Assignment 2 - TFML

Nirjhar Nath
nirjhar@cmi.ac.in
BMC202239

Solution 1:

We will start by proving the following claim:

Claim: For any $c \in \mathbb{R}$,

$$|c| = \min_{a \geq 0} a \;\text{ s.t. } c \leq a \text{ and } c \geq -a.$$

Proof: For any $c \in \mathbb{R}$, we have the following two cases:
Case 1: Suppose $c \geq 0$. Then, we have $|c| = c$. Consider the smallest non-negative a such that $c \leq a$ and $c \geq -a$. Since $c \geq 0$, it satisfies the following inequality:

$$-c \leq 0 \leq c \leq a.$$

The minimum value of a such that this inequality holds is c, i.e.,

$$c = |c| = \min_{a \geq 0} a \;\text{ s.t. } c \leq a \text{ and } c \geq -a.$$

Case 2: Suppose $c < 0$. Then, we have $|c| = -c$. Consider the smallest non-negative a such that $c \leq a$ and $c \geq -a$. Since $c < 0$, it satisfies the following inequality:

$$c < 0 < -c \leq a.$$

The minimum value of a such that this inequality holds is $-c$, i.e.,

$$-c = |c| = \min_{a \geq 0} a \;\text{ s.t. } c \leq a \text{ and } c \geq -a.$$

This proves our claim.

Now we can define a vector of auxiliary variables $\mathbf{s} = (s_1, \ldots, s_m)$, where

$$|\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i| = \min_{s_i \geq 0} s_i \;\text{ s.t. } \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \leq s_i \text{ and } \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \geq -s_i.$$

We want to minimize $\sum_{i=1}^{m}$ such that $\forall\ i \in [m]$,

$$\mathbf{w}^T \mathbf{x}_i - s_i \leq y_i \text{ and } -\mathbf{w}^T \mathbf{x}_i - s_i \leq -y_i.$$

We can translate the above into matrix form as follows: Let A be the $2m \times (m + d)$ matrix, composed as follows:

$$A = \begin{bmatrix} X & -I_m \\ -X & -I_m \end{bmatrix}$$

where X is the matrix whose rows are the feature vectors $\mathbf{x}_i$, and I_m is the $m \times m$ identity matrix. The vector of variables $\mathbf{v} \in \mathbb{R}^{m+d}$ combines the weight vector w and the auxiliary variables $\mathbf{s}$:

$$\mathbf{v} = (w_1, \ldots, w_d, s_1, \ldots, s_m)^T$$

The constraint vector $\mathbf{b} \in \mathbb{R}^{2m}$ is formed by concatenating the target values y_i and their negatives:

$$\mathbf{b} = (y_1, \ldots, y_m, -y_1, \ldots, -y_m)^T$$

Finally, the cost vector $\mathbf{c} \in \mathbb{R}^{d+m}$ that corresponds to the objective function is:

$$\mathbf{c} = (\mathbf{0}_d; \mathbf{1}_m)$$

where $\mathbf{0}_d$ is a zero vector of length d (the number of features) and $\mathbf{1}_m$ is a vector of ones of length m (the number of observations).

The resulting linear programming problem can thus be expressed as:

$$\min \mathbf{c}^T \mathbf{v} \quad \text{s.t.} \quad A\mathbf{v} \leq \mathbf{b}.$$

■

Solution 2:

Consider positive examples of the form $(\alpha, \beta, 1)$ where $\alpha^2+\beta^2+1 \leq R^2$. We note that the target vector $\mathbf{w}^* = (0, 0, 1)$ satisfies the condition $y(\mathbf{w}^* \cdot \mathbf{x}) \geq 1$ for all such pairs $(\mathbf{x}, y)$. Our goal is to construct a sequence of R^2 examples on which the Perceptron algorithm makes R^2 mistakes.

To build this sequence, we start with the example $(\alpha_1, 0, 1)$ where $\alpha_1 = \sqrt{R^2 - 1}$. For each round t, we select a new example so that the following two conditions are met:

$$\alpha^2 + \beta^2 + 1 = R^2 \quad \text{and} \quad \mathbf{w}_{(t)} \cdot (\alpha, \beta, 1) = 0.$$

If these two conditions hold, the Perceptron will continue to make mistakes. We will show that, as long as $t \leq R^2$, it is possible to satisfy both conditions.

