

GML 25 - 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 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 presentation
 - 40 minutes of 4 short presentations (7 min presentation, 3 min questions)
 - Introduce yourself and group members by names
 - Present your results.

Question 0: Concept map for class material so far

Using the lecture slides and chapters 2, in Martin Wainwright's book, generate a concept map connecting the following words and then prepare a presentation of the map that helps you and your fellow students to understand how these concepts and proof ingredients are all related

concentration, uniform convergence/law, sub-Gaussian, train and test error, Azuma-Hoeffding inequality, McDiarmid inequality, Rademacher complexity, Chernoff bound, bounded R.V., empirical processes, generalization gap, empirical risk minimizer, model/hypothesis class, moment generating function, VC dimension, Massart's Lemma, population risk, symmetrization, excess risk

If you have time, also think about the connection to (but don't present): bias, variance, law of large numbers, asymptotic normality, maximum-likelihood estimator.

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 - generally we're taught the intuition in entry ML courses that a linear classifier with larger margin should perform better.

- Explain the intuition behind the statement that large-margin classifiers should lead to a smaller generalization error. Is this true in all classification settings (depending on sample size n , data generating distribution)?

Now parse Theorem 1 in Question 2.

- Explain the statement non-technically and intuitively. How is your intuition in a) reflected in Theorem 1? How can we use the theorem to upper bound the population risk of max-margin solutions?

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*

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

for some constant $c > 0$.

Definition 1 (ramp loss).

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

Lemma 1 (Rademacher contraction). *For any $\mathbb{T} \subset \mathbb{R}^n$ and $\ell : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 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 (last lecture) 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?