GML 23 - Lecture 12 (Interactive Session): Minimax lower bound for semi-supervised learning

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 ≈ 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:
  - 40 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.

Semi-Supervised Learning (SSL) is a learning framework in which the learning algorithm leverages both labeled and unlabeled data from the same distribution, in contrast to only labeled data (supervised learning) or only unlabeled data (unsupervised learning). Numerous empirical studies show that SSL can effectively harness information from both datasets, outperforming UL and SL. However, no such consensus exists yet in theory of machine learning.

In this interactive exercise, we will attempt to answer the following question:

Can semi-supervised classification algorithms simultaneously improve over the minimax rates of both supervised and unsupervised learning?

We will study this question in the context of linear classification for symmetric 2-Gaussian mixture models (GMMs). GMMs are a family of distributions that consist of two identical spherical Gaussians with opposite means and equal weights. We will consider the following model:

$$P_Y = \text{Unif}\{-1, 1\} \text{ and } P^{\theta^*}_{X|Y} = \mathcal{N}(Y\theta^*, I_d),$$

where the ground truth mean $\theta^* \in \mathbb{R}^d$ satisfies $\|\theta^*\|_2 = s$. s measures the "noisiness" of the classification problem: the larger s, the further apart the clusters (draw a picture). We will call the family of such distributions $\mathcal{P}^{(s)}_{2-\text{GMM}}$.

We will consider algorithms that take as input n_l labeled datapoints $\mathcal{D}_l \sim (P^{\theta^*}_{XY})^{n_l}$ or n_u unlabeled datapoints $\mathcal{D}_u \sim (P^{\theta^*}_{X})^{n_u}$, or both, and output an estimator $\hat{\theta} \in \mathbb{R}^d$. The test label of a point x will be predicted as $\text{sgn}(\langle \hat{\theta}, x \rangle)$. In this exercise, we focus on the **estimation error**

$$\mathcal{R}_{\text{estim}}(\theta, \theta^*) \coloneqq \|\theta - \theta^*\|_2.$$

We will prove the following tight lower bound on the estimation error:

Theorem 1 (SSL Minimax Rate for Estimation Error)**.** *For any* $0 < s \leq 1$ *the following holds when* $n_u \gtrsim (1/s)^2$, $n_l \gtrsim \frac{\log n_u}{s^2}$ *and* $d \geq 2$*:*

$$\inf_{\mathcal{A}_{SSL}} \sup_{\|\theta^*\|=s} \mathbb{E}[\mathcal{R}_{estim}(\mathcal{A}_{SSL}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \gtrsim \min\left\{s, \sqrt{\frac{d}{n_l + s^2 n_u}}\right\}.$$

We will prove Theorem 1 via Fano's method. The proof is divided into the following exercises:

Question 1: Fano's method for GMMs

Consider an arbitrary set of predictors $\mathcal{M} = \{\theta_i\}_{i=0}^M$. Prove the following:

$$\begin{aligned}&\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\theta^*\|_2=s} \mathbb{E}_{\mathcal{D}_l, \mathcal{D}_u}[\mathcal{R}_{\text{estim}}(\mathcal{A}_{\text{SSL}}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \geq \\ &\frac{1}{2} \min_{i,j \in [M], i \neq j} \|\theta_i - \theta_j\|_2 \left(1 - \frac{1 + n_l \max_{i \in [M]} D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) + n_u \max_{i \in [M]} D(P_X^{\theta_i} \| P_X^{\theta_0})}{\log M}\right),\end{aligned} \tag{1}$$

where $D(\cdot\|\cdot)$ denotes the KL divergence. *Hint:* first, prove that given a collection of distributions $\{P_1, ..., P_M\}$ and their mixture distribution $\overline{Q} = \frac{1}{M}\sum_{i=1}^M P_i$, it holds that

$$\frac{1}{M}\sum_{i=1}^M D(P_i \| \overline{Q}) \leq \frac{1}{M}\sum_{i=1}^M D(P_i \| Q)$$

for any other distribution Q (Exercise 15.11 in MW).

Question 2: Upper bounds on KL divergence for GMMs

Assume that you are given a packing $\{\theta_i\}_{i=0}^M$ which is constructed as follows: given positive absolute constants c_0 and C_0, we take a c_0-packing $\tilde{\mathcal{M}} = \{\psi_1, ..., \psi_M\}$ on the unit sphere S^{d-2} such that $|\tilde{\mathcal{M}}| \geq e^{C_0 d}$. For an absolute constant $\alpha \in [0, 1]$, we now construct the following packing:

$$\mathcal{M} = \left\{\theta_i = s \begin{bmatrix} \sqrt{1 - \alpha^2} \\ \alpha \psi_i \end{bmatrix}, \quad \psi_i \in \tilde{\mathcal{M}}\right\},$$

and define $\theta_0 = \big[s, 0, ..., 0\big]$.

1) Visualize the given packing and study its properties. Where are θ_0 and θ_i located? What is the distance between different elements of the packing? Is there an intuition for this particular choice? Discuss with your partner why this choice of a packing is better for use in (1) as compared to, for instance, a uniform packing on the sphere S^{d-1}.

2) Compute the KL divergence between two GMMs with identity covariance matrices, i.e. show that

$$D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) = \frac{1}{2}\|\theta_i - \theta_0\|_2^2 \leq \alpha^2 s^2, \text{ for all } i \in [M]. \tag{2}$$

Question 3: Proof of Theorem 1

Assume that additionally to (2), we have proven the following upper bound for the KL divergence between marginal distributions:

$$D(P_X^{\theta_i} \| P_X^{\theta_0}) \leq C\|\frac{1}{s}\theta_i - \frac{1}{s}\theta_0\|_2^2 \leq 2C\alpha^2 s^4. \tag{3}$$

Utilizing these two results as well as Question 1, prove Theorem 1. *(You might need to optimize over one of the constants.)*