Mathematics of Machine Learning
Homework 9 - Solutions

Instructors: Afonso S. Bandeira & Nikita Zhivotovskiy
Course Coordinator: Pedro Abdalla Teixeira

May 7, 2021

Try to solve the questions before looking to the answers. Every item must be proved rigorously. Starred problems are harder.

Problem 1

Let $X = \mathbb{R}^2$ and $Y = \{0,1\}$. Let $\mathcal{F}$ be the set of classifiers corresponding to all concentric circles in the plane centered at the origin, precisely

$$f_r(x) := \begin{cases} 1, \|x\| \leq r \\ 0, \text{otherwise} \end{cases}$$

Prove that $\mathcal{F}$ is PAC-learnable and give an upper bound to the sample complexity.

Solution 1

Let f^* be the unknown classifier that we want to learn and let C denote the circle induced by f^*. A simple learning algorithm is sufficient here: Given the sample S, the algorithm returns the tightest circle $\hat{C}$ containing all the samples with label 1. To analyse the risk, we define $s^* := \inf\{s : \mathbb{P}(s \leq \|x\|_2 \leq r) < \varepsilon\}$ and the annulus $A := \{x : s^* \leq \|x\|_2 \leq r\}$, so clearly $\mathbb{P}(x \in A) \geq \varepsilon$. Observe that the error occurs when no sample falls in the annulus A. We now proceed as in homework 7,

$$\mathbb{P}(\hat{R}(f) > \varepsilon) = \mathbb{P}(\hat{C} \cap A = \emptyset) = \mathbb{P}(\forall i \in [n], X_i \notin A) \leq (1 - \varepsilon)^n.$$

It follows for $n = \lceil \frac{1}{\varepsilon} \log \frac{1}{\delta} \rceil$, the probability that the risk is larger than ε is at most δ, so the class is PAC-learnable with sample complexity at most $\lceil \frac{1}{\varepsilon} \log \frac{1}{\delta} \rceil$.

Problem 2

Assume the data is linearly separable with the margin γ. Let $\hat{f}_S$ be the classifier returned by the Perceptron algorithm after training over the sample S (drawn from some unknown distribution) with size n and running through it until the algorithm makes a pass over the sample with no mistakes. Give a bound for the expected risk via the Leave-One-Out argument.

Solution 2

Let $S^{'}$ be a sample of size $n+1$. For notation simplicity we refer to the Perceptron algorithm as P. Consider a point $x \in S^{'}$. If $\hat{f}_{S'/\{x\}}$ missclassifies x, then x is a support vector. Since we run the Perceptron algorithm until we get no mistakes on the sample, the classifier will be the same when classifying any point in the Leave One Out analysis. Observe that it is not the case if we pass only one time, i.e, if we pass only one time the order of the points matters. We now apply Theorem 15.3 from the lecture notes to obtain

$$\text{LOO}(S^{'}) = \frac{\sum_{i=1}^{n} \hat{f}_i(X_i) \neq f(X_i)}{n+1} \leq \frac{r^2}{\gamma^2(n+1)},$$

where r is the maximum $\|x\|$ (radius) of the sample $S^{'}$. By Theorem 14.1 from the lecture notes we get

$$\mathbb{E}_{X_1,\dots,X_n} R(P(S)) = \mathbb{E}_{X_1,\dots,X_{n+1}} \text{LOO}(P(S^{'})) \leq \frac{\mathbb{E}_{X_1\dots,X_{n+1}} r^2}{\gamma^2(n+1)}.$$

Problem 3

Let $F(t) := \mathbb{P}(X \leq t)$ be the cumulative distribution function of a random variable X and let $\hat{F}_n$ be the empirical cumulative distribution function with respect to an i.i.d sample $X_1, \dots, X_n$, i.e, consider i.i.d random variables $X_1, \dots, X_n$ with the same distribution of X, the empirical cumulative distribution with respect to such sample is given by

$$\hat{F}_n(t) := \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}}{n}.$$

(a) Prove that $\hat{F}_n$ converges uniformly to F in probability. That is,

$$\sup_{t\in\mathbf{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{P} 0.$$

(Hint: Apply the uniform law of large numbers together with the bound for the shatter function of intervals)

(b) * Prove that $\hat{F}_n$ converges uniformly to F almost surely. (Hint: Use Borell-Cantelli lemma)

Solution 3

(a) Let $\mathcal{A}$ denote the collection of events of the type $x \le t$ for $t \in \mathbb{R}$. By Theorem 17.2 in the lecture notes we know that

$$\mathbb{P}(\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \ge \varepsilon) \le 8\mathcal{A}(n)e^{-n\varepsilon^2/32}.$$

We also know that $\mathcal{A}(n) \le n + 1$ (lecture notes), therefore it is easy to see that $\mathbb{P}(\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \ge \varepsilon)$ goes to zero as n goes to infinity.

(b) By letter "a" we obtain that

$$\mathbb{P}(\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \ge \varepsilon) \le 8\mathcal{A}(n)e^{-n\varepsilon^2/32} \le 8e^{-n\varepsilon^2/64},$$

where in the last inequality we used a crude bound: For sufficient large n, we have that $e^{\log(n+1)} \le e^{\frac{n\varepsilon^2}{64}}$. Motivated by the last observation we define n^* to be the smallest n such that $\log(n+1) \le \frac{n\varepsilon^2}{64}$. Now we have that

$$\sum_{n=1}^{\infty} \mathbb{P}(\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \ge \varepsilon) \le \sum_{n=1}^{n^*} 8(n+1)e^{-n\varepsilon^2/32} + \sum_{n=n^*}^{\infty} 8e^{-n\varepsilon^2/64}.$$

The first term in the right hand side is finite as the sum runs over finite terms. The second term in the right hand side is also finite because $8e^{-n\varepsilon^2/64}$ is integrable in the interval (n^*, ∞). We conclude that $\sum_{n=1}^{\infty} \mathbb{P}(\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \geq \varepsilon)$ is finite. By Borell-Cantelli lemma, we get convergence almost surely.

The result above (letter "b") is known as the classical Glivenko-Cantelli theorem. It illustrates a deep connection between empirical process theory and statistical learning theory. It also reveals the power of concentration inequalities, once we prove concentration with exponential decay, we can use crude bounds to get precise asymptotic results.