Machine Learning Homework: Week 5 & 6

周书予
2000013060@stu.pku.edu.cn

April 12, 2022

1

Statement

Prove the following propositions about the algorithm of AdaBoost.

1. $\alpha_t = \arg\min_{\alpha} Z_t = \arg\min_{\alpha} \sum_{i=1}^{n} D_t(i)\exp(-\alpha y_i h_t(x_i))$.
2. $\prod_{t=1}^{T} Z_t = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i f(x_i)\right)$.
3. $\sum_{i=1}^{n} D_{t+1}(i)\mathbb{I}[y_i \neq h_t(x_i)] = \frac{1}{2}$.

Solution

1. Recall that $\varepsilon_t = \sum_{i=1}^{n} D_t(i)\mathbb{I}[y_i \neq h_t(x_i)]$, by applying **AM–GM inequality** we have

$$Z_t = \sum_{i=1}^{n} D_t(i)\exp(-\alpha y_i h_t(x_i)) = \varepsilon_t \exp(\alpha) + (1-\varepsilon)\exp(-\alpha) \geqslant 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

where the equality holds if and only if

$$\varepsilon_t \mathrm{e}^{\alpha} = (1-\varepsilon_t)\mathrm{e}^{-\alpha} \Leftrightarrow \alpha = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} = \frac{1}{2}\ln\frac{1+\gamma_t}{1-\gamma_t} = \alpha_t$$

where $\gamma_t = 1-2\varepsilon$ in the assignment of α_t. This suggests that $\alpha_t = \arg\min_{\alpha} Z_t$ as desired.

2. Notice that $D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$, by iteratively substitute the term of $D_t(i)$ in the expression of Z_T ($Z_T = \sum_{i=1}^{n} D_T(i)\exp(-\alpha_T y_i h_T(x_i))$), we can eventually obtain the following equality.

$$\begin{aligned}
\prod_{t=1}^{T} Z_t &= \prod_{t=1}^{T-1} Z_t \sum_{i=1}^{n} D_T(i)\exp(-\alpha_T y_i h_T(x_i)) \\
&= \prod_{t=1}^{T-2} Z_t \sum_{i=1}^{n} D_{T-1}(i)\exp(-\alpha_T y_i h_T(x_i) - \alpha_{T-1} y_i h_{T-1}(x_i)) \\
&= \cdots \\
&= \sum_{i=1}^{n} D_0(i)\exp\left(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i f(x_i)\right)
\end{aligned}$$

3.

$$
\begin{aligned}
\sum_{i=1}^{n} D_{t+1}(i)\mathbb{I}[y_i \neq h_t(x_i)] &= \sum_{i=1}^{n} \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}\mathbb{I}[y_i \neq h_t(x_i)] \\
&= \frac{\sum\limits_{i=1}^{n} D_t(i)\exp(\alpha_t)\mathbb{I}[y_i \neq h_t(x_i)]}{\sum\limits_{i=1}^{n} D_t(i)\exp(\alpha_t)\mathbb{I}[y_i \neq h_t(x_i)] + \sum\limits_{i=1}^{n} D_t(i)\exp(-\alpha_t)\mathbb{I}[y_i = h_t(x_i)]} \\
&= \frac{\varepsilon_t \mathrm{e}^{\alpha_t}}{\varepsilon_t \mathrm{e}^{\alpha_t} + (1-\varepsilon_t)\mathrm{e}^{-\alpha_t}}
\end{aligned}
$$

Since α_t is chosen so that $\varepsilon_t \mathrm{e}^{\alpha_t} = (1-\varepsilon_t)\mathrm{e}^{-\alpha_t}$, we can prove that $\sum\limits_{i=1}^{n} D_{t+1}(i)\mathbb{I}[y_i \neq h_t(x_i)] = \frac{1}{2}$.

2

Statement

Suppose (x, y) is a data point from the sample space, $\mathbf{h}(x) = (h_1(x), \cdots, h_T(x)) \in \mathbb{R}^T$ is a classifier. Denote $\alpha = (\alpha_1, \cdots, \alpha_T)^{\mathrm{T}}$ with $\sum_{t=1}^{T} \alpha_t = 1$ and $\alpha_i > 0$ as the normal vector of hyperplane $\mathcal{P} = \{\xi \in \mathbb{R}^T | \alpha^{\mathrm{T}}\xi = 0\}$, and $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ a linear combinator of classifiers. Show that $|f(x)|$ gives the **Chebyshev distance (L_∞ distance)** from $\mathbf{h}(x)$ to the hyperplane $\mathcal{P}$.

Solution

The problem is obviously equivalent to the following programming problem.

$$
\begin{aligned}
\text{minimize}_{\beta \in \mathbb{R}^T} \quad & \max_{t=1}^{T} |\beta_t| \\
\text{s.t.} \quad & \sum_{t=1}^{T} \alpha_t(\beta_t - h_t(x)) = 0
\end{aligned}
$$

- The solution is not greater than $\left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|$, since when $\beta_t = \sum_{i=1}^{T} \alpha_i h_i(x)$ for all t, the result is $\left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|$.
- The solution is not smaller than $\left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|$, otherwise $\sum_{t=1}^{T} \alpha_t \beta_t \leqslant \sum_{t=1}^{T} \alpha_t |\beta_t| < \sum_{t=1}^{T} \alpha_t \left|\sum_{t'=1}^{T} \alpha_{t'} h_{t'}(x)\right| = \left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|$ rises a contradiction with the programming condition.