Machine Learning HW#2

Evgeny Marshakov

Problem 1

1. We want to develop AdaBoost algorithm for the general loss function

$$L(\alpha) = \sum_{i=1}^{m} \Phi(-y_i g_t(x_i))$$

where $g = \sum_{t=1}^{T} \alpha_t h_t$. Let us consider the following distributions at each iteration

$$D_{t+1}(i) = \frac{\Phi'(-y_i g_t(x_i))}{\sum_{i=1}^{m} \Phi'(-y_i g_t(x_i))} =: \frac{\Phi'(-y_i g_t(x_i))}{Z_t}$$

Then at each iteration $t \geq 1$ the best base classifier h_t can be characterized as follows

$$e_t = \arg\min_k \left. \frac{dL(\alpha_{t-1} + \eta e_k)}{d\eta} \right|_{\eta=0} = -\sum_{i=1}^{m} y_i h_k(x_i) \Phi'(-y_i g_t(x_i)) \propto \tag{1}$$

$$-\sum_{i=1}^{m} y_i h_k(x_i) \frac{\Phi'(-y_i g_t(x_i))}{\sum_{i=1}^{m} \Phi'(-y_i g_t(x_i))} = -\sum_{i=1}^{m} y_i h_k(x_i) D_t(x_i) = (2\varepsilon_t - 1) \tag{2}$$

So the best classifier h_t is the one that minimizes the error ε_t on the training data. To find the weight of the last classifier (α_t or η, these are just different notations) we need to solve the following equation

$$\frac{dL(\alpha_{t-1} + \eta e_t)}{d\eta} = 0 \tag{3}$$

2. - The function $\Phi_1(-x) = 1_{x \leq 0}$ is obviously non differentiable (at the point $x = 0$) and non strictly increasing convex ($\Phi_1(x) = 0, \ \forall x < 0$).
   - The second derivative of the function $\Phi_2(x) = (1 + x)^2$ is equal to $\Phi_2''(x) = 2 > 0$ so this function is differentiable over all $\mathbb{R}$ and is strictly convex, but non increasing ($\Phi_2(-2) = \Phi_2(0)$).

- The function $\Phi_3(-x) = \max\{0, 1-x\}$ is obviously non differentiable (at the point $x = 1$).
- The function $\Phi_4(x) = \log_2(1 + e^x)$ has the following second derivative

$$\Phi_4''(x) = \frac{1}{\ln 2}\frac{e^x}{(1+e^x)^2} > 0,\ \forall x \in \mathbb{R}$$

So this function is strictly convex. The first derivative of this function is the following

$$\Phi_4'(x) = \frac{1}{\ln 2}\frac{e^x}{1+e^x} > 0,\ \forall x \in \mathbb{R}$$

so this function is strictly increasing. Obviously, for all $x \in \mathbb{R}$ $1 + e^x > 1 \Rightarrow \log_2(1 + e^x) > 0$. Since $\Phi_4(0) = 1$ and the first derivative is positive, for all $x \geq 0$ $\Phi_4(x) \geq 1$.

We see that only the last function satisfies all conditions.

3. Let us derive the boosting algorithm for the function $\Phi_4(-x) = \log_2(1 + e^{-x})$. As we use the distributions from th exercise 1.1, at iteration t we choose classifier that minimizes the error ε_t on the training data. Let us substitute our function to (2). After substituting we obtain an equation on α_t that cannot be solved analytically , so we use one step of Newton algorithm to approximate it.
Let us find the first derivative of the function $F(\alpha_{t-1} + \eta e_t)$ an the point $\eta = 0$.

$$F(\alpha_{t-1} + \eta e_t) = \sum_{i=1}^{m} \log(1 + e^{-y_i \sum s=1^{t-1} h_t(\alpha_s + \eta e_s)})$$

Substituting $\eta = 0$ we obtain that

$$d_1 = \frac{dF(\alpha_{t-1} + \eta e_t)}{d\eta}\bigg|_{\eta=0} = \sum_{i=1} \frac{-y_i h_t(x_i) e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s}}{1 + e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s}}.$$

