CS 6350: Machine Learning Fall 2016

Gopal Menon
Homework 4

November 1, 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.

   Each product can be made up of 1 to N parts. When a product is made up of 1 part, it can be from any of N parts and so there are $\binom{N}{1}$ ways of choosing the one part. Similarly, there are $\binom{N}{2}$ ways of making a product by choosing from 2 parts. So the total number of ways of making the product would be the sum of all these. Or in other words, the size of the hypothesis set H would be

   $$|H| = \sum_{i=1}^{N} \binom{N}{i}$$

   Another way of finding $|H|$ is to consider N slots and each slot tells you whether the part corresponding to the slot number has been used or not. We have 2 choices for each slot (whether to use the part or not), so for N slots, we have 2^N choices. However, we need to not count the case where none of the parts are used (since a product cannot be made without using a part) and so we would need to subtract 1 from the total. That is

   $$|H| = 2^N - 1$$

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

Given a set of N whole parts and N broken parts, considering each part, the product may be built by

(a) not choosing this part
(b) choosing this whole part
(c) choosing only piece 1 of this part
(d) choosing only piece 2 of this part

The assumption is that using both pieces of a broken part is the same as using the whole part. Similar to the above case, we can consider N slots and each slot tells you whether the part corresponding to the slot number has been not used, used as a whole, only part 1 is used or only part 2 is used. As done above, we need to subtract 1 from the total as this accounts for the case where no part is used. So

$$|H| = 4^N - 1$$

(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 an error less than ϵ, with a probability of $1-\delta$, we would need m training samples, where m is given by

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

In the case given above $|H| = 4^6 - 1 = 4096 - 1 = 4095$, $\epsilon = 0.01$ and $\delta = 1.00 - 0.99 = 0.01$.

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

$$\begin{aligned} m &> \frac{1}{0.01}\left(\ln(4095) + \ln\left(\frac{1}{0.01}\right)\right) \\ &> 100 \times (\ln(4095) + \ln(100)) \\ &> 100 \times (8.32 + 4.60) \\ &> 1292 \end{aligned}$$

This means that the robot will need to look at at least 1293 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 trial 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.

In the Chernoff bound, $\frac{S}{m}$ is equivalent to the empirical error $err_S(h)$ and p is equivalent to the generalization error $err_D(h)$. Using these equivalent values in the second Chernoff bound, we get the relationship between the empirical error and the generalization error for a single hypothesis

$$Pr\left[Err_S(h) < (1-\gamma)Err_D(h)\right] \leq e^{\frac{-mp\gamma^2}{2}}$$

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 empirical error will be less than the generalization error by a factor of $1 - \gamma$, will be

$$Pr\left[\exists h; Err_S(h) < (1-\gamma)Err_D(h)\right] \leq |H|\, e^{\frac{-mp\gamma^2}{2}}$$

if we set the value of $\frac{1}{1-\gamma}$ to be $1 + \epsilon$ with the condition that $0 \leq \epsilon \leq 1$, we get

$$\begin{aligned} \frac{1}{1-\gamma} &= 1+\epsilon \\ \implies \frac{1}{1+\epsilon} &= 1-\gamma \\ \implies \gamma &= 1 - \frac{1}{1+\epsilon} \\ &= \frac{1+\epsilon-1}{1+\epsilon} \\ &= \frac{\epsilon}{1+\epsilon} \end{aligned}$$

We can than rewrite the above probability relation as

$$\begin{aligned}
Pr\left[\exists h; (1-\gamma)Err_D(h) > Err_S(h)\right] &\leq |H|\, e^{\frac{-mp\gamma^2}{2}} \\
Pr\left[\exists h; \frac{1}{1+\epsilon}Err_D(h) > Err_S(h)\right] &\leq |H|\, e^{\frac{-mp\gamma^2}{2}} \\
Pr\left[\exists h; Err_D(h) > (1+\epsilon)Err_S(h)\right] &\leq |H|\, e^{\frac{-mp\gamma^2}{2}} \\
&\leq |H|\, e^{\frac{-mp(\frac{\epsilon}{1+\epsilon})^2}{2}} \\
&\leq |H|\, e^{\frac{-mp}{2}\frac{\epsilon^2}{(1+\epsilon)^2}}
\end{aligned}$$

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

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

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

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

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.

