Mathematical Tools in Machine Learning **(24.10.19)**

Solutions to Exercise Session 2

Instructor: Fadoua Balabdaoui *Assistant:* Loris Michel

Problem 9.1: L_1 loss as linear program

Define a vector of auxiliary variables $s = (s_1, \ldots, s_m)$. Following the hint, maximizing the empirical risk is equivalent to minimizing the linear objective $\sum_{i=1}^m s_i$ under the following constraints:

$$(\forall i \in [m]) \;\; \mathbf{w}^T \mathbf{x_i} - s_i \leq y_i, \;\; -\mathbf{w}^T \mathbf{x_i} - s_i \leq -y_i.$$

It is left to translate the above into a matrix form. Let $A \in \mathbb{R}^{2m\times(m+d)}$ be the matrix $A = [X - I_m; -X - I_m]$, where $X_{i\cdot} = x_i$ for every $i \in [m]$. Let $\mathbf{v} \in \mathbb{R}^{d+m}$ be the vector of variables $(w_1, \ldots, w_d, s_1, \ldots, s_m)$. Define $b \in \mathbb{R}^{2m}$ to be the vector $\mathbf{b} = (y_1, \ldots, x_m, -y_1, \ldots, -y_m)^T$. Finally, let $c \in \mathbb{R}^{d+m}$ be the vector $c = (0_d; \mathbb{1}_m)$. It follows that the optimization problem of minimizing the empirical risk can be expressed as the following LP:

$$\begin{aligned} min \quad & c^T \mathbf{v} \\ s.t. \quad & A\mathbf{v} \leq \mathbf{b}. \end{aligned}$$

Problem 9.3: Tightness of convergence time for the Perceptron

Following the hint, let $d = m$, and for every $i \in [m]$, let $\mathbf{x}_i = \mathbf{e}_i$. Let us agree that $\operatorname{sgn}(0) = -1$. For $i = 1, \ldots, d$, let $y_i = 1$ be the label of x_i . Denote by $\mathbf{w}^{(t)}$ the weight vector which is maintained by the Perceptron. A simple inductive argument shows that for every $i \in [d]$, $\mathbf{w}_i = \sum_{j<i} \mathbf{e}_j$. It follows that for every $i \in [d]$, $\langle \mathbf{w}^{(i)}, \mathbf{x}_i \rangle = 0$. Hence, all the instances $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are misclassified (and then we obtain the vector $w = (1, \ldots, 1)$ which is consistent with $x_1, \ldots, x_m$). We also note that the vector $\mathbf{w}^* = (1, \ldots, 1)$ satisfies the requirements listed in the question.

Problem 9.5: Modified Perceptron vs Perceptron

Since for every $w \in \mathbb{R}^d$ and every $x \in \mathbb{R}^d$, we have

$$\operatorname{sgn}(\langle w, x \rangle) = \operatorname{sgn}(\langle \eta w, x \rangle),$$

we obtain that the modified Perceptron and the Perceptron produce the same predictions. Consequently both algorithms produce the same number of iterations.

Problem 5.1: Sufficient condition

We simply follow the hint, by Lemma B.1 (in the Appendix of the book)

$$\begin{aligned} \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] &= \mathbb{P}_{\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1 - 7/8] \\ &\geq \frac{\mathbb{E}[L_{\mathcal{D}}(A(S))] - (1 - 7/8)}{7/8} \\ &\geq \frac{1/8}{7/8} \\ &= 1/7. \end{aligned}$$

Problem 5.3: Improvement of lower-bound (optional)

We modify the proof of the NFL theorem, based on the assumption that $|\mathcal{X}| \geq km$ (while referring to some equations from the book). Choose C to be a set with size km. Then, Equation (5.5) (in the book) becomes

$$L_{\mathcal{D}_i}(h) = \frac{1}{km} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \geq \frac{1}{km} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \geq \frac{k-1}{pk} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]}.$$

Consequently, the right-hand side of Equation (5.6) becomes

$$\frac{k-1}{k} \min_{r \in [p]} \frac{1}{T} \sum_{r=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]}.$$

It follows that

$$\mathbb{E}[L_{\mathcal{D}}(A(S))] \geq \frac{k-1}{2k}.$$