CS 6350: Machine Learning Fall 2016

Gopal Menon
Homework 4

October 29, 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.

   Since we have N available parts, the product can be made with any of the following number of parts, which would be the hypotheses space H.

   $$H = \{1, 2, 3, \ldots, N-1, N\}$$

   This means that the size of the hypotheses space H is

   $$\begin{aligned} |H| &= 1 + 2 + 3 + \ldots + N - 1 + N \\ &= \frac{N(N+1)}{2} \end{aligned}$$

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

   Since each of the N parts can be used as is or broken into two, we have a total of $3N$ possible parts. The reason is that we have N original parts and if each part is broken into two, we have $2N$ broken parts. The total sum of the number of parts is $N + 2N$ or $3N$. The product can be made out of unbroken parts or with broken parts or a combination of both. If broken parts are used, then both halves of the broken part are to be used. This means that the total number of parts from which the product can be made, is reduced to $2N$. Based on an argument similar to the one given above for the previous answer, the size of the hypothesis space H will be

$$|H| = \frac{2N(2N+1)}{2}$$

(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?

In order to learn a hypothesis with n dimensional features, with an error less than ϵ, with a probability of $1-\delta$, we would need m training samples, where m is given by

$$m > \frac{1}{\epsilon}\left(log_e(|H|) + log_e\left(\frac{1}{\delta}\right)\right)$$

In the case given above, since there are $2N$ different parts that can be chosen to make up the product, the input space can be thought of as having $2N$ dimensions since each dimension will be a binary variable which signifies whether a part has been used or not. So $|H| = \frac{2N(2N+1)}{2} = \frac{2\times 6(2\times 6+1)}{2} = \frac{12\times 13}{2} = 78$, $\epsilon = 0.01$ and $\delta = 1.00 - 0.99 = 0.01$.

Substituting these values into the above in-equation, we get

$$\begin{aligned} m &> \frac{1}{0.01}\left(log_e(78) + log_e\left(\frac{1}{0.01}\right)\right) \\ &> 100 \times (log_e(78) + log_e(100)) \\ &> 100 \times (4.36 + 4.60) \\ &> 896.19 \end{aligned}$$

This means that the robot will need to look at at least 897 examples to learn a hyphothesis.

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 $0 \leq \gamma \leq 1$.

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

The empirical error is defined as $err_S(h) = \frac{|\{f(x) \neq h(x)\}|}{m}$ and the generalization error is defined as $err_D(h) = Pr\left[f(x) \neq h(x)\right]$, where S is the training set, h is the hypothesis that is learnt, $f(x)$ is the unknown true function, m is the number of training samples and D is the distribution from which samples are drawn.

Using the Chernoff bounds, we can find the expression for the probability that the expected error (or in other words the generalization error) for a single hypothesis will differ from the empirical error by some factor based on ϵ, where $0 \leq \epsilon \leq 1$. This expression will be

$$Pr\left[err_D(h) > (1+\epsilon)err_S(h)\right] \leq e^{\frac{-mp\epsilon^2}{3}}$$

For the case where the hypothesis h is chosen from the set of all possible hypotheses H, the learning algorithm will choose the hypothesis that has minimum empirical error. Using the union bound, we can say that the probability that there exists hypothesis $h \in H$, such that the expected error will differ from the empirical error, will be

$$Pr\left[\exists h; err_D(h) > (1+\epsilon)err_S(h)\right] \leq |H|\, e^{\frac{-mp\epsilon^2}{3}}$$

In order to minimize the generalization error, we need the above probability to be less than δ. This gives us

$$\begin{aligned} |H|\, e^{\frac{-mp\epsilon^2}{3}} &\leq \delta \\ log_e\left(|H|\right) - \frac{mp\epsilon^2}{3} log_e e &\leq log_e\left(\delta\right) \\ \frac{mp\epsilon^2}{3} &\geq log_e\left(|H|\right) - log_e\left(\delta\right) \\ m &\geq \frac{3}{p\epsilon^2}\left(log_e\left(|H|\right) + log_e\left(\frac{1}{\delta}\right)\right) \end{aligned}$$

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.)

   The growth function $m_{\mathcal{C}}(N)$ for a hypothesis class $\mathcal{C}$ is defined as the maximum number of dichotomies than can be generated by $\mathcal{C}$ on any N points.

