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.

4.

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_t} L(H_t, X) &= 0 \\
\frac{\partial}{\partial \alpha_t} \left(e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t}\epsilon_t \right) &= 0 \\
-e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t}\epsilon_t &= 0 \\
e^{2\alpha_t} &= \frac{1 - \epsilon_t}{\epsilon_t} \\
\alpha_t &= \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t}
\end{aligned}
$$

5.

$$
\begin{aligned}
L(Y|X;W,\mathbf{b}) &\approx L(y|\mathbf{x};W,\mathbf{b}) \\
&= -\log \hat{p}(y|\mathbf{x};W,\mathbf{b}) + \frac{\lambda}{2}\left\|W\right\|^2 \\
&= -\log \frac{e^{W_y\cdot\mathbf{x}+\mathbf{b}_y}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} + \frac{\lambda}{2}\left\|W\right\|^2 \\
\frac{\partial}{\partial W_{ci}} L(y|\mathbf{x};W,\mathbf{b}) &= -p(y=c)\mathbf{x}_i + \frac{\mathbf{x}_i e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} + \lambda W_{ci} \\
&= \left(\frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)\right)\mathbf{x}_i + \lambda W_{ci} \\
\frac{\partial}{\partial \mathbf{b}_c} L(y|\mathbf{x};W,\mathbf{b}) &= \frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)
\end{aligned}
$$

$\therefore$ The update equasions are the below.

$$
\begin{aligned}
W_{ci} &\leftarrow W_{ci} - \eta\left(\left(\frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)\right)\mathbf{x}_i + \lambda W_{ci}\right) \\
\mathbf{b}_c &\leftarrow \mathbf{b}_c - \eta\left(\frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)\right)
\end{aligned}
$$

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