**CS446: Machine Learning** **Fall 2014**

Problem Set 5

Chen Zhang *Handed In: 11/06/2014*

# 1 Problem 1

## 1.1 Part A

### 1.1.1 Unit 1

Define $w = [-1 \quad 1]$, $\theta = 0$. This is an easy solution to separate the given positive and negative examples.

### 1.1.2 Unit 2

Define $w = [-0.5 \quad 0.25]$, $\theta = 0$. This is the solution if we solve the optimization problem.

### 1.1.3 Unit 3

Since SVM optimization gives us a separator which could separate the positive and negative points, with a largest distance between the separator and the two group of points. We can use this property to calculate geometrically this separator and resulting $w$ and $\theta$. We know that the distance between a pair of parallel lines which go through two points is smaller or equal than the distance between those two points. The maximum distance is achieved when this pair of parallel lines are vertical to the line connecting those two points. Then we can draw two parallel lines which go through $x^{(1)}$ and $x^{(6)}$ and vertical to the connecting line of these two points. Thus the desired separator is the line sitting in the middle of this pair of parallel lines and also parallel to them. Now we have the separating line which will give us the largest margin, then we could calculate the corresponding $w$ and $\theta$. Then by scaling this $w$ and $\theta$, we could have the smallest possible $w$ and $\theta$.

## 1.2 Part B

### 1.2.1 unit 1

$I = [x^{(1)},\ x^{(6)}]$

### 1.2.2 unit 2

$\alpha_1$(corresponding to $x^{(1)}$)$= \frac{5}{32}$. $\alpha_6$(corresponding to $x^{(6)}$)$= \frac{-5}{32}$. All rest $\alpha$ corresponding to other points are zero.

### 1.2.3 unit 3

Since the $w$ and $\theta$ calculated are $w = [-0.5\ 0.25]$ and $\theta = 0$, we have $Objective\ function\ value = min\ \frac{1}{2}||w||^2 = ||[-0.5,\ 0.25]||^2 = \frac{5}{32}$

## 1.3 Part C

C is the weight of the loss function in the objective function of the optimization. It reflects how big a role we want the loss function to play in the optimization process, namely the calculation of the separator.
When C is infinitely large, we are considering the weight of loss function to be infinitely large. This way the optimization will be dominated by this loss function. Note that when C is infinitely large, the optimization is valid only when $\xi_i = 0$. This means that the resulting separator will be strict ($\xi = 0$) as well as that the margin need to be as big as possible ($min\ \frac{1}{2}||w||^2$). Thus when C is infinity, the solution to this optimization problem gives the hyperplane that achieves the largest margin (i.e. the hyperplane of (a)-2).
When $C = \infty$, as stated above, the support vectors are composed of the points which will give us a strict separating hyperplanes with the largest margin. Namely, these support vector are on the 'edge' of the group of points and has largest distance between the support vectors. When $C = 0$, the loss function is not taken into consideration at all. This means that the support vectors result in a hyperplane which has no guarantee to separate the data. In this case, the support vectors are away from the 'edge' of the group of points and the margin we get is not a 'valid' margin under our definition, since the resulting separator is not guaranteed to separate the points at all. When $C = 1$, it represent a case in between $C = \infty$ and $C = 0$. Both the loss function and the margin are considered in the optimization. The resulting separator will be a combined consideration of both having small loss and having big margin. The support vectors will not be on the 'edge' of group of points but may be close to the 'edge'. The resulting margin also is in between the case of $C = \infty$ and $C = 0$. But strictly speaking this margin is also not 'valid' under our definition.

# 2 Problem 2

## 2.1 Part A

The dual representation can be written as $w = \sum_{mistake\ examples} r\alpha_j y_j x_j$. In the equation $r$ is the learning rate, $\alpha_j$ is the number of mistakes made on data $j$, $y_j$ is the label of data $j$ and $x_j$ is the data itself. After we have this expression for $w$, we can plug in the following expression to determine the label $f(x) = Th_\theta(\sum_{i=1}^{n} w_i x_i(x))$.

