CS446: Machine Learning **Fall 2015**

Problem Set 5

Handed Out: October 30^{th}*, 2015* *Due: November* 07^{th}*, 2015*

- Feel free to talk to other members of the class in doing the homework. I am more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- Please, no handwritten solutions.
- The homework is due at 11:59 PM on the due date. Please submit your solution manuscript as a single pdf file via Compass (`http://compass2g.illinois.edu`).
- No code is needed for any of these problems. You can do the calculations however you please. You need to turn in only the report.

1. **[SVM - 50 points]**

We have a set of six labeled examples D in the two-dimensional space, $D = \{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(6)}, y^{(6)})\}$, $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{1, -1\}$, $i = 1, 2, ..., 6$ listed as follows:

i	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
1	-1.2	1.6	1
2	-1.6	2	1
3	4	1	-1
4	-3	0	1
5	3	-0.8	-1
6	2	0	-1

Figure 1: Training examples for SVM in question 1.(a)

(a) [20 points] We want to find a linear classifier where examples $\mathbf{x}$ are positive if and only if $\mathbf{w} \cdot \mathbf{x} + \theta \geq 0$.

1. [3 points] Find an easy solution $(\mathbf{w}, \theta)$ that can separate the positive and negative examples given.

Define $\mathbf{w} =$ ________________

Define $\theta =$ ________________

2. [10 points] Recall the Hard SVM formulation:

$$\mathbf{min}_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 \tag{1}$$

$$\text{s.t } y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \theta) \geq 1, \forall(\mathbf{x}^{(i)}, y^{(i)}) \in D \tag{2}$$

What would the solution be if you solve this optimization problem? (Note: you don't actually need to solve the optimization problem; we expect you to use a simple geometric argument to derive the same solution SVM optimization would result in).

Define $\mathbf{w} =$ ________________

Define $\theta =$ ________________

3. [7 points] Given your understanding of SVM optimization, how did you derive the SVM solution for the points in Figure 1?

(b) [17 points] Recall the dual representation of SVM. There exists coefficients $\alpha_i > 0$ such that:

$$\mathbf{w}^* = \sum_{i \in I} \alpha_i y^{(i)} \mathbf{x}^{(i)} \tag{3}$$

where I is the set of indices of the support vectors.

1. [5 points] Identify support vectors from the six examples given.

Define $I =$ ________________

2. [6 points] For the support vectors you have identified, find α_i such that the dual representation of $\mathbf{w}^*$ is equal to the primal one you found in (a)-2.

Define $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_{|I|}\} =$ ________________

3. [6 points] Compute the value of the hard SVM objective function for the optimal solution you found.

Objective function value = ________________

(c) [13 points] Recall the objective function for soft representation of SVM.

$$\mathbf{min}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{j=1}^{m}\xi_i \tag{4}$$

$$\text{s.t } y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + \theta) \geq 1 - \xi_i, \xi_i \geq 0, \forall(\mathbf{x}^{(i)}, y^{(i)}) \in D \tag{5}$$

where m is the number of examples. Here C is an important parameter. For which value of C, the solution to this optimization problem gives the hyperplane that achieves the largest margin (i.e., the hyperplane you have found in (a)-2? Comment on the impact on the margin and support vectors when we use $C = \infty$, $C = 1$, and $C = 0$. Interpret what C controls.

2. [**Kernels - 10 points**]

(a) [**5 points**] Write down the dual representation of the Perceptron algorithm.

(b) [**5 points**] Given two examples $\vec{\mathbf{x}} \in \mathbb{R}^2$ and $\vec{\mathbf{z}} \in \mathbb{R}^2$, let

$$K(\vec{\mathbf{x}}, \vec{\mathbf{z}}) = (\vec{\mathbf{x}}^T\vec{\mathbf{z}})^3 + 400(\vec{\mathbf{x}}^T\vec{\mathbf{z}})^2 + 100\vec{\mathbf{x}}^T\vec{\mathbf{z}}.$$

Prove that this is a valid kernel function.

