CS446: Machine Learning **Fall 2015**

Problem Set 5

Shivam Upadhyay *Handed In: November 11, 2015*

1. SVM

(a) 1. $\mathbf{w} = [-1, 0]^T$
$\theta = 0$

2. $\mathbf{w} = [-0.5, 0.25]^T$
$\theta = 0$

3. SVM optimization is finding the maximal margin between the linearly separable data, geometrically if we look at the figure, we can see that points closest to the hyperplane separating the data are $[(-1.2, 1.6), +], [(2, 0), -]$. We can now find the slope of the separating hyperplane by, first finding the slope between the chosen points, $\frac{1.6}{-3.2} = -\frac{1}{2}$, and the midpoint, $(0.4, 0.8)$, so the line with the farthest distance between the two points (the support vectors), has a slope of 2 with a point $(0.4, 0.8)$, giving the line $y = 2x$, which gives $w = [-2, 1]^T, \theta = 0$.
Then, for minimising the value of w it can be halved repeatedly, until we get $y(w^T x + \theta) = 1$ for both support vectors, for which we will get the value $w = [-0.5, 0.25]$.So,this is the smallest value of w I can get.

(b) 1. $I = \{1, 6\}$

2. $\alpha = \left\{\frac{5}{32}, \frac{5}{32}\right\}$

3. Objective function value $= \frac{5}{32}$.

(c) The important parameter C is the penalty parameter as stated in the lecture notes it controls the trade-off between large margin and small hing loss, i.e it tells how much miss classification should SVM avoid.

Case 1: $C = \infty$, we will get the hyperplane that was obtained in part (a)-2.

Case 2: $C = 1$, The chances of miss classification are higher in this case, margins may contain the support vectors in this case, that's why we introduce the slack variableξ_i.

Case 3: $C = 0$, The margin is even higher in this case and hence even more chances of miss classification, but at the same time will give the larger margin separating hyperplane

2. Kernels

(a) Dual representation of perceptron algorithm

```
Initialize $\alpha \leftarrow \vec{0}$ of length $n$, where $n$ is the number of examples.
Initialize $\theta \leftarrow 0$.
Repeat until for predetermined number of rounds there is no mistake or no mistake at all
    For each training example $(x, y)$:
        if $y[(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x\rangle) + \theta] < 0$: ($\langle x_i, x\rangle$ represents the inner product)
            $\alpha_i \leftarrow \alpha_i + 1$ (where i is the index of the current example $(x, y)$)
            $\theta \leftarrow \theta + y$
```

(b) Given two examples $\vec{x} \in R^2$ and $\vec{z} \in R^2$, let $\vec{x} = (x_1, x_2)^T$ and $\vec{z} = (z_1, z_2)^T$ therefore,

$$K(\vec{x}, \vec{z}) = (x_1 z_1 + x_2 Z_2)^3 + 400(x_1 z_1 + x_2 Z_2)^2 + 100(x_1 z_1 + x_2 Z_2)$$

Now, we know that we a valid kernel by adding other valid kernels that is, if we can prove that kernels $K_1 = (x_1 z_1 + x_2 Z_2)^3$, $K_2 = 400(x_1 z_1 + x_2 Z_2)^2$, and $K_3 = 100(x_1 z_1 + x_2 Z_2)$ are valid, then kernel $K = K_1 + K_2 + K_3$ is also valid, let us first consider K_1,

$$K_1 = x_1^3 z_1^3 + 3x_1^2 x_2 z_1^2 z_2 + 3x_1 x_2^2 z_1 z_2^2 + x_2^3 z_2^3$$
$$K_1 = (x_1^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, x_2^3)(z_1^3, \sqrt{3}z_1^2 z_2, \sqrt{3}z_1 z_2^2, z_2^3)$$
$$K_1 = \phi_1(\vec{x})\phi_1(\vec{z})^T$$

Therefore K_1 is valid kernel.

$$K_2 = 400(x_1 z_1 + x_2 Z_2)^2 = 400(x_1^2, \sqrt{2}x_1 x_2, x_2^2)(z_1^2, \sqrt{2}z_1 z_2, z_2^2)$$

$$K_2 = \phi_2(\vec{x})\phi_2(\vec{z})^T$$

So, K_2 is a valid kernel, similarly:

$$K_3 = 100(x_1 z_1 + x_2 Z_2) = 100(x_1, x_2)(z_1, z_2)^T$$

$$K_3 = \phi_3(\vec{x})\phi_3(\vec{z})^T$$

Therefore, $K(\vec{x}, \vec{z}) = K_1 + K_2 + K_3$ is also a valid kernel.

