CS446: Machine Learning	**Fall 2015**
Problem Set 5	
Dominik Durner	*Handed In: November 10, 2015*

Question 1

1.a.1

$$w = (-1, 0)^T$$
$$\Theta = 0$$

1.a.2

Explanation in 1.a.3!

$$w = (-0.5, 0.25)^T$$
$$\Theta = 0$$

1.a.3

I just looked at the two closest points of each side. This is done since, I want to have at least a margin of 1 to the hyperplane. Hence, the margin is ≥ 1 for the nearest points. If all other points are now having a further distance to the hyperplane, they also satisfy the constrain ≥ 1. Therefore I found that x_1 and x_6 are the closest point for both classes. Now I want to separate those points by calculating a perpendicular line to those two points which separates the data. Accordingly, I calculated the middle point of those vectors which is $(0.4, 0.8)$. The slope of the original connecting line of the points $\frac{3.2}{-1.6} = -0.5$. Hence, the slope of the perpendicular line is 2. With the point $(0.4, 0.8)$ and the slope we get the following hyperplane equation.

$$y = mx + t \Rightarrow y = 2x + 0 = 2x$$

Hence, I get following values for w and Θ.

$$w = (-0.5, 0.25)^T$$
$$\Theta = 0$$

Now, we found the direction of the hyperplane, we need to satisfy the margin constraint that every points distance is at least 1 to the hyperplane. Given the previous fact we know the

following equations.

$$d_1 = \frac{y_1(wx_1)}{|w|} \Leftrightarrow |w| = \frac{y_1(wx_1)}{d_1}$$
$$d_6 = \frac{y_6(wx_6)}{|w|} \Leftrightarrow |w| = \frac{y_6(wx_6)}{d_6}$$

The distance of the points is $d_1 = \sqrt{(-1.2 - 0.4)^2 + (1.6 - 0.8)^2} = d_6$. Since we want to have the distance 1 for our closest points (supported vectors), we know that $y_1(wx_1) = 1$ and $y_6(wx_6) = 1$. Therefore, we simply can calculate the optimal $\hat{w} = c * w = (c * w_x, c * w_y)^T$.

$$\sqrt{(2c)^2 + c^2} = \sqrt{5c^2} = \frac{1}{d_1}$$
$$\Leftrightarrow c = \frac{1}{d_1\sqrt{5}} = 0.25$$

Hence we get $\hat{w} = (-0.5, 0.25)^T$ as the weight vector which satisfies the SVM constrains.

1.b.1

$$I = \{x_1, x_6\}$$

1.b.2

$$\hat{w} = (-0.5, 0.25)^T = \alpha_1 * (-1.2, 1.6)^T + \alpha_6 * (2, 0)^T$$
$$\Rightarrow \alpha_1 = 0.15625 \text{ (need to fully describe the y value)}$$
$$\Rightarrow (-0.3125, 0)^T = \alpha_6 * (2, 0)^T \text{ (since y doesn't change we can fix x)}$$
$$\Rightarrow \alpha_6 = -0.15625$$

$$\hat{w} = (-0.5, 0.25)^T = 0.15625 * (-1.2, 1.6)^T - 0.15625 * (2, 0)^T$$

1.b.3

$$0.5|w|^2 = 0.5 * ((-0.5)^2 + 0.25^2) = 0.15625$$

1.c

C is the regularization parameter that allows us to add a control on how soft we want to be in the optimization. The slack variables ξ_i allows some samples to have a margin of less than 1. Thus by adding $C\xi$ to the optimization function we are considering the cost of allowing example to have a margin smaller than 1. Hence, we can say C controls the relative importance of maximizing the margin.

With $C = \infty$ we really say that it is very important to have a margin of at least 1 to all points. This can easily be seen, since any error in one of the samples would generate an $+\infty$ outcome of our optimization function. Therefore, we would totally avoid having wrong values and would be getting a hard SVM! Hence, this would give us the same result as seen in question 1.a! Hence, the closest points with distance 1 would be supporting ones.

With $C = 1$: If $0 < \xi < 1$ the values would be still correct classified, but would still lie inside the margin. Those vectors are supported vectors since $y(wx + \Theta) = 1 - \xi$. If the $\xi > 1$ the values might be classified wrong. Also those vectors would be included as support vectors. Therefore, $C = 1$ gives us a relaxation for the SVM but still considers errors of the original idea of separating the training data. $C = 1$ might help us to generalize without loosing focus on the original problem.

