Introduction to Statistical Machine Learning
Homework 3

Yota Toyama

November 17, 2016

1. A decision tree can classify linearly separable data. A boundary made by such a tree looks like stairs approximating $\mathbf{w}^T\mathbf{x} + w_0 = 0$. And, in the worst case, its depth is $\lceil \log \lceil \frac{N}{2} \rceil \rceil + 1$ because we can separate a space of $\mathbf{x}$ into $\lceil \frac{N}{2} \rceil$ thin regions and balance the tree along $\mathbf{x}_1$.

2. A decision tree can classify data points which are not linearly separable by separating a space of $\mathbf{x}$ into N thin regions along $\mathbf{x}_1$. And, in the worst case, its depth is $\lceil \log N \rceil$ when the tree is balanced in the same way as in the problem 1.

3.

$$\sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T+1)} = \sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} \frac{1}{Z} W_i^{(T)} e^{-\alpha_{T+1} y_i h_{T+1}(\mathbf{x}_i)}$$

$$= \frac{1}{Z} \sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T)} e^{\frac{1}{2} \log \frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}}$$

$$= \frac{1}{Z} \sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T)} \sqrt{\frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}}$$

$$= \frac{1}{Z} \sqrt{\frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}} \sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T)}$$

$$= \frac{\sqrt{\epsilon_{T+1}(1-\epsilon_{T+1})}}{Z}$$

$$Z = e^{-\alpha_{T+1}}(1 - \epsilon_{T+1}) + e^{\alpha_T} \epsilon_{T+1}$$

$$= \sqrt{\frac{\epsilon_{T+1}}{1-\epsilon_{T+1}}}(1 - \epsilon_{T+1}) + \sqrt{\frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}} \epsilon_{T+1}$$

$$= 2\sqrt{\epsilon_{T+1}(1-\epsilon_{T+1})}$$

$$\therefore \sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T+1)} = \frac{1}{2}$$

Assume $h_{T+2} = h_{T+1}$.

$$\sum_{i \text{ s.t. } y_i \neq h_{T+1}(\mathbf{x}_i)} W_i^{(T+1)} = \frac{1}{2}$$

$$\sum_{i \text{ s.t. } y_i \neq h_{T+2}(\mathbf{x}_i)} W_i^{(T+1)} = \frac{1}{2}$$

$$\epsilon_{T+2} = \frac{1}{2}$$

$$\epsilon_{T+2} \geq \frac{1}{2} \; \text{\textafk}$$

$$\therefore h_{T+2} \neq h_{T+1}$$

4.

$$\frac{\partial}{\partial \alpha_t} L(H_t, X) = 0$$

$$\frac{\partial}{\partial \alpha_t} \left(e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t}\epsilon_t \right) = 0$$

$$-e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t}\epsilon_t = 0$$

$$e^{2\alpha_t} = \frac{1 - \epsilon_t}{\epsilon_t}$$

$$\alpha_t = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t}$$

5.

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max \left\{ 0, 1 - y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + w_0) \right\}$$

$$\Leftrightarrow \quad \max_{\mathbf{w}, \xi} -\frac{1}{2} \|\mathbf{w}\|^2 - C \sum_{i=1}^{N} \xi_i$$

$$\begin{cases} y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + w_0) - 1 + \xi_i \geq 0 \\ \xi_i \geq 0 \end{cases}$$

Using Langrange multipliers,

$$\min_{\alpha_i,\mu_i} \max_{\mathbf{w},\xi} -\frac{1}{2}\|\mathbf{w}\|^2 - C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left(y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+w_0)-1+\xi_i\right) + \sum_{i=1}^{N}\mu_i\xi_i$$

$$\Leftrightarrow \quad \begin{cases} y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+w_0)-1+\xi_i \geq 0 \\ \xi_i \geq 0 \\ \alpha_i \geq 0 \\ \mu_i \geq 0 \\ \alpha_i(y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+w_0)-1+\xi_i)=0 \\ \mu_i\xi_i = 0 \end{cases}$$

Let $L = -\frac{1}{2}\|\mathbf{w}\|^2 - C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left(y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+w_0)-1+\xi_i\right) + \sum_{i=1}^{N}\mu_i\xi_i$.

$$\frac{\partial L}{\partial \mathbf{w}} = -\mathbf{w} + \sum_{i=1}^{N}\alpha_i y_i\phi(\mathbf{x}_i) = 0$$

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = -C + \alpha_i - \mu_i = 0$$

6. Please, see a Jupyter notebook file submitted together.