Mathematics of Machine Learning
Homework 13 - Solutions

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

June 4, 2021

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

Problem 1

In the lecture notes, we proved that, in the realizable case, the Halving algorithm makes at most $O(\log_2 |\mathcal{F}|)$ mistakes where $\mathcal{F}$ is the class of classifiers. Modify Halving algorithm for the non-realizable case when there exits a classifier f^* that makes only m mistakes.

Solution 1

We start with an initial class $\hat{\mathcal{F}} = \mathcal{F}$. For $i = 1, \dots, |\mathcal{F}|$, the loop i goes as follows: We choose the majority vote according to the label of the sample received, after we remove the "wrong" classifiers, we continue until $\hat{\mathcal{F}}$ be empty. Then we restore the class $\hat{\mathcal{F}}$ and set $i = i + 1$ (go to the next loop), we stop restoring the classifier if it makes more than m mistakes. Clearly, in each loop the algorithm makes $O(\log_2 |\mathcal{F}|)$ mistakes, then the total number of mistakes is at most $(m + 1)O(\log_2 |\mathcal{F}|) = O(m \log_2 |\mathcal{F}|)$.

Problem 2

Suppose that we have a learning algorithm with regret R_T bounded by $\frac{1}{\eta} + \eta T$. If the time horizon was known, then we would choose $\eta = \sqrt{\frac{1}{T}}$ to minimize the regret bound as we did at the end of the proof of Theorem 21.2 in the lecture notes. The doubling trick is a technique that allows the learner to have a similar regret bound without knowing the time horizon. It goes as follows: Divide the time into intervals, i.e, for every $k \in \mathbb{N}$, the k-th interval is $2^k, \dots, 2^{k+1} - 1$. We run the algorithm in the beginning of each interval by setting $\eta_k := 2^{-k/2}$. Prove that the learning algorithm, implemented with the doubling trick, cannot have a regret R_2^T worse than $\frac{\sqrt{2}}{\sqrt{2}-1} R_T$.

Solution 2

Observe that the regret in the interval k is at most $\frac{1}{\eta_k} + \eta_k 2^k$. We have at most $\lfloor \log_2 T \rfloor + 1$ intervals. So we write

$$R_2^T \leq \sum_{k=0}^{\lfloor \log_2 T \rfloor} \frac{1}{\eta_k} + \eta_k 2^k = \sum_{k=0}^{\lfloor \log_2 T \rfloor} 2^{k/2+1} = \frac{2^{3/2}\sqrt{T} - 2}{\sqrt{2} - 1} \leq \frac{\sqrt{2}}{\sqrt{2} - 1} R_T.$$

In the last step we used that $R_T = 2\sqrt{T}$.

Problem 3

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function. Recall that the directional derivative of f in the direction of the vector v is defined as

$$\nabla_v f(x) := \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.$$

(a) Prove that the directional derivative of f in the direction of the vector v can be written in terms of the inner product between v and the gradient of f.

(b) Use letter "a" to justify the sentence: " The gradient of a function points out to the direction in which the function increases".

(c) Prove that if f is convex and if the gradient at the point x_o is zero, then x_0 is a global minimum, i.e, $f(x_0) \leq f(x)$ for all points x in the domain of f.

Solution 3

(a) By definition of the gradient,

$$0 = \lim_{t \to 0^+} \frac{f(x + tv) - f(x) - \nabla f^T(x)tv}{\|tv\|_2} = \nabla_v f(x) - \nabla f^T(x)v.$$

Therefore, $\nabla_v f(x) = \langle \nabla f(x), v \rangle$.

(b) By letter "a", for a unit vector v, $|\nabla_v f(x)| = |\langle \nabla f(x), v \rangle|$. The latter is at most $\|\nabla f(x)\|_2$ by Cauchy-Schwarz. We have an equality if and only if $v = \pm \frac{|\nabla f(x)|}{\|\nabla f(x)\|_2}$. Therefore the directional derivative is maximized/minimized in direction/opposite direction of the gradient of the function.

(c) First we claim that for a function $h : \mathbb{R} \longrightarrow \mathbb{R}$ to be convex is necessary and sufficient that $h(y) \geq h(x) + h'(x)(y - x)$ for all points $x, y \in \mathbb{R}$. Indeed, if h is convex, then for every $\lambda \in [0, 1]$, we have

$$h(\lambda x + (1 - \lambda)y) \leq \lambda h(x) + (1 - \lambda)h(x).$$

It immediately implies that

$$h(y) \geq h(x) + \lim_{\lambda \to 0^+} \frac{h(x + \lambda(y - x)) - h(x)}{\lambda} = h(x) + h'(x)(y - x).$$

The converse can be obtained if we set $z = \lambda x + (1 - \lambda)y$ and add the inequality $\lambda h(x) \geq \lambda(h(z) + h'(z)(x - z))$ to a second inequality, replacing x by y, $(1 - \lambda)h(y) \geq (1 - \lambda)(h(z) + h'(z)(y - z))$. Now, we apply the claim above to the function $g(\lambda) := f(\lambda y + (1 - \lambda)x)$, in fact g is convex because it is a composition of convex functions, from the claim $g(1) \geq g(0) + g'(0)$ gives that

$$f(y) \geq f(x) + \nabla f(x)^T(y - x).$$

It immediately implies that if $\nabla f(x_0) = 0$, then $f(x_0) \leq f(y)$ for all $y \in \mathbb{R}^n$.