Homework 3

1 Decision Tree

1.1 Problem 1

A decision tree can separate those points correctly. We can construct this decision tree as follows. For any point (x_{i1}, x_{i2}). First look at the first feature x_{i1}. Find the interval it falls into and then do a corresponding split according to the second feature x_{i2}. The figure 1 gives an intuitive explanation. The depth of the decision tree depth on the way of constructing. We can construct a tree with depth 2. The first part has N branches and it does a split according to the value on first dimension and in second part each node has 2 branches and it will split according to second dimension. If we use a binary tree to do classification the depth of the corresponding binary tree is N.

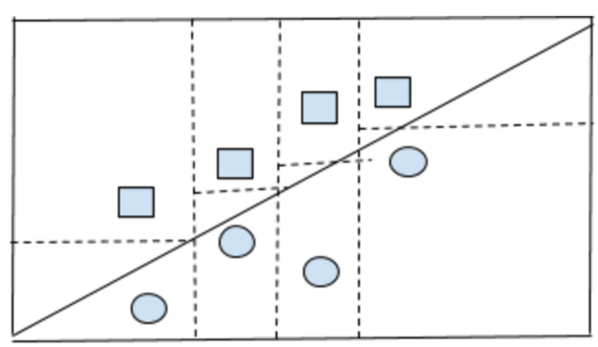

Figure 1: tree of a linear classifier

1.2 Problem 2

Those points can be classified correctly with a decision tree. A decision tree can do any division on the feature space. The strategy is as follows. First do a split on the first dimension of $\mathbf{x}$ to get k intervals and ensure in each interval there exists only one value. Suppose there exists n_i different values of second dimension in ith interval(i.e n_i points in the interval). Any labeling can be done by at most n_i splits. The depth of the tree still depends on the form of the tree. If it does not have to be a binary tree, the depth is 2 with ith part doing a split according to feature in ith dimension($i = 1, 2$). If it should be a binary tree the depth will be equal to p which is the number of distinct feature value in 1st dimension.

2 Boosting

2.1 Problem 3

Suppose the weighted error is $L(h_{T+1})$

$$\begin{aligned} L(h_{t+1}) &= \sum_{i=1}^{N} \frac{W_i^{T+1}}{Z_{T+1}} L_{0/1}(y_i, h_{T+1}(x_i)) \\ &= \sum_{i:y_i \neq h_{T+1}(x_i)} \frac{W_i^{T+1}}{Z_{T+1}} \\ &= \sum_{i:y_i \neq h_{T+1}(x_i)} \frac{W_i^T \cdot e^{\alpha_{T+1}}}{Z_{T+1}} \end{aligned} \tag{1}$$

As $\epsilon_{T+1} = \sum_{i:y_i \neq h_{T+1}(x_i)} W_i^T$, $\alpha_{T+1} = \frac{1}{2} log \frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}$, we can calculate Z_{T+1} as follows:

$$\begin{aligned} Z_{T+1} &= \sum_i W_i^{T+1} \\ &= \sum_i W_i^T \cdot e^{-\alpha_{T+1} y_i h_{T+1}(x_i)} \\ &= e^{-\alpha_{T+1}}(1 - \epsilon_{T+1}) + e^{\alpha_{T+1}} \epsilon_{T+1} \\ &= 2\sqrt{\epsilon_{T+1}(1 - \epsilon_{T+1})} \end{aligned} \tag{2}$$

Thus,

$$\begin{aligned} L &= \frac{\epsilon_{T+1} \cdot \sqrt{\frac{1-\epsilon_{T+1}}{\epsilon_{T+1}}}}{2\sqrt{\epsilon_{T+1}(1 - \epsilon_{T+1})}} \\ &= \frac{1}{2} \end{aligned} \tag{3}$$

2.2 Problem 4

Suppose exponential loss of $H_t (= \sum_i \alpha_i h_i)$ is L. Then,

$$
\begin{aligned}
L &= \sum_{i=1}^{N} W_i^t \\
&= \sum_{i=1}^{N} W_i^{t-1} \cdot e^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{i: y_i \neq h_t(x_i)} W_i^{t-1} \cdot e^{\alpha_t} + \sum_{i: y_i = h_t(x_i)} W_i^{t-1} \cdot e^{-\alpha_t} \\
&= e^{\alpha_t} \cdot \epsilon_t + e^{-\alpha_t} \cdot (1 - \epsilon_t)
\end{aligned} \tag{4}
$$

To minimize L, we need to set $\frac{dL}{d\alpha_t}$ to zero.

$$
\begin{aligned}
&\frac{dL}{d\alpha_t} = 0 \\
&\implies \frac{dL}{d\alpha_t} = e^{-\alpha_t}(\epsilon_t - 1) + e^{\alpha_t} \cdot \epsilon_t = 0 \\
&\implies \alpha_t = \frac{1}{2} log \frac{1 - \epsilon_t}{\epsilon_t}
\end{aligned} \tag{5}
$$

3 SVM

This is a solution to problem in non-separable case. Its equivalent dual form is as follows.

$$
\begin{aligned}
&max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \\
&s.t. \sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha \leq C \\
&w = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i
\end{aligned} \tag{6}
$$

We can set following matrices to represent it in standard form: $\mathbf{f} = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$, $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix}$, $\mathbf{B} = (y_1, \cdots, y_N)$, $\mathbf{H} = \begin{pmatrix} y_1 y_1 K(x_1, x_1) & \cdots & y_1 y_N K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ y_N y_1 K(x_N, x_1) & \cdots & y_N y_N K(x_N, x_N) \end{pmatrix} = diag(y_1, \cdots, y_N)\mathbf{K}diag(y_1, \cdots, y_N)$, $\mathbf{A} =$

$$\begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & & \\ & \ddots & \\ & & -1 \end{pmatrix}, \mathbf{a} = \begin{pmatrix} c \\ \vdots \\ c \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

With the above notations, we can represent the original problem in matrix form as follows.

$$\begin{aligned} & argmin \frac{1}{2}\alpha^T \mathbf{H}\alpha + \mathbf{f}^T \alpha \\ & s.t. \mathbf{A}\alpha \leq \mathbf{a} \\ & \mathbf{B}\alpha = \mathbf{0} \end{aligned} \tag{7}$$