3. Boosting:

(a)(b)(c) For $t = 0$, let us first compute the value of initial distribution $D_0(i)$, we know that,

$$D_{t+1} = \frac{D_t(i)}{z_t} \times 2^{-\alpha_0} \qquad if \qquad y_i = h_0(x_i)$$

$$D_{t+1} = \frac{D_t(i)}{z_t} \times 2^{\alpha_0} \qquad if \qquad y_i \neq h_0(x_i)$$

i	Label	Hypothesis 1				Hypothesis 2			
		D_0	$x_1 \equiv$ $[x > 5]$	$x_2 \equiv$ $[y > 6]$	$h_1 \equiv$ $[x_1]$	D_1	$x_1 \equiv$ $[x > 3]$	$x_2 \equiv$ $[y > 8]$	$h_2 \equiv$ $[x_2]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	$-$	$1/10$	$-$	$+$	$-$	$1/16$	$-$	$+$	$+$
2	$-$	$1/10$	$-$	$-$	$-$	$1/16$	$+$	$-$	$-$
3	$+$	$1/10$	$+$	$+$	$+$	$1/16$	$+$	$-$	$-$
4	$-$	$1/10$	$-$	$-$	$-$	$1/16$	$+$	$-$	$-$
5	$-$	$1/10$	$-$	$+$	$-$	$1/16$	$-$	$+$	$+$
6	$+$	$1/10$	$+$	$+$	$+$	$1/16$	$+$	$-$	$-$
7	$+$	$1/10$	$+$	$+$	$+$	$1/16$	$+$	$+$	$+$
8	$-$	$1/10$	$-$	$-$	$-$	$1/16$	$+$	$-$	$-$
9	$+$	$1/10$	$-$	$+$	$-$	$1/4$	$+$	$+$	$+$
10	$-$	$1/10$	$+$	$+$	$+$	$1/4$	$+$	$-$	$-$

Table 1: Table for Boosting results

initially, $D_0(i) = 1/m = 1/10$ and from the set of examples(positive) we can see see that $\theta_x > 5$ and $\theta_y > 6$, thus mistakes were made on the 9^{th} and 10^{th} examples, giving $\epsilon_0 = 0.2$ while choosing the minimum error. Thus hypothesis $h_1 = |x_1|$. Next, $\alpha_0 = \frac{1}{2}log_2(\frac{1-\epsilon_0}{\epsilon_0}) = \frac{1}{2}log_2(4) = 1$
Then, for first 8 examples $D_1 = D_0/z_t \times 2^{-1} = 1/(20 \times z_t)$ and for last two examples $D_1 = D_0/z_t \times 2^1 = 1/(5 \times z_t)$
Moreover, we know that normalisation factor $z_t = \sum\limits_i D_t(i) = 1$. Therefore, $\frac{8}{20 \times z_t} + \frac{2}{5 \times z_t} = 1$, making $z_t = \frac{4}{5}$.
Thus, for first 8 examples $D_1 = \frac{1}{16}$ and for last two $D_1 = \frac{1}{4}$ Similarly, now $\theta_x > 3$ and $\theta_y > 8$ gives minimum error for updated propbablity distribution. Thus, we will choose the hypothesis that minimises the error. Therefore, for $t = 1$ hypothesis will be $h_2 = |x_2|$. Rest all the entries are made in the table.

(d) There will be four mistakes on examples with index 1, 3, 5, and 6, giving $\epsilon_1 = \frac{4}{16} = \frac{1}{4}$.
So then $\alpha_1 = \frac{1}{2}log_2(\frac{1-0.25}{0.25}) \approx 0.79248$.
This makes the final $H(x) = sgn \sum\limits_t \alpha_t h_t(x) = sgn[1(x > 5) + 0.79248(y > 8)]$.

4. Probability

(a) i. The expected number of children in a family in town A and town b are:

- **Town A:** The expected number of children per family in town A is just 1. Because they just have single children irrespective of the gender
- **Town B:**In town B,let $P(B)$ denote the probability of a child being born is a boy,then $P(B)_1$ be the probability that a first children is born is a boy, which is 0.5.Probability that it's not a boy is $1 - P(B) = P(G) = 0.5$. Then $P(B)_2$ is the probability that they had a girl first, then a boy, which is $0.5 * 0.5$. In general for n^{th} child being born as a boy is $P(B)_n = 0.5^n$. So

$E[X] = \sum_{i=1}^{\infty} i * 0.5^i = 2$. So the expected number of children per family in town B is 2.

ii. Let B be the variable denoting number of boys, and Y denoting number of girls after a generation for Town A and Town B

Town A, doesn't have more than 1 children per family, therefore the expected number of boys per family in town A are just 0.5 (as $P(B) = P(G) = 0.5$), and the expected number of girls are 0.5 as well. Thus the gender of a child is equally probable for being a boy or a girl. $P(B = 1) = 0.5$, $P(B = 0) = 0.5$, and likewise $P(G = 1) = 0.5$, so the expected value for boys and girls are both $1 * 0.5 = 0.5$.

