CS446: Machine Learning **Spring 2017**

Problem Set 4

Handed Out: February 27^{th}, 2017 *Due: March 11^{th}, 2017*

- 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. You will submit your solution manuscript as a single pdf file.
- Please present your algorithms in both pseudocode and English. That is, give a precise formulation of your algorithm as pseudocode and *also* explain in one or two concise paragraphs what your algorithm does. Be aware that pseudocode is much simpler and more abstract than real code.
- The homework is due at 11:59 PM on the due date. We will be using Compass for collecting the homework assignments. Please submit your solution manuscript as a pdf file via Compass (`http://compass2g.illinois.edu`). Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment.
- You may not use late submission credit hours for this problem set.

1. [**PAC Learning - 35 points**] In this problem, we are going to prove that the class of two concentric circles in the plane is PAC learnable. This hypothesis class is formally defined as $\mathcal{H}_{2cc} = \{h_{r_1,r_2} : r_1, r_2 \in \mathbb{R}_+ \text{ and } r_1 < r_2\}$, where

$$h_r(x) = \begin{cases} 1 & \text{if } r_1 \leq \|x\|_2 \leq r_2 \\ 0 & \text{otherwise} \end{cases}$$

For this problem, assume a sample of m points is drawn I.I.D. from some distribution $\mathcal{D}$ and that the labels are provided from some target function $h^*_{r_1^*,r_2^*} \in \mathcal{H}_{2cc}$.

(a) [**5 points**] Describe an algorithm that takes a training sample of m points as described above and returns a hypothesis $\hat{h}_{r_1,r_2} \in \mathcal{H}_{2cc}$ that makes zero mistakes on the training sample. To simplify the analysis that follows (in (b)), represent your hypothesis as two circles with radii r_1, r_2 such that: $r_1^* \leq r_1 < r_2 \leq r_2^*$.

Recall that PAC learning involves two primary parameters: ϵ and δ. ϵ is sometimes called the *accuracy parameter*; we say that if the true error of a learner is larger than ϵ, then the learning has "failed". Our hope is to directly prove that we can find some sample size m such that the probability of drawing a sample of size at least m from $\mathcal{D}$ which causes the learner to "fail" is less than δ (which is sometimes called the *confidence parameter*).

(b) Given the hypothesis that you learned in (a) your hypothesis will only make mistakes on positive examples (we ask that you justify that below). For this problem, ϵ is equal to the probability of drawing a point x from $\mathcal{D}$ that is labeled as a positive example and lies in the area between either $r_1^* \leq \|x\| < r_1$ or $r_2 < \|x\|_2 \leq r_2^*$ in other words,

$$\epsilon = \Pr_{x \sim \mathcal{D}}[r_1^* \leq \|x\|_2 < r_1 \text{ or } r_2 < \|x\|_2 \leq r_2^*]$$

i. [**5 points**] Explain why this is the case.

ii. [**5 points**] What is the probability of drawing a sample of m points from $\mathcal{D}$ where none of the points lie in the areas $r_1^* \leq \|x\|_2 < r_1$ or $r_2 < \|x\|_2 \leq r_2^*$?

(c) [**15 points**] Given parameters δ and ϵ, find value for m such that the probability of drawing a sample of size at least m that has true error larger than ϵ is less than δ.

- **Hint**: The following inequality might be useful:

$$1 - x \leq e^{-x}$$

(d) [**5 points**] We could have found a bound on m using another method. Derive this bound; how does it compare to the bound we found in the last step? (**Hint**: what is the VC Dimension of $\mathcal{H}_{2cc}$?).

2. [**VC Dimension - 5 points**] We define a set of concepts

$$H = \{sgn(ax^2 + bx + c); a, b, c, \in R\},$$

where $sgn(\cdot)$ is 1 when the argument $\cdot$ is positive, and 0 otherwise. What is the VC dimension of H? Prove your claim.

Grading note: You will not get any points without proper justification of your answer.

3. [**Kernels - 15 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

$$K(\vec{\mathbf{x}}, \vec{\mathbf{z}}) = (\vec{\mathbf{x}}^T \vec{\mathbf{z}})^3 + 49(\vec{\mathbf{x}}^T \vec{\mathbf{z}} + 4)^2 + 64\vec{\mathbf{x}}^T \vec{\mathbf{z}}.$$

Prove that this is a valid kernel function.

(c) [**5 points**] We wish to learn a Boolean function represented as a **monotone** DNF (DNF without negated variables) using kernel Perceptron. For this problem, assume that the size of each term in the DNF is of size k, s.t. $k \leq n$, the size dimensionality of the input. In order to complete this task, we will first define a kernel that maps an example $\mathbf{x} \in \{0, 1\}^n$ into a new space of monotone conjunctions of **exactly** k different variables from the n-dimensional space. Then, we will use the kernel Perceptron to perform our learning task.

