Mehryar Mohri
Foundations of Machine Learning 2022
Courant Institute of Mathematical Sciences
Homework assignment 3
March 29, 2022
Due: April 12, 2022 11:55 PM EDT

A Boosting

1. Let $\mathcal{H}$ and $\mathcal{R}$ be two hypothesis sets of functions mapping $\mathcal{X}$ to the reals and let ℓ be a loss function define as

$$\ell(yh(x), r(x)) = \begin{cases} 1_{yh(x) \le 0}, & r(x) > 0, \\ c, & r(x) \le 0, \end{cases}$$

where c is a positive constant less than 1/2. For simplicity, define $b = 2\sqrt{\frac{1-c}{c}}$.

   (a) Let Ψ_1 and Ψ_2 be two loss functions define as

$$\Psi_1(yh(x), r(x)) = \max\{e^{r(x)-yh(x)}, ce^{-br(x)}\},$$

   and

$$\Psi_2(yh(x), r(x)) = e^{r(x)-yh(x)} + ce^{-br(x)}.$$

   Show that Ψ_1 is convex in $(yh(x), r(x))$ and it upper-bounds ℓ. Show that Ψ_2 is convex in $(yh(x), r(x))$ and it upper-bounds Ψ_1.

   (b) Suppose that $\mathcal{H} = \{h_1, h_2, \cdots, h_N\}$ and $\mathcal{R} = \{r_1, r_2, \cdots, r_N\}$ for some $N > 1$. We denote by $\mathcal{F}$ the convex hull of the set of base function pairs $\{(h_1, r_1), (h_2, r_2), \cdots, (h_N, r_N)\}$. Let $\Psi_{1,\mathcal{F}}$ be the family of functions defined by $\Psi_{1,\mathcal{F}} = \{(x, y) \mapsto \min\{\Psi_1(y\mathbf{h}(x), \mathbf{r}(x)), 1\}, (\mathbf{h}, \mathbf{r}) \in \mathcal{F}\}$. Show that the Rademacher complexity of $\Psi_{1,\mathcal{F}}$ admits the following upper bound:

$$\mathfrak{R}_m(\Psi_{1,\mathcal{F}}) \le \mathfrak{R}_m(\mathcal{H}) + (b+1)\mathfrak{R}_m(\mathcal{R}).$$

   (*Hint*: use Talagrand's lemma.)

   (c) Show that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $(\mathbf{h}, \mathbf{r}) \in \mathcal{F}$:

$$R(\mathbf{h}, \mathbf{r}) := \mathrm{E}_{(x,y)\sim\mathcal{D}}[\ell(y\mathbf{h}(x), \mathbf{r}(x))] \le \frac{1}{m}\sum_{i=1}^m \Psi_1(y_i\mathbf{h}(x_i), \mathbf{r}(x_i)) + 2\mathfrak{R}_m(\mathcal{H}) + 2(b+1)\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

   (d) Fix $\rho > 0$. Show that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $(\mathbf{h}, \mathbf{r}) \in \mathcal{F}$:

$$R(\mathbf{h}, \mathbf{r}) \le \frac{1}{m}\sum_{i=1}^m \Psi_1(y_i\mathbf{h}(x_i)/\rho, \mathbf{r}(x_i)/\rho) + \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) + \frac{2(b+1)}{\rho}\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

   Conclude that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $(\mathbf{h}, \mathbf{r}) \in \mathcal{F}$:

$$R(\mathbf{h}, \mathbf{r}) \le \frac{1}{m}\sum_{i=1}^m \Psi_2(y_i\mathbf{h}(x_i)/\rho, \mathbf{r}(x_i)/\rho) + \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) + \frac{2(b+1)}{\rho}\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

   (e) Define the objective $F(\boldsymbol{\alpha})$ for a boosting-type algorithm as

$$F(\boldsymbol{\alpha}) = \frac{1}{m}\sum_{i=1}^m e^{\mathbf{r}(x_i) - y_i\mathbf{h}(x_i)} + ce^{-b\mathbf{r}(x_i)} + \beta\sum_{j=1}^N \alpha_j,$$

   where $\mathbf{h} = \sum_{j=1}^N \alpha_j h_j$, $\mathbf{r} = \sum_{j=1}^N \alpha_j r_j$, and β is a non-negative constant. Show that it is a convex function of $\boldsymbol{\alpha}$. Briefly explain why part (d) suggests that we solve the optimization problem $\min_{\boldsymbol{\alpha}\ge 0} F(\boldsymbol{\alpha})$.

(f) Determine the best direction at iteration t if you apply coordinate descent to F. You should adopt a notation similar to the one used in class and define a distribution D_t for any $t \in [T]$ over the pairs (i, n) with $i \in [m]$ and $n \in \{1, 2\}$. Denote by Z_t the corresponding normalization factor. Distributions $D_{1,t}$ and $D_{2,t}$ are defined by $D_t(i, 1)/Z_{1,t}$ and $D_t(i, 2)/Z_{2,t}$ respectively, where $Z_{1,t}$ and $Z_{2,t}$ are the normalization factors. For any $t \in [T]$ and $j \in [N]$, define

$$\epsilon_{t,j} = \frac{1}{2}\big[1 - \mathrm{E}_{i \sim D_{1,t}}[y_i h_j(x_i)]\big], \; \bar{r}_{j,1} = E_{i \sim D_{1,t}}[r_j(x_i)], \; \bar{r}_{j,2} = E_{i \sim D_{2,t}}[r_j(x_i)].$$

Determine the best direction in terms of $Z_{1,t}$, $Z_{2,t}$, $\epsilon_{t,j}$, $\bar{r}_{j,1}$, $\bar{r}_{j,2}$, and c.

(g) Give the pseudocode of the algorithm. The best step η along a given direction that preserves the non-negativity of $\boldsymbol{\alpha}$ can be found by line search. You do not need to explicitly write down how to do line search in the pseudocode.