With $C = 0$ we can have arbitrarily large values of ξ and they do not impact our optimization function. So the margin can be arbitrarily large since there is no punishment for having values inside the margin. Hence, every vector could be a possible support vector.

Question 2

2.a

1: Initialize $w = (0...0)^T$ and $\Theta = 0$
2: Initialize $M = \{\}$ (M needs to be a multiset!)
3: For each training example $(x, y) \in S$:
3.a: If $y \neq sign((\sum_{(z_i,y_i)\in M} \eta y_i z_i x) + \Theta)$:
3.a.1: $M = M \cup \{(x, y)\}$
3.a.2: $\Theta = \Theta + \eta y$
3.a: endIf
3: endFor
4: return $w = \sum_{(z_i,y_i)\in M} \eta y_i z_i$ and Θ

2.b

$$(xz)^3 = x_1^3 z_1^3 + 3x_1^2 z_1^2 x_2 z_2 + 3x_2^2 z_2^2 x_1 z_1 + x_2^3 z_2^3 =$$
$$= (x_1^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_2^2 x_1, x_2^3) \cdot (z_1^3, \sqrt{3}z_1^2 z_2, \sqrt{3}z_2^2 z_1, z_2^3) = \phi(x)\phi(z)$$

Hence, $(xz)^3$ is a valid kernel.

$$(xz)^2 = x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2 =$$
$$= (x_1^2, \sqrt{2}x_1 x_2, x_2^2) \cdot (z_1^2, \sqrt{2}z_2 z_1, z_2^2) = \phi(x)\phi(z)$$

Hence, $(xz)^2$ is a valid kernel.

We know from the lecture that $k'(x, z) = k_1(x, z) + k_2(x, z)$ is a valid kernel as long as k_1 and k_2 are valid. Furthermore, the scalar multiplication of a kernel $k''(x, z) = c \cdot k_s(x, z)$ is also known as a valid kernel (see lecture notes to Kernel). Since the trivial kernel $k'''(x, z) = x \cdot z$ is also known to be valid we know that

$$K(x, z) = (xz)^3 + 400(xz)^2 + 100(xz)$$

is a valid kernel.

i (1)	Label (2)	Hypothesis 1 D_0 (3)	$x_1 \equiv [x > 5]$ (4)	$x_2 \equiv [y > 6]$ (5)	$h_1 \equiv [x_1]$ (6)	Hypothesis 2 D_1 (7)	$x_1 \equiv [x > 8]$ (8)	$x_2 \equiv [y > 8]$ (9)	$h_2 \equiv [x_2]$ (10)
1	−	0.1	-	+	-	0.0625	-	+	+
2	−	0.1	-	-	-	0.0625	-	-	-
3	+	0.1	+	+	+	0.0625	-	-	-
4	−	0.1	-	-	-	0.0625	-	-	-
5	−	0.1	-	+	-	0.0625	-	+	+
6	+	0.1	+	+	+	0.0625	-	-	-
7	+	0.1	+	+	+	0.0625	+	+	+
8	−	0.1	-	-	-	0.0625	-	-	-
9	+	0.1	-	+	-	0.25	-	+	+
10	−	0.1	+	+	+	0.25	-	-	-

Table 1: Table for Boosting results

Question 3

3.a

See table above!

3.b

See table above!

3.c

The error for hypothesis 1 is $\epsilon = 0.2$

$$\alpha_0 = 0.5 * ln(\frac{1-\epsilon}{\epsilon}) = 0.5 * ln(\frac{0.8}{0.2}) = ln(2)$$
$$z_0 = 2\sqrt{\epsilon(1-\epsilon)} = 2\sqrt{0.2 * 0.8} = 0.8$$
$$D_1 = \frac{D_0(i)}{z_t} e^{-\alpha_0} = \frac{0.1 * 0.5}{0.8} = 0.0625 \text{ if } y_i = h_0(x_i)$$
$$D_1 = \frac{D_0(i)}{z_t} e^{\alpha_0} = \frac{0.1 * 2}{0.8} = 0.25 \text{ if } y_i \neq h_0(x_i)$$