## 2.2 Part B

$K_1(\vec{x}, \vec{z})$ is a valid kernel meaning that $K_1(\vec{x}, \vec{z}) = \vec{\phi_1}(\vec{x})\vec{\phi_1}(\vec{z})$. Similarly $K_2(\vec{x}, \vec{z}) = \vec{\phi_2}(\vec{x})\vec{\phi_2}(\vec{z})$. Now if we define $K(\vec{x}, \vec{z}) = \alpha K_1(\vec{x}, \vec{z}) + \beta K_2(\vec{x}, \vec{z})$, it is easy to see that $K(\vec{x}, \vec{z}) = \alpha\vec{\phi_1}(\vec{x})\vec{\phi_1}(\vec{z}) + \beta\vec{\phi_2}(\vec{x})\vec{\phi_2}(\vec{z})$. Again if we define $\vec{\xi} = [\sqrt{\alpha}\vec{\phi_1},\ \sqrt{\beta}\vec{\phi_2}]$(since $\alpha > 0\ \beta > 0$), it is easy to see that

| $i$ | Label | Hypothesis 1 | | | | Hypothesis 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $D_0$ | $x_1 \equiv$ $[x > 6\ ]$ | $x_2 \equiv$ $[y > 6\ ]$ | $h_1 \equiv$ $[x > 6\ ]$ | $D_1$ | $x_1 \equiv$ $[x > 8\ ]$ | $x_2 \equiv$ $[y > 8\ ]$ | $h_2 \equiv$ $[y > 8\ ]$ |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| *1* | $-$ | 0.1 | - | $+$ | - | 0.0624 | - | $+$ | $+$ |
| *2* | $-$ | 0.1 | - | - | - | 0.0624 | - | - | - |
| *3* | $+$ | 0.1 | $+$ | $+$ | $+$ | 0.0624 | - | - | - |
| *4* | $-$ | 0.1 | - | - | - | 0.0624 | - | - | - |
| *5* | $-$ | 0.1 | - | $+$ | - | 0.0624 | - | $+$ | $+$ |
| *6* | $+$ | 0.1 | $+$ | $+$ | $+$ | 0.0624 | - | - | - |
| *7* | $+$ | 0.1 | $+$ | $+$ | $+$ | 0.0624 | $+$ | $+$ | $+$ |
| *8* | $-$ | 0.1 | - | - | - | 0.0624 | - | - | - |
| *9* | $+$ | 0.1 | - | $+$ | - | 0.23 | - | $+$ | $+$ |
| *10* | $-$ | 0.1 | $+$ | $+$ | $+$ | 0.23 | - | - | - |

Table 1: Table for Boosting results

$K(\vec{x}, \vec{z}) = \vec{\xi}(\vec{x})\vec{\xi}(\vec{z})$. This means that $K(\vec{x}, \vec{z})$ is a inner product of two vectors, thus it is a valid kernel.

## 2.3 Part C

$K(\vec{x}, \vec{z}) = (\vec{x}^T\vec{z})^3 + 400(\vec{x}^T\vec{z})^2 + 100\vec{x}^T\vec{z} = [(\vec{x}^T\vec{z})^{\frac{3}{2}},\ 20(\vec{x}^T\vec{z}),\ 10\vec{x}] * [(\vec{x}^T\vec{z})^{\frac{3}{2}},\ 20(\vec{x}^T\vec{z}),\ 10\vec{z}]$. If we define $\vec{\phi}_1 = [(\vec{x}^T\vec{z})^{\frac{3}{2}},\ 20(\vec{x}^T\vec{z}),\ 10\vec{x}]$ and define $\vec{\phi}_2 = [(\vec{x}^T\vec{z})^{\frac{3}{2}},\ 20(\vec{x}^T\vec{z}),\ 10\vec{z}]$. Then $K(\vec{x}, \vec{z}) = \vec{\phi}_1 * \vec{\phi}_2$. Thus it is a inner product of two vectors, then it is a valid kernel.

