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

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

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 defined as

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

   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:

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

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

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

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

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

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

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

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

   (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 $\min_{\boldsymbol{\alpha}\geq 0} F(\boldsymbol{\alpha})$.

(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.

Solution:

(a) We have

$$\begin{aligned}
\ell(yh(x), r(x)) &= \max\Big\{1_{yh(x)\le 0} 1_{-r(x)<0}, c1_{r(x)\le 0}\Big\} \\
&\le \max\Big\{1_{\max\{yh(x),-r(x)\}\le 0}, c1_{r(x)\le 0}\Big\} \\
&\le \max\Big\{1_{\frac{yh(x)-r(x)}{2}\le 0}, c1_{r(x)\le 0}\Big\} \\
&= \max\Big\{1_{yh(x)-r(x)\le 0}, c1_{br(x)\le 0}\Big\} \\
&\le \Psi_1(yh(x), r(x)) \\
&\le \Psi_2(yh(x), r(x)).
\end{aligned}$$

The convexity is straightforward since $e^{r(x)-yh(x)}$ and $e^{-br(x)}$ are both convex in $(yh(x), r(x))$.

(b) Observe that

$$\begin{aligned}
\min\{\Psi_1(y\mathbf{h}(x), \mathbf{r}(x)), 1\} &= \min\Big\{\max\Big\{e^{\mathbf{r}(x)-y\mathbf{h}(x)}, ce^{-b\mathbf{r}(x)}\Big\}, 1\Big\} \\
&\le \min\Big\{e^{\mathbf{r}(x)-y\mathbf{h}(x)}, 1\Big\} + \min\Big\{ce^{-b\mathbf{r}(x)}, 1\Big\}.
\end{aligned}$$

Note that the function $u \mapsto \min\{e^u, 1\}$ is 1-Lipschitz and the function $u \mapsto \min\{ce^{bu}, 1\}$ is b-Lipschitz. Then, by Talagrand's lemma,

$$\mathfrak{R}_m(\Psi_{1,\mathcal{F}}) \le \mathfrak{R}_m((x, y) \mapsto \mathbf{r}(x) - y\mathbf{h}(x) : (\mathbf{h}, \mathbf{r}) \in \mathcal{F}) + b\mathfrak{R}_m((x, y) \mapsto -\mathbf{r}(x) : (\mathbf{h}, \mathbf{r}) \in \mathcal{F}).$$

The first term in the right-hand side satisfies

$$\begin{aligned}
\mathfrak{R}_m((x, y) \mapsto \mathbf{r}(x) - y\mathbf{h}(x) : (\mathbf{h}, \mathbf{r}) \in \mathcal{F}) &= \mathrm{E}_S\mathrm{E}_\sigma\left[\sup_{(\mathbf{h},\mathbf{r})\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i(\mathbf{r}(x_i) - y_i\mathbf{h}(x_i))\right] \\
&\le \mathrm{E}_S\mathrm{E}_\sigma\left[\sup_{(\mathbf{h},\mathbf{r})\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i\mathbf{r}(x_i)\right] + \mathrm{E}_S\mathrm{E}_\sigma\left[\sup_{(\mathbf{h},\mathbf{r})\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m -\sigma_i y_i\mathbf{h}(x_i)\right] \\
&= \mathrm{E}_S\mathrm{E}_\sigma\left[\sup_{(\mathbf{h},\mathbf{r})\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i\mathbf{r}(x_i)\right] + \mathrm{E}_S\mathrm{E}_\sigma\left[\sup_{(\mathbf{h},\mathbf{r})\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i\mathbf{h}(x_i)\right] \\
&= \mathfrak{R}_m(\mathrm{conv}(\mathcal{R})) + \mathfrak{R}_m(\mathrm{conv}(\mathcal{H})) \\
&= \mathfrak{R}_m(\mathcal{R}) + \mathfrak{R}_m(\mathcal{H}).
\end{aligned}$$

Similarly,

$$\mathfrak{R}_m((x, y) \mapsto -\mathbf{r}(x) : (\mathbf{h}, \mathbf{r}) \in \mathcal{F}) = \mathfrak{R}_m(\mathcal{R}).$$

Combining the above completes the proof.

