# CSE 512 – Homework I

Hang Zhao ID: 112524698

March 16, 2020

Due **Mar 15 (Sunday), 2020 [11:59 pm]**.

## 1 Theory

I: A "warm up" problem
Consider instances of X drawn from the uniform distribution D on $[-1, 1]$. Let f denote the actual labeling function mapping each instance to its label $y \in \{-1, 1\}$ with probabilities

$$Pr(y = 1|x > 0) = 0.9$$

$$Pr(y = -1|x > 0) = 0.1$$
$$Pr(y = 1|x <= 0) = 0.1$$
$$Pr(y = -1|x <= 0) = 0.9$$

The hypothesis h predicts the label for each instance as defined below"

$$h(x) = \begin{cases} \text{1, if x > 0} \\ \text{-1, otherwise.} \end{cases}$$

Measure the success of this predictor by calculating the training error for h.

$$P[h(X) \neq f(X)] = P[Y = -1|X > 0] * P[h(X) = 1|X > 0] + P[Y = 1|X \leq 0] * P[h(X) = -1|X \leq 0]$$
$$= 0.1 * 0.5 + 0.1 * 0.5$$
$$= 0.1$$

II: Bayes Optimal Predictor

Show that for every probability distribution D, the Bayes optimal predictor $f_D$ is, in fact, optimal. That is, show that for every classifier $g : X \to 0, 1$,

$$L_D(f_D) \leq L_D(g).$$

Define $x \in X$ and p as the probability that a label is positive given x. Then we have:

$$P[f_D(X) \neq y|X = x] = 1_{[p<\frac{1}{2}]} \cdot P[Y = 1|X = x] + 1_{[p\geq\frac{1}{2}]} \cdot P[Y = 0|X = x]$$

$$= 1_{[p<\frac{1}{2}]} \cdot p + 1_{[p\geq\frac{1}{2}]} \cdot (1-p)$$

If p is larger or equal to $\frac{1}{2}$, then the first term will become to 0, and the second term becomes to 1 - p. If p is smaller than $\frac{1}{2}$, then the first term will become to p, and the second term becomes 0. We can conclude that $1_{[p<\frac{1}{2}]} \cdot p + 1_{[p\geq\frac{1}{2}]} \cdot (1-p)$ always get the smaller value of p and (1-p). We can denote it as $min\{p, 1-p\}$.

Define g as a classifier learned from training dataset, then we can get:

$$P[g(X) \neq Y|X=x] = P[g(X)=1|X=x] \cdot P[Y=0|X=x] + P[g(X)=0|X=x] \cdot P[Y=1|X=x]$$
$$= P[g(X)=1|X=x] \cdot (1-p) + P[g(x)=0|X=x] \cdot p$$

since p and (1- p) both greater or equal to $min\{p, 1-p\}$. We can have:

$$\geq P[g(X)=1|X=x] \cdot min\{p, 1-p\} + P[g(X)=0|X=x] \cdot min\{p, 1-p\}$$
$$= min\{p, 1-p\} \cdot (P[g(X)=1|X=x] + P[g(X)=0|X=x])$$
$$= min\{p, 1-p\}$$

Thus we can conclude that:

$$L_D(g) = E_{(x,y)\sim D}[1_{[g(x)\neq y]}]$$
$$\geq E_{x\sim D_x} min\{p, 1-p\}$$

Hence we know that:

$$L_D(f_D) \leq L_D(g).$$

III: Perceptron with a Learning Rate

Let us modify the perceptron algorithm as follows:
In the update step, instead of setting $w^{(t+1)} = w^{(t)} + y_i x_i$ every time there is a misclassification, we set $w^{(t+1)} = w^{(t)} + \eta y_i x_i$ instead, for some $0 < \eta < 1$. Show that this modified perceptron will:
(a) perform the same number of iterations as the original, and
(b) converge to a vector that points to the same direction as the output of the original perceptron.