# 3 Problem 3

## 3.1 Part A B C

For $D_0$, a uniform distribution is used, such that $D_0 = 0.1$ for all the data points. In this case, the best weak learner is $x > 6$, and there are two mistakes. Then $\epsilon_0 = 0.2$, $\alpha_0 = \frac{1}{2}\ln\frac{1-0.2}{0.2} = 0.693$
For $D_1$, for the right cases, $D_1 = \frac{D_0}{Z_0}\exp[-\alpha_0] = \frac{D_0}{Z_0}\exp[-0.693]$. For the wrong cases, $D_1 = \frac{D_0}{Z_0}\exp[\alpha_0] = \frac{D_0}{Z_0}\exp[0.693]$. It can be calculated that $Z_0 = 0.822$. With this the new distribution for right cases are $D_1 = \frac{0.1*\exp[-0.693]}{0.822} = 0.0642$, while for the wrong cases are $D_1 = \frac{0.1*\exp[0.693]}{0.822} = 0.23$.
Using this new distribution, we can calculate the error for the best weak learner in round 2 ($y > 8$) as $\epsilon_2 = 0.0624 * 2 = 0.2496$. Thus $\alpha_2 = \frac{1}{2}\ln\frac{1-\epsilon_2}{\epsilon_2} = 0.55$

## 3.2 Part D

Since $\epsilon_0 = 0.2$, thus $\alpha_0 = 0.693$. Similarly, $\epsilon_1 = 0.2496$, thus $\alpha_1 = 0.55$. Thus the final hypothesis is $sign[0.693 * sign[x - 6] + 0.55 * sign[y - 8]]$.

## 3.3 Part E

$Error(t + 1) = \sum_{i \in \; mistake \; at \; t+1} D_{t+1}(i) = \sum_{i \in \; mistake \; at \; t+1} \frac{D_t(i)}{Z_t} \exp[-\alpha_t y_i h_t(x_i)]$. If at t+1, we use the same distribution at t, then we have $\sum_{i \in \; mistake \; at \; t+1} \frac{D_t(i)}{Z_t} \exp[-\alpha_t y_i h_t(x_i)] = \sum_{i \in \; mistake \; at \; t} \frac{D_t(i)}{Z_t} \exp[\alpha_t]$

By using the expression of $\alpha_t$ we have $\sum_{i \in \; mistake \; at \; t} \frac{D_t(i)}{Z_t} \exp[\alpha_t] = \sum_{i \in \; mistake \; at \; t} \frac{D_t(i)}{Z_t} \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \sum_{i \in \; mistake \; at \; t} \frac{D_t(i)}{2\sqrt{\epsilon_t(1-\epsilon_t)}} \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \sum_{i \in \; mistake \; at \; t} \frac{D_t(i)}{2\epsilon_t} = \frac{1}{2\epsilon_t} \sum_{i \in \; mistake \; at \; t} D_t(i) = \frac{1}{2}$

# 4 Problem 4

## 4.1 Part A

### 4.1.1 Unit 1

Since we know that in town A, each family will have one child. Thus the expected number of children in a family is 1. In town B, each family will keep having new child until a boy is born. Thus in town B, the probability of having X children is $0.5^X$. Thus the expected value is $E[B] = 1 * 0.5^1 + 2 * 0.5^2 + 3 * 0.5^3 + ....n * 0.5^n$. When $n- > \infty$, $E[B] = 2$
In conclusion we have $E[A] = 1$ and $E[B] = 2$.

### 4.1.2 Unit 2

Since expected number of children in town A is 1 and the possibility of having boy or girl is the same, thus it is easy to see $R[A] = 1 : 1$. In town B, each family is expected to have two children and one of them must be a boy. Thus the other one must be a girl. Thus it is straightforward to see $R[B] = 1 : 1$