Define a kernel $K(\mathbf{x}, \mathbf{z}) = \sum_{c \in C} c(\mathbf{x})c(\mathbf{z})$, where C is a family of monotone conjunctions containing **exactly** k different variables, and $c(\mathbf{x}), c(\mathbf{z}) \in \{0, 1\}$ is the value of c when evaluated on example $\mathbf{x}$ and $\mathbf{z}$ separately. Show that $K(\mathbf{x}, \mathbf{z})$ can be computed in time that is linear in n.

4. **[SVM - 25 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	-2	0	1
2	-2.4	-1.6	1
3	1.3	2.6	-1
4	-0.3	-2.5	1
5	3	0.2	-1
6	0	2	-1

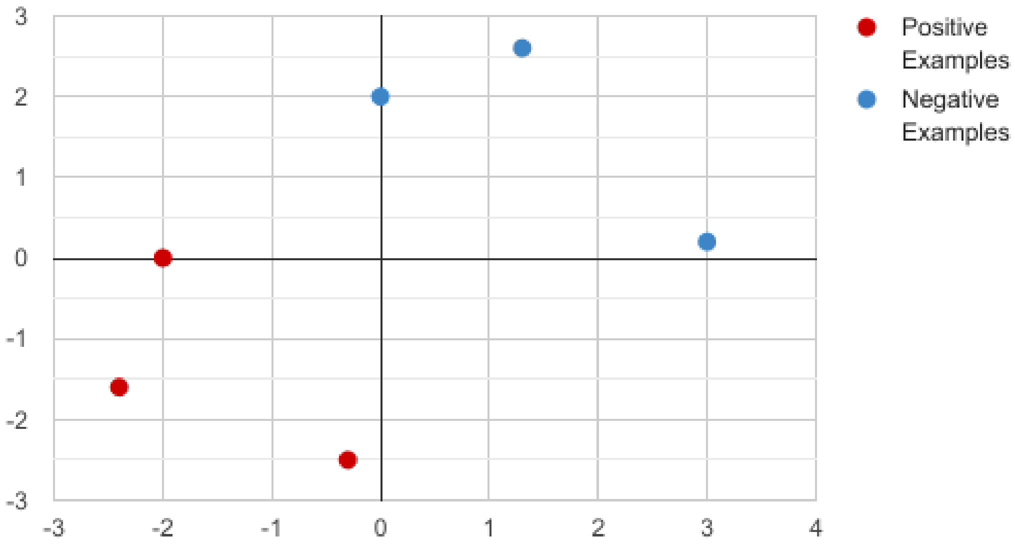

Figure 1: Training examples for SVM in question 1.(a)

(a) [4 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. [1 points] Find an easy solution $(\mathbf{w}, \theta)$ that can separate the positive and negative examples given.

   Define $\mathbf{w} =$ ________________

   Define $\theta =$ ________________

2. [4 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. [5 points] Given your understanding of SVM optimization, how did you derive the SVM solution for the points in Figure 1?

(b) [15 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. [5 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. [5 points] Compute the value of the hard SVM objective function for the optimal solution you found.

   Objective function value $=$ ________________

(c) [10 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 trivial value of C, the solution to this optimization problem gives the hyperplane that 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.

5. **[Boosting - 20 points]** Consider the following examples $(x, y) \in \mathbb{R}^2$ (i is the example index):

i	x	y	Label
1	0	8	$-$
2	1	4	$-$
3	3	7	$+$
4	-2	1	$-$
5	-1	13	$-$
6	9	11	$-$
7	12	7	$+$
8	-7	-1	$-$
9	-3	12	$+$
10	5	9	$+$

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) $f_1 \equiv [x > \theta_x]$ or (b) $f_2 \equiv [y > \theta_y]$, 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.

(a) **[5 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*, f_1 and f_2, in the fourth and fifth columns of Table 1.

(b) **[5 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.

i	Label	Hypothesis 1				Hypothesis 2			
		D_0	$f_1 \equiv$ $[x > \_\ ]$	$f_2 \equiv$ $[y > \_\ ]$	$h_1 \equiv$ $[\ \_\_\_\_\ ]$	D_1	$f_1 \equiv$ $[x > \_\ ]$	$f_2 \equiv$ $[y > \_\ ]$	$h_2 \equiv$ $[\ \_\_\_\_\ ]$
(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

(c) [**5 points**] Now compute D_1 for each example, find the new best weak learners f_1 and f_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) [**5 points**] Write down the final hypothesis produced by AdaBoost.

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