   If we choose N to be 2, and if we choose k to be 0, 1 or 2, then all possible dichotomies cannot be generated. So VC dimension has to be less than 2. A similar argument shows that VC dimension has to be less than 1, since all possible dichotomies cannot be generated for this case. So the VC dimension is 0.

   (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 $\mathcal{X}$.

   Since the value of 1 is assigned to at most k elements, then VC dimension can be k. The reason is that if we choose k elements from $\mathcal{X}$, if we assume that the functions in the hypothesis class will assign a value of 1 to all elements between 0 and k, then this will cover all dichotomies (or in other words all possible binary values) if we consider the k elements to be the bits in a k-bit binary number.

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 $\mathcal{H}$?

   Since the two intervals are disjoint, there are five regions in the instance space that are defined by the intervals:

   (a) before the first interval

   (b) inside the first interval

   (c) between the first and second intervals

   (d) inside the second interval

   (e) after the second interval

   So the maximum number of points that can possibly be split into all possible dichotomies is 5. If the 5 points can be split into all possible dichotomies then the

dichotomy $\{1, 0, 1, 0, 1\}$ should be one of them (I am representing a positive label by 1 and a negative label by 0). However, this dichotomy is not possible since there are only two intervals where the label can be 1 and the label $\{1, 0, 1, 0, 1\}$ requires three such intervals.

Consider 4 points. These 4 points can be split into the dichotomies $\{\{0,0,0,0\}, \{0,0,0,1\}, \{0,0,1,0\}, \{0,0,1,1\}, \{0,1,0,0\}, \{0,1,0,1\}, \{0,1,1,0\}, \{0,1,1,1\}, \{1,0,0,0\}, \{1,0,0,1\}, \{1,0,1,0\}, \{1,0,1,1\}, \{1,1,0,0\}, \{1,1,0,1\}, \{1,1,1,0\}, \{1,1,1,1\}\}$. These are all possible dichotomies that can be generated using 4 points and the hypothesis space $\mathcal{H}$.

So $d_{VC}(\mathcal{H}) = 4$.

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 -1 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 $+1$.

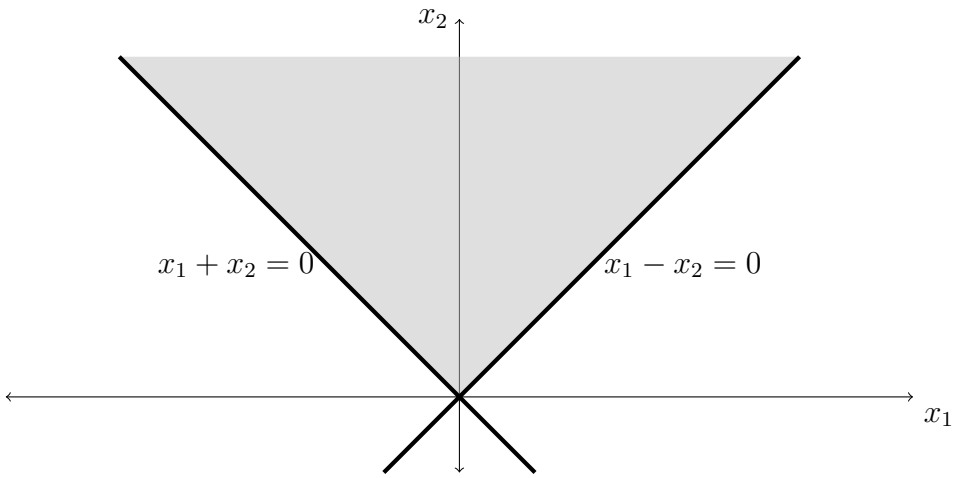

Figure 1: An example with $a = 0, b = 0$. All points in the gray region (extending infinitely) shows the region that will be labeled as positive.

   What is the VC dimension of this class?

   The concept class contains individual hypothesis functions that consist of two arms that can intersect anywhere on $\Re^2$ and can cover an infinite arc from any angle from 0° to 360°. Consider the case of three collinear points which have labels $\{+1, -1, +1\}$. This classification clearly cannot be satisfied by any of the hypothesis in the hypothesis class. Any two points (2^1 points) with any distribution of $+1$ and -1 can be covered.

