leq \delta$, w/ prob. at least $1 - \delta$, uniformly over $h \in \mathcal{H}$ we have

$$R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_{k(h)})}{n}} + 2\mathcal{R}_n(\mathcal{H}_{k(h)})$$

13 / 16

Interlude: Proof of SRM Corollary

1. Observe that the high probability bound is over the event

$$\begin{aligned} & \cap_{h \in \mathcal{H}} \left\{ R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_{k(h)})}{n}} + 2\mathcal{R}_n(\mathcal{H}_{k(h)}) \right\} \\ & \supseteq \cap_{h \in \mathcal{H}} \cap_{k: h \in \mathcal{H}_k} E_{k,h} = \cap_{k \in \mathbb{N}} \cap_{h \in \mathcal{H}_k} E_{k,h} \end{aligned}$$

2. Using union bound we then get

$$\mathbb{P}(\cap_{k \in \mathbb{N}} \cap_{h \in \mathcal{H}_k} E_{k,h}) = 1 - \mathbb{P}(\cup_{k \in \mathbb{N}} \cup_{h \in \mathcal{H}_k} E_{k,h}^C) \geq 1 - \sum_{k \in \mathbb{N}} \mathbb{P}(\cup_{h \in \mathcal{H}_k} E_{k,h}^C)$$

3. Since further we assumed $\mathbb{P}(\cap_{h \in \mathcal{H}_k} E_{k,h}) \geq 1 - \delta_k$ (i.e. uniform law can be applied), finally $\mathbb{P}(\cap_{k \in \mathbb{N}} \cap_{h \in \mathcal{H}_k} E_{k,h}) \geq 1 - \sum_{k \in \mathbb{N}} \delta_k \geq 1 - \delta$

14 / 16

Proof of uniform margin bound theorem

1. For hard-SVM we define $\mathcal{F}_k = \{\langle w, \cdot \rangle : \|w\| \leq B_k\}$. Define

$$E_{k,f} = \{\mathbb{P}[y \neq \text{sign}(f(x))] \leq R_n^1(w) + \frac{2DB_k}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta_k)}{n}}\}$$

2. Since we know that $\mathbb{P}(\cap_{f \in \mathcal{F}_k} E_{k,f}) \geq 1 - \delta_k$ via the margin bound choosing $\gamma = 1$, we can use the same argument as for SRM:

- Choose $B_k = e^k$ and $\delta_k = \frac{\delta}{2k^2}$ where $\sum_k \delta_k \leq \delta$
- Then $k(w) = \lceil \log(\|w\|) \rceil$ and hence $B_{k(w)} \leq \|w\|e$ and $\frac{1}{\delta_{k(w)}} = \frac{2k(w)^2}{\delta} \leq \frac{2(\log \|w\| + 1)^2}{\delta}$
- Using the same argument as in the proof of SRM corollary, we can also show that with prob. at least $1 - \delta$ uniformly over all w

$$\mathbb{P}[yw^\top x < 0] \leq R_n^1(w) + \frac{2DB_{k(w)}}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta_{k(w)})}{n}}$$

3. Plugging in the quantities and w_{SVM} then yields the theorem. \square

15 / 16

References

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

Guarantees for Machine Learning L1-4 Concept Map

