CIS519: Applied Machine Learning **Fall 2020**

Homework 3

Handed Out: November 2, 2020 *Due: November 16, 2020 at 11:59pm*

- 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 include at the top of your document the list of people you consulted with in the course of working on the homework.
- While we encourage discussion within and outside the class, cheating and copying code is strictly not allowed. Copied code will result in the entire assignment being discarded at the very least.
- Please use Piazza if you have questions about the homework. Also, please come to the TAs recitations and to the office hours.
- Handwritten solutions are not allowed. All solutions must be typeset in Latex. Consult the class' website if you need guidance on using Latex. If you don't have a lot of experience with Latex (or even if you do), we recommend using Overleaf (`https://www.overleaf.com`) to write your solutions. You will submit your solutions as a single pdf file (in addition to the package with your code; see instructions in the body of the assignment).
- The homework is due at 11:59 PM on the due date. We will be using Gradescope for collecting the homework assignments. You should have been automatically added to Gradescope. If not, please ask a TA for assistance. Please do **not** hand in a hard copy of your write-up. Post on Piazza and contact the TAs if you are having technical difficulties in submitting the assignment.
- Here are some resources you will need for this assignment
  - Latex materials: `https://www.seas.upenn.edu/~cis519/fall2019/assets/HW/HW3/hw3-material.zip`

1 Short Questions [20 Points]

(a) **(4 points)** Consider the following SVM formulation with the squared hinge-loss

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i [\max(0, 1 - y_i \mathbf{w}^\intercal \mathbf{x}_i)]^2$$

(1) What is i in the expression above. (Choose **one** of the following)

(a) index of features of examples

(b) index of examples in training data

(c) index of examples which are support vectors

(d) index of examples in test data

(2) One of the following optimization problems is equivalent to the one stated above. Which one? (Choose **one** of the following)

(a)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq -\xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(b)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i^2 \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 0 \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(c)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 1 - \xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(d)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i^2 \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 1 - \xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(b) **(9 points)** Consider the instance space $X = \{1, 2, \ldots N\}$ of natural numbers. An *arithmetic progression* over X is a subset of X of the form $X \cap \{a + bi : i = 0, 1, 2, \ldots\}$ where a and b are natural numbers. For instance, for $N = 40$ the set $\{9; 16; 23; 30; 37\}$ is an arithmetic progression over $\{1, 2, \ldots N\}$ with $a = 9, b = 7$.

Let C be the concept class consisting of *all arithmetic progressions over* $\{1, 2, \ldots N\}$. That is, each function $c_{a,b} \in C$ takes as input a natural number and says 'yes' if this natural number is in the $\{a, b\}$ arithmetic progression, 'no' otherwise.

A neural network proponent argues that it is necessary to use a deep neural networks to learn functions in C. You decide to study how expressive the class C is in order to determine if this argument is correct.

(1) **(2 points)** Determine the order of magnitude of the VC dimension of C (Circle one of the options below).

(A) $O(\log N)$ (B) $O(N)$ (C) $O(N^2)$ (D) ∞

(2) **(5 points)** Give a brief justification for the answer above.

(3) **(2 points)** Use your answer to (a) to settle the argument with the deep network proponent. (Circle on of the options below)

(A) Deep networks are needed (B) Simpler Classifiers are sufficient

(c) **(7 points)** We showed in class that the training error of the hypothesis h generated by the AdaBoost algorithm is bounded by

$$e^{-2\gamma^2 T},$$

where T is the number of rounds and γ is the advantage the weak learner has over chance.

Assume that you have a weak learner with an advantage of at least $1/4$. (That is, the error of the weak learner is below $1/4$.) You run AdaBoost on a dataset of $m = e^{14} (\approx 10^6)$ examples, with the following tweak in the algorithm:

> Instead of having the algorithm loop from $t = 1, 2, \ldots, T$, you run the loop with the following condition: If the Adaboost classifier h misclassifies at least one example, do another iteration of the loop.

That is, you run AdaBoost until your hypothesis is consistent with your m examples.

Question: Will your algorithm run forever, or can you guarantee that it will halt after some number of iterations of the loop?