   So $d_{VC}(H) = 1$.

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

   Consider any set of points N that are labelled by H_2. The same set of points may not result in the same number of dichotomies by H_1 as those produced by H_2. The reason is that H_2 may have hypotheses that are not present in H_1, whereas all hypotheses in H_1 will be present in H_2. This means that H_2 can be more expressive than H_1 and can shatter more points. This means that $d_{VC}(H_1) \leq d_{VC}(H_2)$.

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

$$h_a(\mathbf{x}) = \text{sgn}(x_1)$$
$$h_b(\mathbf{x}) = \text{sgn}(x_1 - 2)$$
$$h_c(\mathbf{x}) = -\text{sgn}(x_1)$$
$$h_d(\mathbf{x}) = -\text{sgn}(x_2)$$

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 6.

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 $h_a(\mathbf{x}) = \text{sgn}(x_1), \epsilon_1 = 1/4, \alpha_1 = \frac{\ln 3}{2} = 0.5493, Z_1 = \frac{\sqrt{3}}{2}$

$\mathbf{x} = [x_1, x_2]$	y_i	$h_a(x)$	D_1	$D_1(i)y_i h_t(\mathbf{x_i})$	D_2
[1,1]	-1	1	1/4	-1/4	3/6
[1,-1]	1	1	1/4	1/4	1/6
[-1,-1]	-1	-1	1/4	1/4	1/6
[-1,1]	-1	-1	1/4	1/4	1/6

Table 3: Choose $h_b(\mathbf{x}) = \text{sgn}(x_1 - 2), \epsilon_2 = 1/6, \alpha_2 = 0.8047, Z_2 = 0.7454$

$\mathbf{x} = [x_1, x_2]$	y_i	$h_b(x)$	D_2	$D_2(i)y_i h_t(\mathbf{x_i})$	D_3
[1,1]	-1	-1	3/6	3/6	3/10
[1,-1]	1	-1	1/6	-1/6	5/10
[-1,-1]	-1	-1	1/6	1/6	1/10
[-1,1]	-1	-1	1/6	1/6	1/10

Table 4: Choose $h_c(\mathbf{x}) = -\text{sgn}(x_1), \epsilon_3 = 7/10, \alpha_3 = -0.4236, Z_3 = 0.9165$

$\mathbf{x} = [x_1, x_2]$	y_i	$h_c(x)$	D_3	$D_3(i)y_i h_t(\mathbf{x_i})$	D_4
[1,1]	-1	-1	3/10	3/10	0.5000
[1,-1]	1	-1	5/10	-5/10	0.3751
[-1,-1]	-1	1	1/10	-1/10	0.0714
[-1,1]	-1	1	1/10	-1/10	0.0714

The hypothesis $h_c(x)$ had an error worse than 0.5 and so was not used. The next iteration will use the same weights that were used in the third iteration.

Table 5: Choose $h_d(\mathbf{x}) = -\text{sgn}(x_2), \epsilon_4 = 1/10, \alpha_4 = 1.0986, Z_4 = 0.6$

$\mathbf{x} = [x_1, x_2]$	y_i	$h_d(x)$	D_4	$D_4(i)y_i h_t(\mathbf{x_i})$	D_5
[1,1]	-1	-1	3/10	3/10	0.1667
[1,-1]	1	1	5/10	5/10	0.2778
[-1,-1]	-1	1	1/10	-1/10	0.5000
[-1,1]	-1	-1	1/10	1/10	0.0556

Final Hypothesis: $H_{final}(x) = 0.5493 \times h_a(x) + 0.8047 \times h_b(x) + 1.0986 \times h_d(x)$

Table 6: Apply Final Hypothesis on test data

$\mathbf{x} = [x_1, x_2]$	y_i	$h_a(x)$	$h_b(x)$	$h_d(x)$	$H_{final}(x)$
[1,1]	-1	1	-1	-1	-1
[1,-1]	1	1	-1	1	1
[-1,-1]	-1	-1	-1	1	-1
[-1,1]	-1	-1	-1	-1	-1