(a) For every w and x $\in R^d$, at each time of iteration, $w = \sum_i y_i x_i$ and at each time after multiplying a value between zero and one will not change the direction of w thus the sign of $< w, x >$ and $< \eta w, x >$ is the same, so changing the perceptron algorithm from $w^{(t+1)} = w^{(t)} + y_i x_i$ to $w^{(t+1)} = w^{(t)} + \eta y_i x_i$ will not affect the number of iterations.
(b) For every w and x $\in R^d$, at each time of iteration, $w = \sum_i y_i x_i$ and at each time after multiplying a value between zero and one will not change the direction of w thus the sign of $< w, x >$ and $< \eta w, x >$ is the same, so changing the perceptron algorithm from $w^{(t+1)} = w^{(t)} + y_i x_i$ to $w^{(t+1)} = w^{(t)} + \eta y_i x_i$ will not affect the the direction of the outputs.

IV: Unidentical Distributions

Let X be a domain and let $D_1, D_2, ....D_m$ be a sequence of distributions over X. Let H be a finite class of binary classifier over X and let $f \in H$. Suppose we are getting a sample S of m examples such that the instances are independent but not identically distributed, the $i^{th}$ instance is sampled from $D_i$ and then $y_i$ is set to be $f(x_i)$. Let $\bar{D}_m$ denote the average, i.e, $\bar{D}_m = (D_1 + .... + D_m)/m$. Fix an accuracy parameter $\epsilon \in (0, 1)$. Show that

$$Pr[\exists h \in H s.t. L_{\bar{D}_m,f}(h) > \epsilon \; and \; L_{S,f}(h) = 0] \leq |H|e^{-em}$$

If there exist some $h \in H$ such that $L_{\bar{D}_m,f}(h) > \epsilon$ which means the probability of the predict value is not equal to observed value would be greater than $\epsilon$, thus we can know that the probability of the predict value is equal to observed value would be less than $1 - \epsilon$. We have:

$$\frac{P_{X\sim D_1}[h(X) = f(X)] + ... + P_{X\sim D_m}[h(X) = f(X)]}{m} < 1 - \epsilon$$

Since there are total m probabilities and each of them is less than $1 - \epsilon$.
From the book equation 2.8 we know:

$$D^m(S|_x : L_S(h) = 0) = D^m(S|_x : \forall i, h(x_i) = f(x_i))$$

$$= \prod_{i=1}^{m} D(x_i : h(x_i) = f(x_i)).$$

We can get:

$$P_{S\sim \prod_{i=1}^{m} D_i}[L_S(h) = 0] = \prod_{i=1}^{m} P_{x\sim D_i}[h(X) = f(X)]$$

According to the geometric-arithmetic mean inequality:

$$\frac{x_1 + ... + x_n}{m} \geq (x_1 \cdot ... \cdot x_n)^{\frac{1}{m}}$$

$$((\prod_{i=1}^{m} P_{x\sim D_i}[h(X) = f(X)])^{\frac{1}{m}})^m$$

$$\leq (\frac{\sum_{i=1}^{m} P_{X\sim D_i}[h(X) = f(X)]}{m})^m$$

Then from inequality 2.9 in the book and applied union bound we can get:

$$< (1 - \epsilon)^m$$

$$\leq e^{-\epsilon m}$$

Hence we can get:

$$Pr[\exists h \in H s.t. L_{\bar{D}_m,f}(h) > \epsilon \; and \; L_{S,f}(h) = 0] \leq |H|e^{-em}$$

V: Vapnik-Chervonenkis (VC) Dimension

(a) Let $H^d$ be the class of axis-aligned rectangles in $R^d$. Prove that $VCdim(H^d) = 2d$.
(b) The above example might suggest that the VD dimension of a hypothesis class is bounded above by some multiple of the number of parameters used to define the hypothesis class. But this is not always the case. This question illustrates that a hypothesis class may be very complex, and not even learnable, even if it is defined with a very small number of parameters. Consider the hypothesis class of sine functions:

$$H = \{x \to \lceil sin(\theta x) \rceil : \theta \in R\}$$

and consider $\lceil -1 \rceil = 0$. Prove that $VCdim(H) = \infty$.

