Mehryar Mohri
Advanced Machine Learning 2017
Courant Institute of Mathematical Sciences
Homework assignment 1
April 18th, 2017
Due: May 05, 2017

For all these problems, we will adopt the notation used in class.

A. Learning kernels

We consider the scenario of learning kernel.

1. Prove the equivalence of the (primal) optimization problem

$$\min_{\mathbf{w},\mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^{p} \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^{p} \mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i) \right) \right\}.$$

   and the (dual) problem

$$\max_{\boldsymbol{\alpha}} 2\boldsymbol{\alpha}^\top \mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_p \mathbf{Y} \boldsymbol{\alpha} \end{matrix} \right\|_r$$
$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0,$$

   where $r, q \geq 1$ are conjugate numbers, $\frac{1}{r} + \frac{1}{q} = 1$.

2. What does the problem correspond to for $r = 1$?

B. Deep boosting

Let F be the function defined over $\mathcal{F} = \text{conv}(\bigcup_{k=1}^{p} H_k)$ by

$$F(f) = \widehat{R}_{S,\rho}(f) + \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(H_{k_t}),$$

for any $f = \sum_{t=1}^{T} \alpha_t h_t \in \mathcal{F}$. Define the Voted Risk Minimization (VRM) solution as the function f^* minimizing F:

$$f_{\text{VRM}} = \operatorname*{argmin}_{f \in \mathcal{F}} F(f).$$

Let f^* be the element in $\mathcal{F}$ with the smallest generalization error:

$$R(f^*) = \inf_{f \in \mathcal{F}} R(f).$$

1. Fix $\rho > 0$. Use the margin bound presented in class for deep bosting to derive an upper bound on $R(f_{\text{VRM}}) - R_\rho(f^*)$, where $R_\rho(f^*) = \mathrm{E}_{(x,y)\sim D}[1_{yf^*(x)\leq\rho}]$ is the ρ-margin loss of f^*.

2. Compare this result to generalization bound proven in class for SRM.

C. Structured prediction

1. Show that $\Phi_u \colon v \mapsto e^{u-v}$ upper bounds $v \mapsto u1_{v\leq 0}$ for all $u \geq 0$.

2. Use that to derive a new structured prediction algorithm based on the hypothesis set

$$\mathcal{H}_2 = \{x \mapsto \mathbf{w} \cdot \mathbf{\Psi}(x, y) \colon \mathbf{w} \in \mathbb{R}^N, \|\mathbf{w}\|_2 \leq \Lambda_2\},$$

   for a feature vector $\mathbf{\Psi}$.

3. Assume a bigram feature decomposition $\Psi(x, y) = \sum_{k=1}^{l} \phi(x, k, y_{k-1}, y_k)$. Use that to give an explicit margin bound, assuming that $\|\mathbf{\Psi}\| \leq r$.

4. Describe in detail an efficient algorithm for the the computation of the gradient for your algorithm.

Bonus: Multi-Armed Bandit

Consider the standard multi-armed bandit problem with N arms, but assume now that the learner receives extra information as follows. Assume that there is a function $O : \{1, 2, \ldots, N\} \to 2^{\{1,2,\ldots,N\}}$ (where 2^A indicates the power set of A) such that at any round, pulling arm i results in knowledge of the rewards of all arms in $O(i)$. Assume further that $i \in O(i)$ (i.e. pulling an arm always results in knowledge of that arm's reward) and $i \in O(j) \Leftrightarrow j \in O(i)$.

1. Explain how the function O induces an undirected graph $G = (V, E)$ over the arms of the game.

2. Recall that a *clique* C of a graph is a subset of its vertices such that every vertex is connected to every other vertex in this subset. Moreover, a *clique covering* of a graph is a set of cliques such that their

union is equal to the entire set of vertices in the graph. Let $\mathcal{C}$ be a clique covering of the graph described in the previous question, so that $\cup_{C \in \mathcal{C}} C = V$. Design an algorithm that achieves the following regret bound:

$$\sum_{t=1}^{T} \mu_{I_t} - \mu_* \leq \mathcal{O}\left(\inf_{\mathcal{C}} \left\{ \sum_{C \in \mathcal{C}} \frac{\max_{i \in C} \Delta_i \log(T)}{\min_{j \in C} \Delta_j^2} \right\} \right).$$

3. Explain how this regret bound compares to results for the standard MAB problem and for the full-information setting.