CS 5350/6350: Machine Learning Fall 2016

Gopal Menon

Homework 4

October 22, 2016

1 PAC learning

1. [20 points total] A factory assembles a product that consist of different parts. Suppose a robot was invented to recognize whether a product contains all the right parts. The rules of making products are very simple: 1) you are free to combine any of the parts as they are 2) you may also cut any of the parts into two distinct pieces before using them. You wonder how much effort a robot would need to figure out the what parts are used in the product.

   (a) [5 points] Suppose that a naive robot has to recognize products made using only rule 1. Given N available parts and each product made out of these constitutes a distinct hypothesis. How large would the hypothesis space be? Brief explain your answer.

   (b) [5 points] Suppose that an experienced worker follows both rules when making a product. How large is the hypothesis space now? Explain.

   (c) [10 points] An experienced worker decides to train the naive robot to discern the makeup of a product by showing you the product samples he has assembled. There are 6 available parts. If the robot would like to learn any product at 0.01 error with probability 99%, how many examples would the robot have to see?

2. [20 points, from Tom Mitchell's book] We have learned an expression for the number of training examples sufficient to ensure that every hypothesis will have true error no worse than ϵ plus its observed training error $error_S(h)$. In particular, we used Hoeffding bounds to derive

$$m \geq \frac{1}{2\epsilon^2}(\ln(|H|) + \ln(1/\delta)).$$

   Derive an alternative expression for the number of training examples sufficient to ensure that every hypothesis will have true error no worse than $(1 + \epsilon)error_S(h)$, where $0 \leq \epsilon \leq 1$. You can use general Chernoff bounds to derive such a result.

   Chernoff bounds: Suppose $X_1, \cdots, X_m$ are the outcomes of m independent coin flips (Bernoulli trials), where the probability of heads on any single trail is $Pr[X_i = 1] = p$

and the probability of tails is $Pr[X_i = 0] = 1 - p$. Define $S = X_1 + X_2 + \cdots + X_m$ to be the sum of these m trials. The expected value of S/m is $E[S/m] = p$. The Chernoff bounds govern the probabilty that S/m will differ from p by some factor $0 \leq \gamma \leq 1$.

$$\begin{aligned} Pr[S/m > (1+\gamma)p] &\leq e^{-mp\gamma^2/3} \\ Pr[S/m < (1-\gamma)p] &\leq e^{-mp\gamma^2/2} \end{aligned} \tag{1}$$

2 VC Dimensions

1. [10 points] Suppose you have a finite hypothesis space $\mathcal{C}$. Show that its VC dimension at most $\log_2 |\mathcal{C}|$ (Hint: You also prove this by contradiction.)

2. [10 points] Given some finite domain set, $\mathcal{X}$, and a number $k \leq |\mathcal{X}|$, figure out the VC-dimension of each of the following classes and prove your claims:

   (a) $\mathcal{H}_{=k}^{\mathcal{X}} = \{h \in \{0,1\}^{\mathcal{X}} : |x : h(x) = 1| = k\}$. That is , the set of all functions that assign the value 1 to exactly k elements of $\mathcal{X}$.

   (b) $\mathcal{H}_{\leq k}^{\mathcal{X}} = \{h \in \{0,1\}^{\mathcal{X}} : |x : h(x) = 1| \leq k \quad or \quad |x : h(x) = 0| \leq k\}$. That is , the set of all functions that assign the value 1 or 0 to at most k elements of $\mathcal{X}$.

3. [10 points] Suppose we have an instance space consisting of real numbers and a hypothesis space $\mathcal{H}$ consisting of *two* disjoint intervals, defined by $[a, b]$ and $[c, d]$. That is, a point $x \in \Re$ is labeled as positive if, and only if, either $a \leq x \leq b$ or $c \leq x \leq d$. Determine the VC dimension of $\mathcal{H}$?

4. [15 points] We have a learning problem where each example is a point in $\Re^2$. The concept class H is defined as follows: A function $h \in H$ is specified by two parameters a and b. An example $\mathbf{x} = \{x_1, x_2\}$ in $\Re^2$ is labeled as $+$ if and only if $x_1 + x_2 \geq a$ and $x_1 - x_2 \leq b$ and is labeled $-$ otherwise.