The VC dimension of $\mathcal{C}$,$d_{VC}(\mathcal{C})$, is the largest N for which $m_{\mathcal{C}}(N) = 2^N$. In other words, $d_{VC}(\mathcal{C})$ is the largest N that can be split into all possible dichotomies by the hypothesis class $\mathcal{C}$.

Each hypothesis in the hypothesis class $\mathcal{C}$ can at the maximum generate a distinct dichotomy. That means, there could be a maximum of $|\mathcal{C}|$ dichotomies that can be generated by the hypothesis class $\mathcal{C}$.

This means that $d_{VC}(\mathcal{C})$ is the largest N such that

$$\begin{aligned} m_{\mathcal{C}}(N) = 2^N &= |\mathcal{C}| \\ N \log_2 2 &= \log_2 |\mathcal{C}| \\ N &= \log_2 |\mathcal{C}| \end{aligned}$$

This means that the VC dimension at most $\log_2 |\mathcal{C}|$

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

   $d_{VC}(\mathcal{H})$, the VC dimension of the hypothesis class $\mathcal{H}$, is defined as the largest N for which $m_{\mathcal{H}}(N) = 2^N$. The definition of $m_{\mathcal{H}}$ is similar to the one given above.

   For the case of maximum number of dichotomies, we can consider that each hypothesis in $\mathcal{H}$ creates a distinct dichotomy. To find $d_{VC}(\mathcal{H})$ corresponding to this case, since we know that at exactly k elements are assigned a label 1 and the remaining elements are assigned label 0, for the case of the VC dimension, the maximum number of dichotomies must be a power of 2. Consider a subset of elements of $\mathcal{X}$, where k elements are assigned label 1 and y elements are assigned label 0.

$$\begin{aligned} |\mathcal{H}| &= 2^{k+y} \\ \log_2 |\mathcal{H}| &= k + y \\ y &= \log_2 |\mathcal{H}| - k \end{aligned}$$

   Since $d_{VC}(\mathcal{H})$, the VC dimension of the hypothesis class $\mathcal{H}$, is defined as the largest N for which $m_{\mathcal{H}}(N) = 2^N$

$$\begin{aligned} 2^{d_{VC}(\mathcal{H})} &= 2^{k+y} \\ &= 2^{k+\log_2|\mathcal{H}|-k} \\ &= 2^{log_2|\mathcal{H}|} \\ \implies d_{VC}(\mathcal{H}) &= \log_2 |\mathcal{H}| \end{aligned}$$

   (b) $\mathcal{H}^{\mathcal{X}}_{\leq k} = \{h \in \{0,1\}^{\mathcal{X}} : |x : h(x) = 1| \leq k \quad or \quad |x : h(x) = 0| \leq k\}$. 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 $+1$.

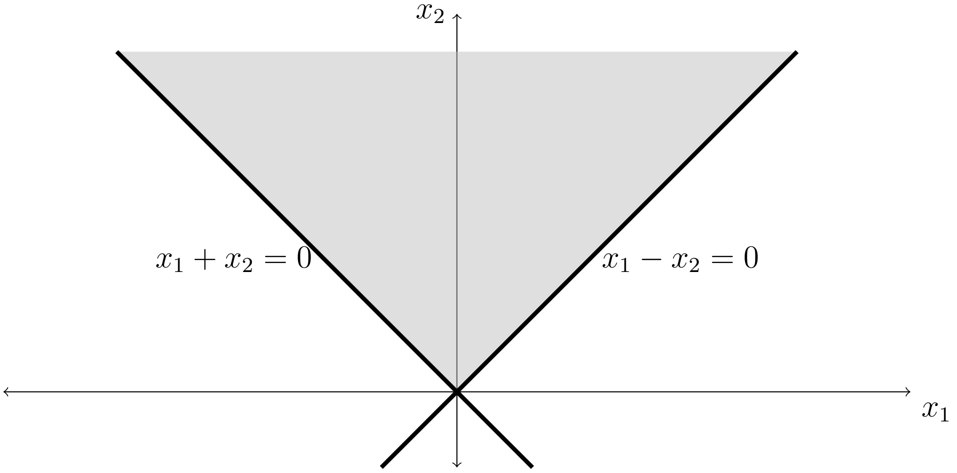

Figure 1: An example with $a = 0, b = 0$. All points in the gray region (extending infinitely) shows the region that will be labeled as positive.

   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.

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

$\mathbf{x} = [x_1, x_2]$	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	