3. [**Boosting - 30 points**] Consider the following examples $(x, y) \in \mathbb{R}^2$ (i is the example index):

i	x	y	Label
1	1	10	$-$
2	4	4	$-$
3	8	7	$+$
4	5	6	$-$
5	3	16	$-$
6	7	7	$+$
7	10	14	$+$
8	4	2	$-$
9	4	10	$+$
10	8	8	$-$

In this problem, you will use Boosting to learn a hidden Boolean function from this set of examples. We will use two rounds of AdaBoost to learn a hypothesis for this data set. In each round, AdaBoost chooses a weak learner that minimizes the error ϵ. As weak learners, use hypotheses of the form (a) $x_1 \equiv [x > \theta_x]$ or (b) $x_2 \equiv [y > \theta_y]$, for some integers θ_x, θ_y (either one of the two forms, not a disjunction of the two). There should be no need to try many values of θ_x, θ_y; appropriate values should be clear from the data.

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

Table 1: Table for Boosting results

(a) [**6 points**] Start the first round with a uniform distribution D_0. Place the value for D_0 for each example in the third column of Table 1. Write the new representation of the data in terms of the *rules of thumb*, x_1 and x_2, in the fourth and fifth columns of Table 1.

(b) [**6 points**] Find the hypothesis given by the weak learner that minimizes the error ϵ for that distribution. Place this hypothesis as the heading to the sixth column of Table 1, and give its prediction for each example in that column.

(c) [**6 points**] Now compute D_1 for each example, find the new best weak learners x_1 and x_2, and select hypothesis that minimizes error on this distribution, placing these values and predictions in the seventh to tenth columns of Table 1.

(d) [**6 points**] Write down the final hypothesis produced by AdaBoost.

(e) [**6 points**] Prove or disprove: AdaBoost determines the distribution at time $t+1$ in such a way that the error of the $t-$th hypothesis is exactly half.

What to submit: Fill out Table 1 as explained, show computation of α and $D_1(i)$, give the final hypothesis, H_{final}, and state your answer to part (e).

4. [**Probability - 20 points**]

(a) [**5 points**] There are two towns A and B, where all families follow the following scheme for family planning:

- Town A: Each family has just one child – either a boy or a girl.
- Town B: Each family has as many children as it wants, until a boy is born, and then it does not have any more children.

Assume that the boy to girl ratio is 1:1 for both towns A and B (number of boys equals number of girls), and the probability of having a boy child is 0.5, the same as that of having a girl child. Answer the following questions:

i. What is the expected number of children in a family in towns A and B?

ii. What is the boy to girl ratio at the end of one generation in towns A and B?

(b) [**5 points**]

i. For events A and B, prove

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

ii. For events A, B, and C, rewrite $P(A, B, C)$ as a product of several conditional probabilities and one unconditional probability involving a single event. Your conditional probabilities can use only one event on the left side of the conditioning bar. For example, $P(A|C)$ and $P(A)$ would be okay, but $P(A, B|C)$ is not.

(c) [**3 points**] Let A be any event, and let X be a random variable defined by

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

X is sometimes called the *indicator* random variable for the event A. Show that $\mathbb{E}[X] = P(A)$, where $\mathbb{E}[X]$ denotes the expected value of X.

(d) [**7 points**] Let X, Y, and Z be random variables taking values in $\{0, 1\}$. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z: For example, $P(X = 0, Y = 1, Z = 0) = 1/10$ and

	$Z = 0$		$Z = 1$	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$
$Y = 0$	1/15	1/15	4/15	2/15
$Y = 1$	1/10	1/10	8/45	4/45

$P(X = 1, Y = 1, Z = 1) = 4/45$.

i. Is X independent of Y ? Why or why not?

ii. Is X conditionally independent of Y given Z? Why or why not?

iii. Calculate $P(X = 0|X + Y > 0)$.