In town B, $P(B = 1) = 1$, since they refuse to stop having children until a son comes. But since they stop right after, $P(B > 1) = 0$. This means $E[B] = 1$. For girls, $P(G = 1) = 0.5 * 0.5$ since it means the first birth was a girl, and was immediately followed by a boy. Continuing, $P(G = 2) = 0.5^3$, and so on.
So $E[G] = 1 * 0.5^2 + 2 * 0.5^3 + \ldots = \sum_{i=1}^{\infty} i * 0.5^{i+1} = 1$.

Putting it all together, the ratio in town A is $\frac{0.5}{0.5} = \mathbf{1}$, and the ratio in town B is $\frac{1}{1} = \mathbf{1}$, maintaining the existing ratio in both towns.

(b) i. The probability of two events happening,give one of them has occurred is given by the chain rule and is as follow:

$$P(A \cap B) = P(A)P(B|A)$$

The above equation gives the probability of A and B happening, given A has already occurred.Similarly, probability of A and B when B has occurred is given by:

$$P(A \cap B) = P(B)P(A|B)$$

Equating the above two equations we get:

$$P(A)P(B|A) = P(B)P(A|B)$$

rearranging,

$$P(A|B) = P(A)P(B|A)/P(B)$$

ii. Using the chain rule:

$$P(\bigcap_{k=1}^{n} A_k) = \prod_{k=1}^{n} P(A_k | \bigcap_{j=1}^{k-1} A_j)$$

repeated application of this rule, we get:

$$P(A, B, C) = P(A|B, C)P(B, C)$$

$$P(A,B,C) = P(A|B,C)P(B|C)P(C)$$

(c) There are two events, E_1 and E_2 with probabilities $P(x=0) = P(\neg A)$ and $P(x=1) = P(A)$ with values $v_1 = 0, v_2 = 1$ respectively.
Therefore,Expected value is given by:

$$\mathbb{E} = v_1 P(x=0) + v_2 P(x=1)$$

$$\mathbb{E} = 0 \times P(\neg A) + 1 \times P(A)$$

$$\mathbb{E} = P(A)$$

(d) i. For two events X and Y to be independent, they should satisfy the following condition P(X—Y) = P(X) or P(Y—X) = P(Y)
Let us consider the case where value of X=o given Y=0, if P(X=0) is same as P(X=0—Y=0) X is independent of Y or else its dependent.

$$P(X=0) = 1/15 + 1/10 + 4/15 + 8/45 = 11/18$$

$$P(Y=0) = 1/15 + 1/15 + 4/15 + 2/15 = 8/15$$

$$P(X=0|Y=0) = \frac{P(X=0,Y=0)}{P(Y=0)} = \frac{1/15 + 4/15}{8/15} = 5/8$$

As $P(X=0|Y=0) \neq P(X=0)$ X is **NOT** independent of Y.

ii. If we can prove P(X=0—Y=y, Z=z) = P(X=0—Z=z) $\forall x, z \in \{0,1\}$, then X is conditionally independent given the value of Z.
First we have,

$$P(X=0|Y=0,Z=0) = \frac{1/15}{1/15 + 1/15} = 1/2$$

$$P(X=0|Y=1,Z=0) = \frac{1/10}{1/10 + 1/10} = 1/2$$

$$P(X=0|Y=0,Z=1) = \frac{4/15}{2/15 + 4/15} = 2/3$$

$$P(X=0|Y=1,Z=1) = \frac{8/45}{4/15 + 8/45} = 2/3$$

Second,

$$P(X=0|Z=0) = \frac{1/15 + 1/10}{1/15 + 1/15 + 1/10 + 1/10} = 1/2$$

$$P(X=0|Z=1) = \frac{4/15 + 8/45}{4/15 + 2/15 + 8/45 + 4/45} = 2/3$$

Thus, X is conditionally independent of Y given Z.

iii. For (X+Y ¿ 0), either one of them have to be 1 or both of them have to be 1, Thus,

$$P(X = 0|X + Y > 0) = \frac{1/10 + 8/45}{1/15 + 1/10 + 1/10 + 2/15 + 4/45 + 8/45} = 5/12$$