University of Freiburg
Dept. of Computer Science
Prof. Dr. F. Kuhn
P. Schneider

Advanced Algorithms
Sample Solution Problem Set 6

Issued: Friday, May 31, 2019

Exercise 1: Learning a Linear Classifier

Assume that we are given m feature vectors $\boldsymbol{a_1}, \dots, \boldsymbol{a_m} \in \mathbb{R}^n$ and that each vector $\boldsymbol{a_i}$ has a label $\ell_i \in \{-1, +1\}$. Our goal will be to find non-negative weights $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \geq \boldsymbol{0}$, such that the weighted combination of the features matches the label, i.e., such that such that $\text{sgn}(\boldsymbol{x}^\top \boldsymbol{a_i}) = \ell_i$ for all $i \in \{1, \dots, m\}$. Alternatively, we can define vectors $\boldsymbol{b_i} := \ell_i \boldsymbol{a_i}$ and we then require that $\boldsymbol{x}^\top \boldsymbol{b_i} \geq 0$ for all $i \in \{1, \dots, m\}$.

Concretely, we want to solve the following approximate version of the problem. Assume that there exists a non-negative vector $\boldsymbol{x^*}$ such that $\boldsymbol{b_i}^\top \boldsymbol{x^*} \geq 0$ for all i. W.l.o.g., we can assume that $\boldsymbol{x}^*$ is normalized such that $\boldsymbol{1}^\top \boldsymbol{x^*} = 1$, i.e., the entries of $\boldsymbol{x}^*$ sum up to 1. For a given parameter $\delta > 0$, our goal will be to find a vector $\boldsymbol{x}$, which is also normalized such that $\boldsymbol{1}^\top \boldsymbol{x} = 1$ such that $\boldsymbol{b_i}^\top \boldsymbol{x} \geq -\delta$ for all $i \in \{1, \dots, m\}$. In order to achieve this, we use the MWU algorithm as follows.

Assume that we have $\|\boldsymbol{b_i}\|_\infty \leq \rho$ for all $i \in \{1, \dots, m\}$ (i.e., all the absolute entries of the vectors $\boldsymbol{b_i}$ are upper bounded by ρ). We run the algorithm with n experts, one corresponding to each dimension. We interpret the vector $\boldsymbol{x}$ as a probability distribution on the n experts (dimensions) and initialize $\boldsymbol{x_1} := \frac{1}{n} \cdot \boldsymbol{1}$ to be the uniform distribution. In each step $t \geq 1$ of the MWU algorithm, we find a feature vector $\boldsymbol{b_i}$ for which $\boldsymbol{b_i}^\top \boldsymbol{x_t} < -\delta$ (if no such $\boldsymbol{b_i}$ exists, we are done and output the vector $\boldsymbol{x_t}$). We define the loss of expert $j \in \{1, \dots, n\}$ as $-b_{i,j}/\rho$ (where $b_{i,j}$ is the j^{th} entry of vector $\boldsymbol{b_i}$).

Show that after at most $O\big(\frac{\rho^2}{\delta^2} \log n\big)$ steps of the MWU algorithm, we have found a vector $\boldsymbol{x}$ for which $\boldsymbol{x}^\top \boldsymbol{b_i} \geq -\delta$ for all $i \in \{1, \dots, m\}$.

Sample Solution

Our strategy for the proof of correctness is as follows.

(i) As long as we find a vector $\boldsymbol{b_i}$ with $\boldsymbol{b_i}^\top \boldsymbol{x_t} < -\delta$ in some round t we show that we have an expected loss $L^t > \delta/\rho$ in that round (shown later).

(ii) We choose $T = C \cdot \frac{\rho^2}{\delta^2} \log n$ for some large enough constant C.

(iii) For a contradiction we assume that we find a vector $\boldsymbol{b_i}$ with $\boldsymbol{b_i}^\top \boldsymbol{x_t} < -\delta$ in *every* round $1, \dots, T$.

(iv) Then clearly (i), (ii) and (iii) imply that the total loss is $L = \sum_{t=1}^{T} L^t > T \cdot \frac{\delta}{\rho} = C \cdot \frac{\rho}{\delta} \log n$.

(v) From the lecture we know that the regret R, namely the difference of L to the loss L^* of the best expert for rounds $1, \dots, T$, is bounded by $R = L - L^* \leq c' \cdot \sqrt{T \log n} = c'\sqrt{C} \cdot \frac{\rho}{\delta} \log n \leq c \cdot \frac{\rho}{\delta} \log n$ with $c < C$ (if we define $c := c'\sqrt{C}$ and choose $C > c'^2$).

(vi) Due to $C > c$ the best expert must have a strictly positive loss: $L^* \geq L - R > (C - c) \cdot \frac{\rho}{\delta} \log n > 0$.

(vii) From the fact that a solution $\boldsymbol{x^*} \geq 0$ exists with $\boldsymbol{b_i}^\top \boldsymbol{x^*} \geq 0$ for each i we derive that there is an expert with loss at most 0 (shown later), a contradiction to (vi).

(viii) Since assumption (iii) is wrong, in some round in $1, \ldots, T$ we have no vector $\boldsymbol{b_i}$ with $\boldsymbol{b_i}^\top \boldsymbol{x_t} < -\delta$ anymore.

Let us plug the holes in (i) and (vii), starting with (i). For a given round $t \in [T]$, let us define $i(t)$ as the index of the chosen vector $\boldsymbol{b_{i(t)}}$ with $\boldsymbol{b_{i(t)}^\top x_t} < -\delta$.

$$L^t = \boldsymbol{x_t}^\top \boldsymbol{f^t} = \sum_{j=1}^{n} x_j^t f_j^t = -\frac{1}{\rho} \sum_{j=1}^{n} x_i^t b_{i(t),j} = -\frac{1}{\rho} \cdot \boldsymbol{b_{i(t)}^\top x_t} > \frac{\delta}{\rho}.$$

Finally we show (vii). We know that there is a solution $\boldsymbol{x^*} \geq 0$ with $\boldsymbol{b_i}^\top \boldsymbol{x^*} \geq 0$ for all $i \in [m]$.

$$\begin{aligned}
\forall i \in [m] : \boldsymbol{b_i}^\top \boldsymbol{x^*} \geq 0 \quad &\Longrightarrow \quad \sum_{t=1}^{T} \boldsymbol{b_{i(t)}^\top x^*} \geq 0 \\
&\Longleftrightarrow \quad \sum_{t=1}^{T} \sum_{j=1}^{n} x_j^* b_{i(t),j} \geq 0 \\
&\Longleftrightarrow \quad \sum_{j=1}^{n} \sum_{t=1}^{T} x_j^* b_{i(t),j} \geq 0 \\
&\Longrightarrow \quad \exists j \in [n] : \sum_{t=1}^{T} x_j^* b_{i(t),j} \geq 0 \\
&\Longrightarrow \quad \max_{j \in [n]} \sum_{t=1}^{T} x_j^* b_{i(t),j} \geq 0 \\
&\overset{\boldsymbol{x^*} \geq 1}{\Longleftrightarrow} \quad \max_{j \in [n]} \sum_{t=1}^{T} b_{i(t),j} \geq 0 \\
&\Longleftrightarrow \quad \min_{j \in [n]} \sum_{t=1}^{T} -\frac{b_{i(t),j}}{\rho} \leq 0 \\
&\Longleftrightarrow \quad \min_{j \in [n]} \sum_{t=1}^{T} f_j^t \leq 0 \quad \Longleftrightarrow \quad L^* \leq 0.
\end{aligned}$$