3.d

The error for hypothesis 2 is $\epsilon = 0.25$

$$\alpha_1 = 0.5 * ln(\frac{1-\epsilon}{\epsilon}) = 0.5 * ln(\frac{0.75}{0.25}) = ln(\sqrt{3})$$
$$H_{final} = sign(ln(2) * sign([x > 5]) + ln(\sqrt{3}) * sign([y > 8]))$$

Question 4

4.a.i

Be X the random variable for the number of children.

$$E[X_A] = 1 * P(X_A = 1) = 1$$
$$E[X_B] = \sum_{k=1}^{\infty} k * P(X_B = k) =$$
$$= \sum_{k=1}^{\infty} k * P(female)^{k-1} P(male) =$$
$$= \sum_{k=1}^{\infty} k * 0.5^k = 2$$

4.a.ii

A: Since every child is $P(female) = P(male) = 0.5$ and every family just one child the expected value for boys and girls are both 0.5. Hence the ratio is 1:1.
B: Every family has 1 boy. So therefore we need to consider the expected number of girls (Y) for each family. This is the number of girls before the first boy. Hence, we consider the summation of expected value for having k girls.

$$E[Y] = \sum_{k=1}^{\infty} P(female)^k = \sum_{k=1}^{\infty} 0.5^k = 1$$

Therefore the ratio is in town B 1:1 too.

4.b.i

Just apply twice the product rule and we get the following equation.

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

4.b.ii

$$P(A,B,C) = P(A,B|C)P(C) = \frac{P(C|A,B)P(A,B)}{P(C)} P(C) =$$
$$= P(C|A,B)P(A,B) = P(C|A,B)P(A|B)P(B)$$

4.c

$$E[X] = 1 * P(A) + 0 * P(\neg A) = 1 * P(A) = P(A)$$

4.d.i

Independence means $P(A, B) = P(A)P(B)$.

$$P(X = 1, Y = 1) \stackrel{?}{=} P(X = 1)P(Y = 1) \Leftrightarrow \frac{17}{90} \stackrel{?}{=} \frac{7}{18} * \frac{7}{15}$$

Hence, X and Y are not independent since the result is not equal for the X=1 and Y=1 distribution!

4.d.ii

$$\begin{aligned}
P(Z = 1) &= \frac{2}{3} \\
P(Z = 0) &= \frac{1}{3} \\
P(X = 1, Y = 1|Z = 1) &= \frac{4}{30} \\
P(X = 1, Y = 1|Z = 0) &= \frac{3}{10} \\
P(X = 1, Y = 0|Z = 1) &= \frac{2}{10} \\
P(X = 1, Y = 0|Z = 0) &= \frac{1}{5} \\
P(X = 0, Y = 1|Z = 1) &= \frac{8}{30} \\
P(X = 0, Y = 1|Z = 0) &= \frac{6}{20} \\
P(X = 0, Y = 0|Z = 1) &= \frac{4}{10} \\
P(X = 0, Y = 0|Z = 0) &= \frac{2}{10}
\end{aligned}$$

$$P(X = 1|Z = 1) = \frac{1}{3}$$
$$P(X = 1|Z = 0) = \frac{1}{2}$$
$$P(X = 0|Z = 1) = \frac{2}{3}$$
$$P(X = 0|Z = 0) = \frac{1}{2}$$
$$P(Y = 1|Z = 1) = \frac{12}{30}$$
$$P(Y = 1|Z = 0) = \frac{6}{10}$$
$$P(Y = 0|Z = 1) = \frac{18}{30}$$
$$P(Y = 0|Z = 0) = \frac{6}{15}$$

Hence, the variables are conditionally independent since $P(X = x, Y = y|Z = z) = P(X = x|Z = z) * P(Y = y|Z = z)$ for all values of x, y, z.

4.d.iii

$$P(X = 0|X + Y > 0) \Rightarrow \frac{P(X = 0 \wedge X + Y > 0)}{P(X + Y > 0)} =$$
$$= \frac{P(X = 0 \wedge Y = 1)}{P(X = 0 \wedge Y = 1) + P(X = 1 \wedge Y = 0) + P(X = 1 \wedge Y = 1)} =$$
$$= \frac{\frac{5}{18}}{\frac{2}{3}} = \frac{5}{12}$$