(1) Choose one of the following options (**2 points**):

(A) Run forever (B) Halt after some number of iterations

(2) Justify (**5 points**): If you think it may run forever, explain why. If you think it will halt after some number of iterations of the loop, give the best *numeric* bound you can on the maximum # of iterations of the loop that may be executed. (You don't need a calculator here).

2 Boosting [30 points]

In this problem, you will manually run the AdaBoost algorithm for two iterations on the following dataset:

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

i	x	y	Label
1	9	4	+
2	2	1	+
3	5	2	+
4	6	2	+
5	8	4	+
6	1	2	−
7	2	9	−
8	3	4	−
9	6	7	−
10	5	8	−

Each example $(x, y) \in \mathcal{R}^2$ has a positive or negative label. The instances are numbered under column i.

You 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). Each θ can be an integer between 0 and 9 inclusive (for this dataset, lower than 0 and higher than 9 will not change the hypothesis since all of the features values are between 0 and 9). To make the computation easier, use log base 2 and wherever Adaboost has e^x, use 2^x instead.

Fill in Table 1 by following these instructions:

1. **[1 point]** 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.
2. **[10 points]** Find the values for θ_x and θ_y which have the lowest errors. Put the values you find in the top boxes in columns 4 and 5 (e.g. write $x > 9$ if $\theta_x = 9$ had the lowest

error for any x value) and their corresponding errors ϵ. If there are ties, you can pick a value arbitrarily. Then, fill in the rest of columns 4 and 5 with $+/-$ based on how f_1 and f_2 would label the instances.

3. **[1 point]** In column 6, pick the function which had the lowest error (e.g. write $h_1 \equiv f_1$ if f_1 has the lowest error), then copy that hypothesis' predictions in column 6.
4. **[6 points]** Now compute D_1 for each example and write the new values in column 7.
5. **[10 points]** Repeat the above process for hypothesis 2 using the new D_1 values.
6. **[2 points]** Write down the final hypothesis produced by AdaBoost. You can write the answer in terms of h_1 and h_2, but include the actual values for α_1 and α_2.

What to submit: Fill out Table 1 as explained and give the final hypothesis, H_{final}.

3 SVMs [30 points]

In this problem you will manually find the margin of a hard SVM and answer questions about the solution you found.

You have been provided with a set of 6 labeled examples D in two-dimensional space in Figure 1, $D = \{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(6)}, y^{(6)})\}$, $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{1, -1\}, i = 1, 2, ..., 6$.

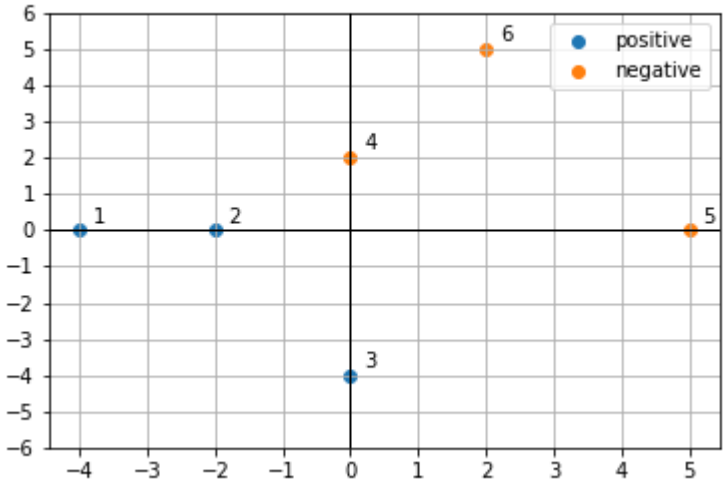

Figure 1: Training examples for SVM

(a) [5 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 point] Find a simple solution $(\mathbf{w}, \theta)$ that can separate the positive and negative examples given. (We are looking for an answer where the values for $\mathbf{w}$ and θ are

integers.) Remember that the vector which defines a line is perpendicular to that line.

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). Explain how you came to this solution.

Define $\mathbf{w} =$

Define $\theta =$

(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.

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

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

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 $C = 0$.
.7in

Submission Instructions

We will be using Gradescope to turn in writeup pdfs. You should have been automatically added to Gradescope. If you do not have access, please ask the TA staff on Piazza.

For this homework assignment, there are two Gradescope assignments:

– "Homework 3 - PDF": This is the assignment where you should upload your writeup as a PDF.