Calculating the second derivative and substituting $\eta = 0$ we obtain

$$d_2 = \frac{d^2F(\alpha_{t-1} + \eta e_t)}{d\eta^2}\bigg|_{\eta=0} = \sum_{i=1}^{m} \left[\frac{(y_i h_t(x_i))^2 e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s}}{1 + e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s}} - \frac{(y_i h_t(x_i))^2 e^{-2y_i \sum_{s=1}^{t-1} \alpha_s h_s}}{(1 + e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s})^2} \right].$$

Consider $\alpha_t = -\frac{d_1}{d_2}$. The update of the distribution is the following:

$$D_{t+1}(i) \leftarrow \frac{\Phi_4'(-y_i g_t(x_i))}{\sum_{i=1}^{m} \Phi_4'(-y_i g_t(x_i))} = \frac{1}{Z_t} \cdot \frac{1}{1 + e^{y_i g_t(x_i)}}$$

where normalization factor is the following

$$Z_t = \sum_{i=1}^{m} \Phi_4'(-y_i g_t(x_i))$$

It differs from the standard AdaBoost in α_t. Moreover, we can find the better solution if we make more steps of Newton algorithm.

Problem 2

We know that linear combination (with positive coefficients) of differentiable and convex functions is differentiable and convex function, resp. Also we know that composition of convex with the linear function is convex. So the function

$$F(\alpha) = \sum_{i=1}^{m} w_i e^{-y_i \sum_{t=1}^{T} \alpha_y h_t}$$

is convex. Substituting this loss function to the general case from the exercise 1.1 we can find that

$$D_1(i) = \frac{w_i}{\sum_{i=1}^{m} w_i}$$

At each iteration we also choose the base classifier which minimizes the error ε_t on the training data. In this case the equation (3) has the following form

$$\frac{dF(\alpha_{t-1} + \eta_t e_t)}{d\eta} = 0 \Leftrightarrow -\sum_{i=1}^{m} y_i h_t(x_i) w_i e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s(x_i)} \cdot e^{-\alpha_t y_i h_t(x_i)} = 0 \Leftrightarrow$$

$$\Leftrightarrow -\sum_{i=1}^{m} D_t(i) e^{-\eta y_i h_t(x_i)} = 0 \Leftrightarrow (1-\varepsilon_t)e^{-\eta} - \varepsilon_t e^{\eta} = 0 \Leftrightarrow \eta = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$$

So we see that WeightedAdaBoost differs from the original AdaBoost in the first distribution

$$\text{AdaBoost}: \; D_1(i) = \frac{1}{m}$$

$$\text{WeightedAdaBoost}: \; D_1(i) = \frac{w_i}{\sum_{i=1}^{m} w_i}$$

Problem 3

We need to prove that

$$\sum_{i=1}^{m} D_{t+1}(i) y_i h_t(x_i) = 0$$

Indeed,

$$\sum_{i=1}^{m} D_{t+1}(i) y_i h_t(x_i) = \sum_{i=1}^{m} \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)} y_i h_t(x_i)}{Z_t} = \frac{1}{Z_t} \frac{dZ_t}{a\alpha_t}$$

since $Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)}$ by definition. From lectures we know that α_t minimizes Z_t, so $\frac{dZ_t}{d\alpha_t} = 0$. Hence these two vectors are uncorrelated.

Problem 4

1. • Set

$$G_1(x) = e^x$$

$$G_2(x) = x + 1$$

It can be easily seen that the both functions are differentiable and that $G_1'(0) = G_2'(0) = 1$, so the function $G(x)$ is also differentiable.

• Since $G_1(x)$ and $G_2(x)$ are convex functions,

$$G(y) - G(x) \geq G'(x)(y - x)$$

For $x \cdot y \geq 0$. Assume that $y \leq 0$ and $x \geq 0$. Then

$$G(y) - G(x) = e^y - (x + 1) \geq (y + 1) - (x + 1) = 1 \cdot (y - x) = G'(x)(y - x)$$

because $G'(x) = 1$ for all $x \geq 0$. The inequality holds since $y + 1$ is tangent line to the convex function e^y at the point $y = 0$. So the function $G(x)$ is convex.

2. We can make the same as in the first exercise.