Using induction, assume that $\mathbf{w}_{(t-1)} = (a, b, t-1)$ for some scalars a and b. Note that $\|\mathbf{w}_{(t-1)}\|^2 = t - 1$, which follows from the Perceptron's mistake bound proof, where we encounter inequalities that hold as equalities in this case. Thus, we have $\alpha^2+\beta^2+(t-1)^2 = (t-1)R^2$.

Without loss of generality, we can rotate $\mathbf{w}_{(t-1)}$ around the z-axis, transforming it to the form $(a, 0, t-1)$ with $a = \sqrt{(t-1)R^2 - (t-1)^2}$. We then choose

$$\alpha = \frac{t-1}{a}.$$

For any value of β,

$$(\langle a, 0, t-1\rangle, (\alpha, \beta, 1)) = 0,$$

which meets the second condition.

To ensure that the first condition holds, we need $\alpha^2 + 1 \leq R^2$. If this is true, we can select β such that $\beta^2 = R^2 - \alpha^2 - 1$. Indeed, we verify as follows:

$$\alpha^2 + 1 = \frac{(t-1)^2}{a^2} + 1 = \frac{(t-1)^2}{(t-1)R^2 - (t-1)^2} + 1 = \frac{(t-1)R^2}{(t-1)R^2 - (t-1)^2} \leq R^2,$$

where this last inequality holds when $R^2 \geq t$.

Thus, this construction allows us to generate a sequence of examples on which the Perceptron algorithm makes R^2 mistakes, reaching the theoretical upper bound. ■

Solution 3:

Given the sequence of hypothesis classes $H_1, H_2, \ldots$ for binary classification, where the learning algorithm implements the ERM rule, we need to prove that an ERM hypothesis can be found in the unrealizable case, where each class H_n is defined by at most $2n$ examples, and can be done so in $O(nm^{O(n)})$ time.

First, consider a sample S of size m. We can partition the hypotheses in H_n into equivalence classes where any two hypotheses within the same class have identical behavior with respect to S. The number of such equivalence classes is bounded by the binomial coefficient $\binom{m}{n}$, which indicates the number of ways to choose n elements from a set of m.

We have,

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}$$

Approximating this for large m and small n, we can simplify this as:

$$\binom{m}{n} \leq \frac{m^n}{n!}$$

Given that $n!$ grows super-exponentially, for practical computational purposes, we can approximate it further as:

$$\binom{m}{n} \leq m^n$$

Thus, $\binom{m}{n}$ falls within $O(m^n)$, which suits the polynomial bound of $O(m^{O(n)})$.

Since each hypothesis in an equivalence class behaves identically with respect to S, only one representative from each class needs to be evaluated for ERM. The empirical risk for each representative can be computed in $O(mn)$ time. Calculating this across potentially m^n equivalence classes yields a total computational time of:

$$O(mn \times m^n) = O(m^{n+1}n)$$

Considering n as a constant relative to m and the polynomial growth of m^n, the complexity is simplified to $O(nm^{O(n)})$.

Therefore, it is computationally feasible to find an ERM hypothesis in the unrealizable case for the class H_n, meeting the time complexity requirement of $O(nm^{O(n)})$, hence completing the proof. ■

Solution 4:

1. Given that A is a non-uniform learner for a class $\mathcal{H}$. For each $n \in \mathbb{N}$, we define $\mathcal{H}_n^A = \{h \in \mathcal{H} : m^{\text{NUL}}(0.1, 0.1, h) \leq n\}$. We have, for any distribution $\mathcal{D}$, with probability at least $0.9 = 1 - 0.1$ over the choice of $S \sim \mathcal{D}^m$, it holds that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + 0.1$$

for every $h \in \mathcal{H}_n$ m such that $m \geq m_{\mathcal{H}}^{\text{NUL}}(0.1, 0.1, h) \geq n$. Also, since $0.1 < \frac{1}{7}$ and $0.1 < \frac{1}{8}$, we have

$$L_{\mathcal{D}}(A(S)) \leq \text{argmin}_{h \in \mathcal{H}_n^A}(L_{\mathcal{D}}(h) + 0.1) < \text{argmin}_{h \in \mathcal{H}_n^A}\left(L_{\mathcal{D}}(h) + \frac{1}{8}\right).$$

