Homework 13
HDP KNU+ Fall 2022

Hints are in the back of this homework set.

Bacteria reacts positively or negatively to d different factors, such as acidity, temperature, availability of food, etc. While d can be huge, only few factors are important for the life of bacteria, say $s \ll d$ of them. We want to determine which factors are important.

To this end, we conduct an experiment with n independent bacteria. For each bacteria, we record all d factors, and whether the bacteria thrived or died.

We model this mathematically by assuming that all d factors are independent $N(0, 1)$ random variables. Assume that bacteria lives if the sum of all *important* factors is positive, and dies if this sum is negative.

In the following two problems, we find the set of important factors from $n = O(s \log d)$ bacteria. That's great! Since the logarithmic function grows slowly, this sample size n almost does not depend on the total total number of factors d, which can be huge.[1]

Problem 1 (Sparse learning)

(a) Express the experiment in the context of supervise learning. Namely, represent the training data as $(X_1, Y_1), \dots, (X_n, Y_n)$, where the vector of factors of i-th bacteria is $X_i \sim N(0, I_d)$, the (unknown) vector $w^* \in \{0, 1\}^d$ encodes which factors are important and which are not, and the state of i-th bacteria is

$$Y_i = \operatorname{sign}\langle w^*, X_i \rangle.$$

Introduce the hypothesis class $\mathcal{H}$ so that

$$|\mathcal{H}| = \binom{d}{s} \le d^s.$$

(b) Assume that $n \ge C s \log d$ with a sufficiently large absolute constant C. Show that the generalization error of the ERM algorithm satisfies

$$R(h_n^*) \le 0.001$$

with probability at least 0.99.

Problem 2 (Sparse learning continued)

(a) To prepare for the next step, prove the following inequality for $g \sim N(0, I_d)$ and any fixed pair of unit vectors $u, v \in \mathbb{R}^d$:

$$0.878 \|u - v\|_2^2 \le \mathbb{E} \left(\operatorname{sign}\langle u, g \rangle - \operatorname{sign}\langle v, g \rangle \right)^2.$$

---

[1] There are two caveats though: (a) our additive model may be too simplistic, and (b) our ERM algorithm can be too slow. For practical algorithms, see *sparse dictionary learning*.

(b) Deduce from the previous two parts (Problem 1(b) and Problem 2(a)) that

$$\|w_n^* - w^*\|_2^2 \le 0.01s$$

where w^* is the unknown vector from (a), and w_n^* is the output of the ERM algorithm.

(c) Interpret (b) as stating that at most $0.01s$ coordinates of w_n^* and w^* can be different. Conclude that we can find the set of s important factors up to 1% of error.

Problem 3 (VC dimension: examples)

(a) Let $\mathcal{H}$ be the class of indicators of half-finite intervals, i.e. $\mathcal{H}$ consists of functions of the form $\mathbf{1}_{(-\infty,a)}$ and $\mathbf{1}_{(b,\infty)}$, where $a, b \in \mathbb{R}$. Prove that

$$\mathrm{vc}(\mathcal{H}) = 2.$$

(b) Let $\mathcal{H}$ be the class of indicators of all convex sets in $\mathbb{R}^2$. Show that

$$\mathrm{vc}(\mathcal{H}) = \infty.$$

Problem 4 (Two bounds on VC dimension)

(a) Prove that for any finite class of Boolean functions $\mathcal{H}$, we have

$$\mathrm{vc}(\mathcal{H}) \le \log_2 |\mathcal{H}| \,.$$

(b) Prove that for any finite-dimensional class of Boolean functions $\mathcal{H}$, we have

$$\mathrm{vc}(\mathcal{H}) \le \dim(\mathcal{H}),$$

where $\dim(\mathcal{H})$ denotes the linear algebraic dimension of $\mathcal{H}$, i.e. the maximal number of linearly independent functions in $\mathcal{H}$.

TURN OVER FOR HINTS

Hints

Hints for Problem 1.

(a) Include in the hypothesis class $\mathcal{H}$ all functions of the form $h(x) = \operatorname{sign}\langle w, x\rangle$, where w is.....(describe it yourself).

(b) Adopting the quadratic loss, write down the expression for $R(h)$. Notice that $R(h^*) = 0$. (Recall that by our assumptions, factors exactly determine the state of the bacteria.) Then apply the generalization bound from Lecture 35, November 23.

Hints for Problem 2.

(a) Open up the squares on each side, use Grothendieck's identity (Lecture 17, October 10) and a linearization of arccosine (Fact on p.3 of Lecture 17, October 10).

(b) Do this for the unit vectors $u = w_n^*/\sqrt{s}$ and $v = w^*/\sqrt{s}$.

Hints for Problem 3.

(b) Consider an arbitrarily large number of points $\{x_1, \dots, x_n\}$ that lie on a circle. Label these points with labels 1 and -1 arbitrarily. Find a convex set that includes all the points labeled 1, and excludes all points labeled -1. (A picture for $n = 5$ would be enough.)

Hints for Problem 4.

(a) Consider the "restricted class" $\mathcal{H}|_{\{x_1,\dots,x_d\}}$ obtained by restriction of each function $h \in \mathcal{H}$ onto the subset $S := \{x_1, \dots, x_d\}$. (Thus the functions in the restricted class have domain S.) If S is shattered by $\mathcal{H}$, then the restricted class consists of all 2^d Boolean functions on the d-element set S.

(b) The restricted class consists of all Boolean functions on a d-element set, so it has linear algebraic dimension d. (Check!)