Theoretical assignment 1; 8 points total + ≥ 7 points extra

Theoretical Deep Learning #2, MIPT

Let $\mathcal{F}$ be a set of functions $f : X \to \mathbb{R}$. Let $X_n = \{x_i\}_{i=1}^n$, where all $x_i \in X$, and $Y_n = \{y_i\}_{i=1}^n$, where all $y_i \in \{-1, 1\}$.

We say "$\mathcal{F}$ shatters X_n" if $\forall Y_n \; \exists f \in \mathcal{F} : \; \forall i \; f(x_i) y_i > 0$.

We say "$\mathcal{F}$ γ-shatters X_n" if $\exists B_n : \; \forall Y_n \; \exists f \in \mathcal{F} : \; \forall i \; (f(x_i) - b_i) y_i > \gamma$.

By definition, VC-dimension is a size of the largest shattered dataset:

$$\mathrm{VC}(\mathcal{F}) = \max(n : \; \exists X_n : \; \mathcal{F} \text{ shatters } X_n).$$

We define fat-shattering dimension analogously:

$$\mathrm{fat}_\gamma(\mathcal{F}) = \max(n : \; \exists X_n : \; \mathcal{F} \; \gamma\text{-shatters } X_n).$$

Problem 1

2 points extra.

Let $d, n \in \mathbb{N}$ and $n > d$. Prove that

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

From this, and from Sauer's lemma it follows that

$$\log \Pi(\mathcal{F}, n) \leq \mathrm{VC}(\mathcal{F})(1 + \log(n / \mathrm{VC}(\mathcal{F}))), \quad n > \mathrm{VC}(\mathcal{F}).$$

Problem 2

3 points + 5 points extra.

Let $x \in X = \mathbb{R}^d$. Let $\mathcal{F}$ be the class of linear classifiers on X:

$$\mathcal{F} = \{\langle w, \cdot \rangle + b, \; w \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

1. **3 points.** Prove that $\mathrm{VC}(\mathcal{F}) \geq d + 1$.
2. **5 points extra.** Prove that $\mathrm{VC}(\mathcal{F}) = d + 1$.

Problem 3

1 point.

In the lecture we have mentioned a class of functions (neural networks) with infinite VC-dimension and finite fat-shattering dimension. One can wonder if VC-dim is always not less than fat-shattering dim. However, this is not true.

Assume $\mathcal{F}$ is a set of all feed-forward fully-connected neural nets of the same architecture:

$$\mathcal{F} = \{x \mapsto W_L \sigma(W_{L-1} \dots \sigma(W_1 x)),\ W_l \in \mathbb{R}^{d_l \times d_{l-1}}\},$$

where $d_L = 1$, $x \in X \subset \mathbb{R}^{d_0}$, and σ is any element-wise non-linearity.

Prove that $\mathrm{VC}(\mathcal{F}) \leq \mathrm{fat}_\gamma(\mathcal{F})$ for any $\gamma > 0$.

Problem 4

0.5 point for each example, 2 points total + ≥ 0 points extra.

Name popular machine learning algorithms with corresponding function classes of infinite VC-dimension (and argue why it is infinite).

For example, SVM with linear kernel is a ML algorithm, the corresponding function class is a class of linear pedictors, and its VC-dim is $d+1$, where d is the number of features.

Problem 5

2 points.

Let $\mathcal{F}$ be a set of functions $f : X \to \mathbb{R}$. Let $X_n = \{x_i\}_{i=1}^n$, where all $x_i \in X$.

We call $\bar{\mathcal{F}}$ ϵ-net of $\mathcal{F}$ under l_1 norm wrt dataset X_n if

$$\forall f \in \mathcal{F}\ \exists \bar{f} \in \bar{\mathcal{F}} :\ \|f(X_n) - \bar{f}(X_n)\|_1 = \sum_{i=1}^n |f(x_i) - \bar{f}(x_i)| < \epsilon.$$

We call a covering number of $\mathcal{F}$ under l_1 norm wrt dataset X_n (and denote it $\mathcal{N}_1(\mathcal{F}, \epsilon, X_n)$) the size of minimal corresponding ϵ-net.

We define Rademacher complexity of function class $\mathcal{F}$ conditioned on dataset X_n as

$$\mathrm{Rad}(\mathcal{F}|X_n) = \mathbb{E}_{\Sigma_n} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right|,$$

where $\Sigma_n = \{\sigma_i\}_{i=1}^n$ and all $\sigma_i \sim U(\{-1, 1\})$.

Assume $\forall f \in \mathcal{F}\ 0 \leq f \leq 1$. Prove that $\forall X_n\ \forall \epsilon > 0$

$$\mathrm{Rad}(\mathcal{F}|X_n) \leq \frac{\epsilon}{n} + \sqrt{\frac{2}{n} \log(2\mathcal{N}_1(\mathcal{F}, \epsilon, X_n))}.$$

Hint: start with considering an ϵ-net $\bar{F}$ of $\mathcal{F}$ under l_1 norm wrt dataset X_n. Then, reduce $\mathrm{Rad}(\mathcal{F}|X_n)$ *to* $\mathrm{Rad}(\bar{\mathcal{F}}|X_n)$.