If $\mathcal{D}$ satisfies the realizability assumption, then with probability $1 - \frac{1}{7}$, we get $L_{\mathcal{D}}(A(S)) < \frac{1}{8}$. Therefore, each class $\mathcal{H}_n$ has finite VC dimension, because if not, then by the No Free Lunch theorem, we have, $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$ with probability $\geq \frac{1}{7}$, which is a contradiction.

2. Given that $\mathcal{H}$ is nonuniformly learnable and $\mathcal{H}_n^A$ has finite VC dimension. By Theorem 7.2 and 7.3, it can be expressed as a countable union of agnostic PAC learnable classes $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n^A$, each having finite VC-dimension.

3. Assume, for the sake of contradiction, that $\text{VCdim}(\mathcal{H}_n) < \infty$ for every n. We next define a sequence of finite subsets $(K_n)_{n \in \mathbb{N}}$ of K in a recursive manner. Let $K_1 \subseteq K$ be a set of size $\text{VCdim}(\mathcal{H}_1) + 1$. Suppose that $K_1, \ldots, K_{r-1}$ are chosen. Since K is infinite, we can pick $K_r \subseteq K \setminus \bigcup_{i=1}^{r-1} K_i$ such that $|K_r| = \text{VCdim}(\mathcal{H}_r) + 1$.

   For each $n \in \mathbb{N}$, there exists a function $f_n : K_n \to \{0, 1\}$ such that $f_n \notin \mathcal{H}_n$. Since K is shattered, we can pick $h \in \mathcal{H}$ which agrees with each f_n on K_n. It follows that for every n, $h \notin \mathcal{H}_n$, contradicting our earlier assumption. Thus, there exists some n for which $\text{VCdim}(\mathcal{H}_n) = \infty$.

4. Let $\chi = \mathbb{R}$. For each $n \in \mathbb{N}$, consider $\mathcal{H}_n$ as the class composed of unions of up to n intervals. Specifically, define

$$\mathcal{H} = \{h_{a_1, b_1, \ldots, a_n, b_n} : \forall i \in [n], a_i \leq b_i\},$$

   where

$$h_{a_1, b_1, \ldots, a_n, b_n}(x) = \sum_{i=1}^{n} \mathbf{1}_{[a_i, b_i]}(x).$$

   The VC dimension of $\mathcal{H}_n$ can be shown to be $2n$: If we consider points $x_1 < \cdots < x_{2n}$, each i-th interval can be used to shatter the consecutive pairs $\{x_{2i-1}, x_{2i}\}$. For points $x_1 < \cdots < x_{2n+1}$, the alternating label pattern $(1, -1, \ldots, 1, -1)$ cannot be achieved with just n intervals, indicating a limitation. This analysis establishes that the overall VC-dimension of $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$ is infinite. As a result, $\mathcal{H}$ is not PAC learnable, although it is non-uniformly learnable.

5. Let $\mathcal{H}_2$ be the class of all functions from the interval $[0, 1]$ to $\{0, 1\}$. The set $[0, 1]$ is shattered by H_2, indicating that $\mathcal{H}_2$ is capable of classifying every subset of $[0, 1]$, which implies that $\text{VCdim}(\mathcal{H}_2) = \infty$. Thus, $\mathcal{H}_2$ is not nonuniformly learnable because, as shown in part (ii), if a class can be expressed as a countable

union of classes each having finite VC dimensions, then it is nonuniformly learnable. However, since $\mathcal{H}_2$ itself shatters an uncountable set with an infinite VC dimension, it cannot be described in such a way. Therefore, $\mathcal{H}_2$ is not nonuniformly learnable.

■