Machine Learning - Homework

Due: November 25th, 2019

Exercise 1 is on 4 points. Exercise 2 is on 3 points. Exercise 3 is on 7 points. Exercise 4 is on 9 points. The maximal mark is 20 points (hence, you do not need to do everything in order to have the maximal mark). Take great care of the redaction: it must be clear and precise.

1. **Hardness of learning.**
   In this exercise, we consider the problem of binary classification with the hypothesis class $\mathcal{H}$ of intersections of 3 homogeneous halfspaces in $\mathbb{R}^d$. Prove that computing an ERM in the realizable case for $\mathcal{H}$ is NP-hard.
   Hint: Recall that a graph $G = (V, E)$ is 3-colorable if there exists a mapping $f : V \to \{1, 2, 3\}$ such that $(u, v) \in E \implies f(u) \neq f(v)$. You may want to use the following reduction of the graph 3-coloring problem: for any graph $G = (V, E)$, where $V = \{v_1, \ldots, v_d\}$, let $m = |V| + |E|$ and $S \in (\mathbb{R}^d \times \{0, 1\})^m$ be the sample containing

   - for every $i \in \{1, \ldots, d\}$, the pair $(e_i, -1)$;
   - for every edge $(v_i, v_j) \in E$, the pair $\left(\dfrac{e_i + e_j}{2}, +1\right)$.

2. **On the VC-dimension.**
   1. Prove that the VC-dimension of a finite class $\mathcal{H}$ is at most $\Big\lfloor \log_2\left(|\mathcal{H}|\right) \Big\rfloor$, where $\lfloor u \rfloor$ denotes the largest integer at most equal to u.
   2. Give an example of an infinite class $\mathcal{H}$ of functions over the real interval $\mathcal{X} = [0, 1]$ such that $\mathrm{VCdim}(\mathcal{H}) = 1$.
   3. Give an example of a finite hypothesis class $\mathcal{H}$ over the domain $\mathcal{X}$ of your choice such that $\mathrm{VCdim}(\mathcal{H}) = \log_2\left(|\mathcal{H}|\right)$.

3. **Perceptron with margin.**
Consider binary classification in $\mathcal{X} = \mathbb{R}^d$ with label set $\mathcal{Y} = \{\pm 1\}$: the sample is $\big((x_1, y_1), \dots, (x_m, y_m)\big) \in \big(\mathbb{R}^d \times \{\pm 1\}\big)^m$. We assume that the data is linearly separable, and even that the *margin*

$$\gamma = \max_{w \in \mathbb{R}^d : \|w^*\| = 1} \ \min_{1 \le i \le m} \frac{y_i \langle w, x_i \rangle}{\|x_i\|}$$

is known and can be used in the algorithm. The aim of the *Perceptron with margin* algorithm is to find a linear separator with almost optimal margin. The aim of the questions 1-6 is to prove that the Perceptron-with-margin algorithm below achieves margin at least $\gamma/2$ in at most $12/\gamma^2$ iterations.

Algorithm: Perceptron-with-margin γ

Input: margin γ
Data: training set $(x_1, y_1), \dots, (x_m, y_m)$
1 $w_0 \leftarrow (0, \dots, 0)$
2 $t \geq 0$
3 **while** $\exists i_t : y_{i_t} \langle w_t, x_{i_t} \rangle \le \frac{\gamma}{2} \|x_{i_t}\| \|w_t\|$ **do**
4 $\quad w_{t+1} = w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|}$
5 $\quad t \leftarrow t + 1$
6 **return** w_t

1. Justify the existence of w^* such that
$$\forall 1 \le i \le m, \qquad \frac{y_i \langle w^*, x_i \rangle}{\|x_i\|} \ge \gamma \ .$$
2. In this question and the following, t is a positive integer for which the condition to continue the while loop of the algorithm (line 3) is satisfied. Prove that $\langle w^*, w_t \rangle \ge \gamma t$.
3. Prove that
$$\|w_{t+1}\|^2 \le \|w_t\|^2 + \gamma \|w_t\| + 1 \ .$$
4. Show that if $\|w_t\| \ge 2/\gamma$, then
$$\|w_{t+1}\|^2 \le \left(\|w_t\| + \frac{3\gamma}{4} \right)^2 \ .$$
5. Deduce that
$$\|w_t\| \le 1 + \frac{2}{\gamma} + \frac{3\gamma t}{4} \ .$$
6. Conclude.
7. For any $\eta \in (0, 1)$, give an algorithm that yields a linear separator with margin at least $(1 - \eta)\gamma$ in at most $K(\eta)/\gamma^2$ iterations, where $K(\eta)$ is a function to be specified.