(a) First we need to show $VCdim(H^d) \geq 2d$ . Consider a 2 dimensional case, we define the classifier as $h_{a_1,b_1,....,a_d,b_d}(x_1,....,x_d) = \prod_{i=1}^{d} 1_{[x_i \in [a_i,b_i]]}$ where $\forall i \in [d], a_i \leq b_i$. Then consider a set of 2d points, they are being set as 1 or -1 on only one of d dimensions and for other dimensions the value will be 0. We can see these 2d points can be shattered by an axis-aligned rectangle. Thus we can know $VCdim(H^d) \geq 2d$.
Next, we need to prove $VCdim(H^d) < 2d+1$ and it cannot be shattered by an axis-aligned rectangle. Consider a set of 2d+1 points, using the minimum and maximum value in each dimension to be the lower bound and upper bound as $a_i$ and $b_i$. Because there are 2d+1 points, based the pigeonhole principle, there must have at least one point inside this axis-aligned rectangle. If we mark it as negative, and the rest of them as positive, then this case cannot be shatter by an axis-aligned rectangle. And we know $VCdim(H^d) < 2d+1$.
Together we can get with the class of axis-aligned rectangles in $R^d$, $VCdim(H^d) = 2d$.
(b) For any base $b \geq 2 (b \in N)$ and any real number of x, we can denote x as:

$$x = a_n a_{n-1} a_{n-2} ... a_0 . b_1 b_2 b_3 b_4 ...$$

and basically it means $x = a_0 + a_1 \cdot b + a_2 \cdot b^2 + .... + a_n b^n + b_1 b^{-1} + b_2 b^{-2} + ...$ . Based on this, we define $0.x_1x_2x_3...$ as the binary expansion of $x \in (0,1)$. Now we can write:

$$sin(2^m \pi x) = sin(2^m \pi (0.x_1 x_2 ...))$$

for $x \in (0,1)$.

$$sin(2^m \pi (0.x_1 x_2 ...)) = sin(2\pi 2^{m-1} (0.x_1 x_2 ...)$$

After shifting m-1 position we have:

$$= sin(2\pi (x_1 x_2 ... x_{m-1} . x_m x_{m+1} ..))$$

and since it is the multiple of $2\pi$, so we can eliminate the value above zero.

$$= sin(2\pi (x_1 x_2 ... x_{m-1} . x_m x_{m+1} ..) - 2\pi (x_1 x_2 .... x_{m-1} .0))$$

$$= sin(2\pi (0.x_m x_{m+1} ....)).$$

If $x_m = 0$, then the maximum value of $0.x_m x_{m+1}.... = 1 * 2^{-2} + 1 * 2^{-3} .... 1 * 2^{-n} = \frac{1}{2}$ and minimum value is 0. So $2\pi(0.x_m x_{m+1}....) \in (0, \pi)$ and $sin(2\pi(0.x_m x_{m+1}....)) > 0$. The ceiling of $sin(2^m \pi x)$ is 1. If $x_m = 1$, then the maximum value of $0.x_m x_{m+1}.... = 1 * 2^{-1} + 1 * 2^{-2} + 1 * 2^{-3} .... 1 * 2^{-n} = 1$ and minimum value is $2^{-1}$. So $2\pi(0.x_m x_{m+1}....) \in [\pi, 2\pi)$ and $sin(2\pi(0.x_m x_{m+1}....)) \leq 0$. The ceiling of $sin(2^m \pi x)$ is 0. Thus we can get the ceiling of $sin(2^m \pi x)$ is $1 - x_m$.
Then in order to prove the VC dimension of H is infinity, we choose n points which can be shatter by H and these n points belong to the range between 0 and 1 inclusive. Then we denote all the points as $x_1, ...., x_n$:

$$x_1 = 0.0000...11$$

$$x_2 = 0.0000...11$$

.......