   For example, if we set $a = 0, b = 0$, the grey region in figure 1 is the region of $\mathbf{x} = \{x_1, x_2\}$ that has label $+1$.

   What is the VC dimension of this class?

5. [**For 6350 Students,** 15 points] Let two hypothesis classes H_1 and H_2 satisfy $H_1 \subseteq H_2$. Prove: $VC(H_1) \leq VC(H_2)$.

3 AdaBoost

[15 points] You are given the following examples in the Table 1. You need to learning a model that minimize the error on this small dataset.

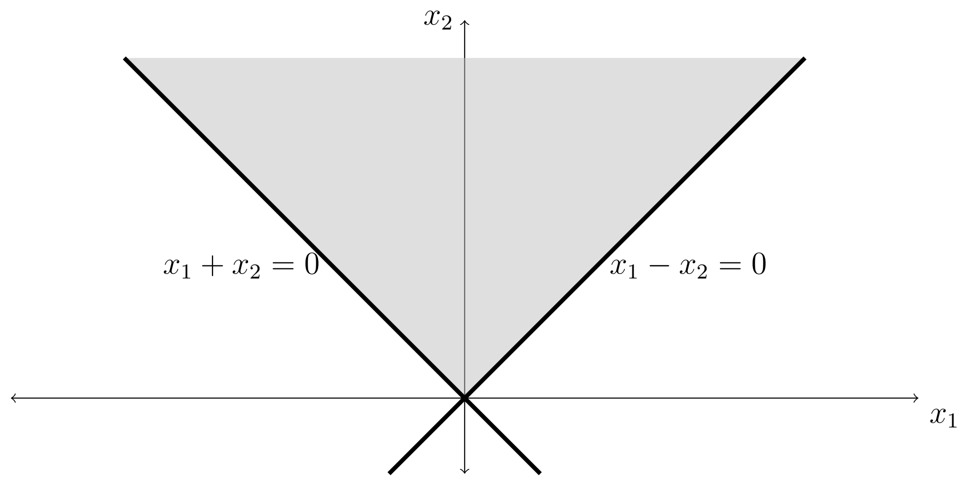

Figure 1: An example with $a = 0, b = 0$. All points in the gray region (extending infinitely) shows the region that will be labeled as positive.

Table 1: training set

$\mathbf{x} = [x_1, x_2]$	y
[1,1]	-1
[1,-1]	+1
[-1,-1]	-1
[-1,-1]	-1

Assuming you are also given the following 4 weak hypothesis classifiers

$$\begin{aligned} h_a(\mathbf{x}) &= \operatorname{sgn}(x_1) \\ h_b(\mathbf{x}) &= \operatorname{sgn}(x_1 - 2) \\ h_c(\mathbf{x}) &= -\operatorname{sgn}(x_1) \\ h_d(\mathbf{x}) &= -\operatorname{sgn}(x_2) \end{aligned}$$

Treat them as your weak classifiers(rule of thumb) for the following question.

Step through the full AdaBoost algorithm (Lecture Boosting slide P34) for 4 rounds by choosing h_t from the above 4 weak classifiers. Remember that you need to **choose a hypothesis** from h_a, h_b, h_c, h_d whose weighted classification error is **better than chance**. However, in this question, for easier grading, we have chosen h_a as the first hypothesis and show the values of ϵ_1, α_1, Z_1, D_1 in Table 2.

For you answer, please follow the table template, report the hypothesis you choose and all the ϵ_t, α_t, Z_t, D_t, and the final hypothesis $H_{final}(x)$ for *four subsequent rounds*.

Table 2: Choose $h_a(\mathbf{x}) = \mathrm{sgn}(x_1), \epsilon_1 = 1/4, \alpha_1 = \frac{\ln 3}{2}, Z_1 = \frac{\sqrt{2}}{3}$

$\mathbf{x} = [x_1, x_2]$	y_i	D_1	$D_1(i)y_ih_t(\mathbf{x_i})$	D_2
[1,1]	-1	1/4	-1/4	
[1,-1]	+1	1/4	1/4	
[-1,-1]	-1	1/4	1/4	
[-1,-1]	-1	1/4	1/4	