4. **Adaboost.**

Let n be a positive integer, and let $\mathcal{X}$ be a subset of $\mathbb{R}^p$ for some $p > 0$. We assume that there exists a positive real number γ and a function Φ (called *weak classifier*) which, given any weighted sample $\mathcal{S} = \big\{(x_i, y_i, w_i) : 1 \le i \le m\big\}$, with $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$, $0 \le w_i \le 1$ and $w_1 + \cdots + w_m = 1$, yields a classification rule $h = \Psi(\mathcal{S}) : \mathcal{X} \mapsto \{-1, 1\}$ such that

$$\sum_{i=1}^{m} w_i \, \mathbb{1}\{h(x_i) \neq y_i\} \le \frac{1}{2} - \gamma \, .$$

Algorithm Adaboost works as follows. For a given number T of iterations:

- **Initialization:** for every $i \in \{1, \ldots, m\}$, let $w_i^1 = 1/m$;
- **Main loop:** for every t from 1 to T:
  - compute $h_t = \Phi\big((x_i, y_i, w_i^t)_{1 \le i \le m}\big)$;
  - compute
  $$\epsilon_t = \sum_{i=1}^{m} w_i^t \, \mathbb{1}\{h_t(x_i) \neq y_i\} \qquad \text{and} \qquad \alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) \; ;$$
  - for every $i \in \{1, \ldots, m\}$, let
  $$w_i^{t+1} = \frac{w_i^t}{Z_t} \times \exp\big(-\alpha_t \, y_i \, h_t(x_i)\big) \; ,$$
  where Z_t is such that $w_1^{t+1} + \cdots + w_m^{t+1} = 1$.
- **Output:** the final classifier is the function $H : \mathcal{X} \mapsto \{-1, 1\}$ defined by
$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t \, h_t(x)\right) \; ,$$
where $\operatorname{sign}(u) = 2 \times \mathbb{1}\{u \ge 0\} - 1$.

We define

$$F(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x)$$

and

$$e = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{H(x_i) \neq y_i\} \, .$$

1. In supervised classification, what is the name of e ?
2. Show that
$$e \le \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{y_i \, F(x_i) \le 0\} \le \frac{1}{m} \sum_{i=1}^{m} \exp\big(-y_i \, F(x_i)\big) \, .$$
3. Show that for every $i \in \{1, \ldots, m\}$,
$$w_i^{T+1} = \frac{\exp\big(-y_i \, F(x_i)\big)}{m \prod_{t=1}^{T} Z_t} \; ,$$
and that
$$\sum_{i=1}^{m} \exp\big(-y_i \, F(x_i)\big) = m \prod_{t=1}^{T} Z_t \, .$$

4. Show that
$$Z_t = \epsilon_t \exp(\alpha_t) + (1 - \epsilon_t)\exp(-\alpha_t) = 2\sqrt{\epsilon_t(1-\epsilon_t)}\ .$$
What is the value of α that minimizes
$$g(\alpha) = \epsilon_t \exp(\alpha) + (1 - \epsilon_t)\exp(-\alpha)\ ?$$
5. Show that
$$\sum_{i=1}^{m} w_i^{t+1}\, \mathbb{1}\{h_t(x_i) \neq y_i\} = \frac{1}{2}\ .$$
How to interpret this equality?
6. For every t between 1 and T, let $\gamma_t = 1/2 - \epsilon_t$. Show that
$$e \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2} \leq \exp\left(-2T\gamma^2\right)\ .$$
7. Give a value of T_0 such that for every $T \geq T_0$, $e = 0$. Should one necessarily choose T of order T_0 ?
8. How can you interpret the sentence: "weak learnability implies strong learnability"?
9. Why is **Ada**boost said to be *adaptive*?