$$x_{n-1} = 0.0011...11$$

$$x_n = 0.0101...01$$

Given the labeling 1 for all instances, we choose $h(x)$ to be the ceiling of $sin(2^1x)$ which will give the first bit of the binary expansion. And if we label 1 for $x_1, ...x_{n-1}$ and label 0 for the last element, we can choose $h(x)$ to be the ceiling of $sin(2^2x)$ which will give us the second bit of the binary expansion and so on. Thus we can say that all the instance can be given any label by $h \in H$ and it can be shattered. So the VC dimension of H is infinite.

VI: Boosting

We have intuitively argued in class that in class that in AdaBoost, the probability distribution is updated in a way that forces the next iteration's weak learner to focus on the mistakes made in the current iteration. Show that the error of the current weak learner $h_t$ on the next iteration's distribution $D^{(t+1)}$ is exactly 0.5. That is, show that for every $t \in 1, 2, ..., T$,

$$\sum_{i=1}^{m} D_i^{(t+1)} I_{h_t(x_i)} \neq \frac{1}{2}$$

According to the definition, we know that:

$$Z_t = \sum_{i:h_t(x_i)=y_i} D_t(i)e^{-\alpha_t} + \sum_{i:h_t(x_i)\neq y_i} D_t(i)e^{\alpha_t}$$

$$= e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t$$

Since we have $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$ and we plugged it into the formula we can have:

$$= \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}(1-\epsilon_t) + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t$$

$$= 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Notice the error part is $\sqrt{\epsilon_t(1-\epsilon_t)}$ and this error of the current weak learner $h_t$ on the next iteration's distribution $D^{(t+1)}$ is exactly 0.5.

VII: Cross-validation

Cross-validation works well in practice, but there are some cases where it might fail. As such, there are no theoretical guarantees. Suppose, in a scenario, that the label is chosen at random according to $P[y = 1] = P[y = 0] = 0.5$. Consider a learning algorithm that outputs the constant predictor $h(x) = 1$ if the number of 1s in the training set labels is odd, and $h(x) = 0$ otherwise. Prove that the difference between the leave-one-out estimate and the true error in such a case is always 0.5.

Define S to be the independent and identical distributed sample data. Define h as a hypothesis mentioned in the question. Because h is a constant function, so $L_{D,f}(h) = \frac{1}{2}$. Now we should consider different situations, if the number 1s in S set labels is odd. Then for some fold $(x, y) \in S$, if the number of 1s in $S\backslash(x, y)$ is odd, then y here must 0. Using the data set $S\backslash(x, y)$ for training and (x,y) for cross validation, the constant predictor $h(x) = 1$, so the leave one out estimate error is 1. If the number of 1s in $S\backslash(x, y)$ is even, then y here must be 1. Using the data set $S\backslash(x, y)$ for training and (x,y) for cross validation, the constant predictor $h(x) = 0$, so the leave one out estimate error is 1. So the average estimate error is 1 and the difference between estimate error and true error is $\frac{1}{2}$. If the number 1s in S set labels is even. Then for some fold $(x, y) \in S$, if the number of 1s in $S \backslash (x, y)$ is even, then y here must 0. Using the data set $S \backslash (x, y)$ for training and (x,y) for cross validation, the constant predictor $h(x) = 0$, so the leave one out estimate error is 0. If the number of 1s in $S \backslash (x, y)$ is odd, then y here must be 1. Using the data set $S \backslash (x, y)$ for training and (x,y) for cross validation, the constant predictor $h(x) = 0$, so the leave one out estimate error is 0. So the average estimate error is 0 and the difference between estimate error and true error is $\frac{1}{2}$.