(c) By Rademacher complexity bound and part (b), with probability at least $1-\delta$, the following holds for all $(\mathbf{h},\mathbf{r}) \in \mathcal{F}$:

$$\begin{aligned}
R(\mathbf{h},\mathbf{r}) &\leq \mathrm{E}_{(x,y)\sim\mathcal{D}}[\min\{\Psi_1(y\mathbf{h}(x),\mathbf{r}(x)),1\}] \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\min\{\Psi_1(y_i\mathbf{h}(x_i),\mathbf{r}(x_i)),1\} + 2\mathfrak{R}_m(\Psi_{1,\mathcal{F}}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\min\{\Psi_1(y_i\mathbf{h}(x_i),\mathbf{r}(x_i)),1\} + 2\mathfrak{R}_m(\mathcal{H}) + 2(b+1)\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\Psi_1(y_i\mathbf{h}(x_i),\mathbf{r}(x_i)) + 2\mathfrak{R}_m(\mathcal{H}) + 2(b+1)\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.
\end{aligned}$$

(d) Suppose that $\tilde{\mathcal{H}} = \{h_1/\rho, h_2/\rho, \cdots, h_N/\rho\}$ and $\tilde{\mathcal{R}} = \{r_1/\rho, r_2/\rho, \cdots, r_N/\rho\}$. We denote by $\tilde{\mathcal{F}}$ the convex hull of the set $\{(h_1/\rho, r_1/\rho), (h_2/\rho, r_2/\rho), \cdots, (h_N/\rho, r_N/\rho)\}$. Note that $(\mathbf{h},\mathbf{r}) \in \mathcal{F}$ is equivalent to $(\mathbf{h}/\rho, \mathbf{r}/\rho) \in \tilde{\mathcal{F}}$

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

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

The conclusion holds since Ψ_2 upper-bounds Ψ_1.

(e) Observe that

$$\begin{aligned}
F(\boldsymbol{\alpha}) &= \frac{1}{m}\sum_{i=1}^{m} e^{\mathbf{r}(x_i)-y_i\mathbf{h}(x_i)} + ce^{-b\mathbf{r}(x_i)} + \beta\sum_{j=1}^{N}\alpha_j \\
&= \frac{1}{m}\sum_{i=1}^{m}\Psi_2(y_i\mathbf{h}(x_i),\mathbf{r}(x_i)) + \beta\sum_{j=1}^{N}\alpha_j.
\end{aligned}$$

This is a convex function of $\boldsymbol{\alpha}$ since Ψ_2 is convex and composition with an affine function of $\boldsymbol{\alpha}$ preserves convexity.

Part (d) suggests to select $\boldsymbol{\alpha}$ as the solution of $\min_{\boldsymbol{\alpha}\in\Delta}\frac{1}{m}\sum_{i=1}^{m}\Psi_2(y_i\mathbf{h}(x_i)/\rho,\mathbf{r}(x_i)/\rho)$. Via a change of variable $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}/\rho$ that does not affect the optimization problem, we can equivalently search for $\min_{\boldsymbol{\alpha}\geq 0}\frac{1}{m}\sum_{i=1}^{m}\Psi_2(y_i\mathbf{h}(x_i),\mathbf{r}(x_i))$ such that $\sum_{j=1}^{N}\alpha_j \leq 1/\rho$. Introducing the Lagrange variable β associated to the constraint $\sum_{j=1}^{N}\alpha_j \leq 1/\rho$, the problem can be rewritten as $\min_{\boldsymbol{\alpha}\geq 0} F(\boldsymbol{\alpha})$.

(f) Define D_t by $D_t(i,1) = \frac{e^{\mathbf{r}_{t-1}(x_i)-y_i\mathbf{h}_{t-1}(x_i)}}{Z_t}$ and $D_t(i,2) = \frac{e^{-b\mathbf{r}_{t-1}(x_i)}}{Z_t}$. Then, we have

$$\begin{aligned}
&F'(\boldsymbol{\alpha}_{t-1},\mathbf{e}_j) \\
&= \frac{1}{m}\sum_{i=1}^{m}\big([r_j(x_i)-y_ih_j(x_i)]e^{\mathbf{r}_{t-1}(x_i)-y_i\mathbf{h}_{t-1}(x_i)} - cbr_j(x_i)e^{-b\mathbf{r}_{t-1}(x_i)}\big) + \beta \\
&= \frac{Z_t}{m}\sum_{i=1}^{m}([r_j(x_i)-y_ih_j(x_i)]D_t(i,1) - cbr_j(x_i)D_t(i,2)) + \beta \\
&= \frac{Z_t}{m}(2Z_{1,t}\epsilon_{t,j} - Z_{1,t} + Z_{1,t}\bar{r}_{j,1} - cbZ_{2,t}\bar{r}_{j,2}) + \beta.
\end{aligned}$$

Hence, the best direction is

$$k = \operatorname*{argmin}_{j\in[N]} 2Z_{1,t}\epsilon_{t,j} + Z_{1,t}\bar{r}_{j,1} - cbZ_{2,t}\bar{r}_{j,2}.$$

(g) This is straightforward based on the previous results.