GML 23 - Lecture 5 (Interactive Session): Proof of margin bound

Instructions

The aim of the interactive sessions is to collectively prove some relevant results from the literature.

- 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 questions)
  - Introduce yourself and group members by names
  - Present your results.

Question 1: Intuition for margin bound

For a classification problem given a training set $\{(x_i, y_i)\}_{i=1}^n$, we define $\min_i y_i w^\top x_i$ as the (unnormalized) margin and consider a set of linear functions with bounded norm $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : \|w\|_2 \leq B\}$.

First, not using formulas, the uniform law or Theorem 1 below - what's the intuition for why enforcing a large margin should lead to better generalization?

a) Show it graphically (no right or wrong)
b) How would a bound look like that captures your intuition?

Now parse Theorem 1 in Question 2.

c) How is your intuition reflected in Theorem 1?

Question 2: Prove margin bound given ingredients

We define the set of linear functions $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : \|w\|_2 \leq B\}$ and for any $f \in \mathcal{F}_B$ define $R_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $R^\gamma(f) = \mathbb{E}_{X,Y}\mathbb{1}_{Y f(X) \leq \gamma}$. Assume $\|x\|_2 \leq D$ with probability 1. In this question we will prove the following Theorem

Theorem 1 (margin bound)**.** *If the assumptions are valid for any fixed γ, w/ prob. at least $1-\delta$, for any f we have*

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

for some constant $c > 0$.

Definition 1 (ramp loss)**.**

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

Lemma 1 (Rademacher contraction)**.** *For any $\mathbb{T} \subset \mathbb{R}^n$ and $\ell : \mathbb{R}^n \to \mathbb{R}^n$ with univariate L-Lipschitz functions it holds that*

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

with $\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\theta_i$. Hint: The ramp loss upper bounds the 0-1 loss and is $\frac{1}{\gamma}$-Lipschitz

a) **Prove Theorem 1** using the ramp loss in eq. 1, contraction inequality (Lemma 1, discussed today) and the expression of the Rademacher complexity for linear function classes (Lemma 2)

Question 3: Prove Rademacher complexity of linear functions

Lemma 2 (for linear function class)**.** *For $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : \|w\|_2 \leq B\}$, the empirical Rademacher complexity is*

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

Hint: Use the fact that $\|x\|_2 = \sqrt{\|x\|_2^2}$ and that $\sqrt{\cdot}$ is a concave function

a) **Prove the Lemma.**

b) Determine the VC dimension for linear classifiers in $\mathbb{R}^d$. How does a bound on the Rademacher complexity using the VC dimension compare with the statement of the lemma?