CSCI 4230: Computational Learning Theory — Assignment 2
The Chinese University of Hong Kong, Spring 2023 — due Wednesday March 8 23:59

Collaborating on homework is encouraged, but you must write your own solutions in your own words and list your collaborators. Copying someone else's solution will be considered plagiarism and may result in failing the whole course.

Please answer clearly and concisely. Explain your answers. Unexplained answers will get lower scores or even no credits.

(1) (20 points) Let $d \geqslant 1$ and $\mathcal{C}$ be the set of axis-aligned hyperrectangles in $\mathbb{R}^d$, i.e.

$$\mathcal{C} = \{R(a,b) \mid a, b \in \mathbb{R}^d\}$$

where $R(a,b) = \{x \in \mathbb{R}^d \mid a_i \leqslant x_i \leqslant b_i \text{ for } 1 \leqslant i \leqslant d\}$ is the hyperrectangle with opposite corners a and b.

Show that $\text{VCDim}(\mathcal{C}) = 2d$. Explain in details how you get the lower and upper bounds.

Hint: Generalize Section 6.3.3 of the Shalev-Shwartz–Ben-David reference book.

(2) (30 points) We now show the VC dimension of linear threshold functions in n-dimensional Euclidean space is $n+1$, generalizing the result in Notes05.

In the following, given a finite set $S = \{s_1, \ldots, s_k\}$ of points in $\mathbb{R}^n$, the convex hull of S contains every convex combination of points in S, that is, the convex hull contains every point that can be written as $\sum_{1 \leqslant i \leqslant k} \lambda_i s_i$ such that $\lambda_i \geqslant 0$ and $\sum_{1 \leqslant i \leqslant k} \lambda_i = 1$.

Let $\mathcal{C}$ be the concept class of linear threshold functions in $\mathbb{R}^n$.

(a) Prove that $\text{VCDim}(\mathcal{C}) \geqslant n+1$. In other words, find a set of $n+1$ points in $\mathbb{R}^n$ that is shattered by $\mathcal{C}$.
Prove that your set is shattered.

(b) Show that $\text{VCDim}(\mathcal{C}) \leqslant n+1$ using Radon's Theorem, which says:
Radon's Theroem: Any set S of $n+2$ points in $\mathbb{R}^n$ can be partitioned into two disjoint subsets S_1 and S_2 whose convex hulls intersect.

(c) Prove Radon's Theorem.
Hint: Useful fact from linear algebra: Any $n+1$ points $x_1, \ldots, x_{n+1}$ in $\mathbb{R}^n$ are linearly dependent, that is, there are real numbers $\lambda_1, \ldots, \lambda_{n+1}$, *not all zeros*, such that $\lambda_1 x_1 + \cdots + \lambda_{n+1} x_{n+1} = 0$.
Hint: You may prove Radon's Theorem by first proving the following statement:
Any $n+2$ points $x_1, \ldots, x_{n+2}$ in $\mathbb{R}^n$ are affinely dependent, that is, there are real numbers $\lambda_1, \ldots, \lambda_{n+2}$, *not all zeros*, such that $\lambda_1 x_1 + \cdots + \lambda_{n+2} x_{n+2} = 0$ and $\lambda_1 + \cdots + \lambda_{n+2} = 0$.

(3) (20 points) Consider PAC learning s-sparse disjunctions. The end of Notes09 outlines the "Further Improved Algorithm" that finds a disjunction $h(x)$ consistent with all negative samples and $1 - \varepsilon/2$ fraction of positive samples, using Greedy Heuristic for Set Cover.

(a) Argue that $h(x)$ involves at most $s \ln(2/\varepsilon)$ literals.

(b) State and prove a variant of Theorem 2.1 of Notes09 and apply your theorem to show the following:
Let c be any s-sparse disjunction. Given $O\left(\frac{1}{\varepsilon}\left(\ln\frac{1}{\delta} + s \ln\frac{1}{\varepsilon} \ln n\right)\right)$ independent samples drawn from $\text{EX}(c, \mathcal{D})$, with probability $\geqslant 1 - \delta$, Further Improved Algorithm outputs a hypothesis $h(x)$ with error $\leqslant \varepsilon$.
(The key part is to justify why $O\left(\frac{1}{\varepsilon}\left(\ln\frac{1}{\delta} + s \ln\frac{1}{\varepsilon} \ln n\right)\right)$ samples suffice.)

(4) (20 points) In this problem, we want to show that properly PAC-learning monotone k-sparse disjunctions is NP-hard, by reducing from Set Cover.

To this end, give a polynomial-time reduction from Set Cover to the following problem:

Sparse Monotone Disjunction
Input: labelled samples $(x^1, y^1), \ldots, (x^m, y^m)$ where $x^i \in \{0,1\}^r$ and $y^i \in \{0,1\}$
Goal: Find sparsest monotone disjunction h consistent with all labels, i.e.
$h(x^i) = y^i$ for $1 \leqslant i \leqslant m$

Sparsest means h involves as few literals as possible.

In other words, your algorithm reads an input instance I of Set Cover and convert it to an input instance I' of Sparse Monotone Disjunction, such that

I has a small Set Cover if and only if I' has a sparse monotone disjunction

Justify why your reduction algorithm works.

(Techniques from Notes11 can further reduce from Sparse Monotone Disjunction to the corresponding proper learning problem, but you are not required to show this latter step.)

(5) (10 points) Show that for every $d \geqslant 0$ and every $m \geqslant 0$, there is a concept class $\mathcal{C}$ of VC dimension d over an instance space X and a subset $S \subseteq X$ such that $|S| = m$ and $|\Pi_{\mathcal{C}}(S)| = \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$.