CSE512 Machine Learning

Chumki Acharya (112683478)

14 June, 2020

Homework-1

Question 1

Solution:

As given, the instances of X are drawn from a uniform distribution on $[-1, 1]$, so the probability of $x \in X$ being less than or greater than 0 would both be 0.5.

$P(x > 0) = 1/2 = 0.5$
$P(x \leq 0) = 1/2 = 0.5$

According to the hypothesis h, we can assume the probability of an instance being wrongly predicted would fall in two cases:

1. when x $>$ 0 and y= -1, but h predicts the label as 1
2. when x $\leq$ 0 and y= 1, but h predicts the label as -1

Thus, the total training error of prediction by hypothesis h is,

TrainingError= $\Pr(y = -1|x > 0) * P(x > 0) + \Pr(y = 1|x \leq 0) * P(x \leq 0) = 0.1 * 0.5 + 0.1 * 0.5 = 0.1$

Therefore, the success of the predictor can be measured as:

$1-$ TrainingError $= 1 - 0.1 = 0.9$

Thus, we can say that the hypothesis has 90% success.

Question 2

Solution:

Let $x \in \mathcal{X}$. We can consider conditional probability to solve this. Assuming α_x be conditional probability of a positive label given x. We have

$$\begin{aligned}\mathbb{P}\left[f_{\mathcal{D}}(X) \neq y|X = x\right] &= \mathbb{1}\left[\alpha_x \geq 1/2\right] \cdot \mathbb{P}[Y = 0|X = x] + \mathbb{1}\left[\alpha_x < 1/2\right] \cdot \mathbb{P}[Y = 1|X = x] \\ &= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot (1 - \alpha_x) + \mathbb{1}_{[\alpha_x < 1/2]} \cdot \alpha_x \\ &= \min\left\{\alpha_x, 1 - \alpha_x\right\}\end{aligned}$$

If we take a classifier g from $\mathcal{X}$ to $\{0,1\}$. We can write

$$\mathbb{P}\left[g(X) \neq Y|X = x\right] = \mathbb{P}[g(X) = 0|X = x] \cdot \mathbb{P}[Y = 1|X = x] \quad + \mathbb{P}[g(X) = 1|X = x] \cdot \mathbb{P}[Y = 0|X = x]$$

Now using conditional probability we can write,

$$\begin{aligned}\mathbb{P}\left[g(X) \neq Y|X=x\right] &= \mathbb{P}[g(X)=0|X=x]\cdot\alpha_x + \mathbb{P}[g(X)=1|X=x]\cdot(1-\alpha_x)\\ &\geq \mathbb{P}[g(X)=0|X=x]\cdot\min\left\{\alpha_x, 1-\alpha_x\right\}\\ &+ \mathbb{P}[g(X)=1|x]\cdot\min\left\{\alpha_x, 1-\alpha_x\right\}\\ &= \min\left\{\alpha_x, 1-\alpha_x\right\}\end{aligned}$$

We know $\mathbb{P}\left[g(X) \neq Y|X=x\right]$ is true for every $x \in \mathcal{X}$.

Now by using law of expectation, we can say

$$\begin{aligned}L_{\mathcal{D}}(g) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[g(x)\neq y]}\right]\\ &= \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}\left[\mathbb{1}_{(g(x)\neq y]}|X=x\right]\right]\\ &\geq \mathbb{E}_{x\sim\mathcal{D}_X}\min\left\{\alpha_x, 1-\alpha_x\right\}\end{aligned}$$

Thus we can say

$$L_{\mathcal{D}}(g) \geq L_{\mathcal{D}}(f_D)$$

Hence

$$L_{\mathcal{D}}(f_D) \leq L_{\mathcal{D}}(g)$$

Question 3

Solution:

Let us assume some hypothesis h in H with true error L$_{\left(\bar{D}_m,f\right)}$(h) $> \epsilon$.

Also, we can say that,

$$\frac{\mathbb{P}_{X\sim D_1}[h(X)=f(X)]+\ldots+P_{X\sim D_1}[h(X)=f(X)]}{m} < 1-\epsilon\ldots.(1) \tag{1}$$

Also, $L_s(h)$ = 0 would mean that the hypothesis h is consistent with S.

Furthermore,

$$\begin{aligned}\mathbb{P}_{S\sim D}\left[\mathrm{L}_s(\mathrm{h})=0\right] &= \textstyle\prod_{i=1}^{m}\mathrm{P}_{X\sim D_i}[\mathrm{h(X)=f(X)}]\\ \mathbb{P}_{S\sim D}\left[\mathrm{L}_s(\mathrm{h})=0\right] &= \left(\left(\textstyle\prod_{i=1}^{m}P_{X\sim D_i}[h(X)=f(X)]\right)^{1/m}\right)^m\end{aligned} \tag{2}$$

Since, the arithmetic mean is greater than or equal to the geometric mean, using this inequality we can substitute the right hand side of the equation as,

$$\mathbb{P}_{S\sim D}\left[\mathrm{L}_s(\mathrm{h})=0\right] \leq \left(\frac{\sum_{i=1}^{m}P_{X\sim D_i}[h(X)=f(X)]}{m}\right)^m \tag{3}$$

Again, using the previous equation (1), we can say

$$\mathbb{P}_{S\sim D}\left[\mathrm{L}_s(\mathrm{h})=0\right] \leq (1-\epsilon)^m \tag{4}$$

Again, using definition,

$$\mathbb{P}_{S\sim D}\left[\mathrm{L}_s(\mathrm{h})=0\right] \leq (\mathrm{e})^{-\epsilon m} \tag{5}$$

Thus, we can infer that,

$$\Pr\left[\ni h \in H : L_{(D_m,f)}(h) > \epsilon \text{ and } L_S(h)=0\right] \leq |H|e^{-\epsilon m} \tag{6}$$

Question 4

Solution:

(a) We know that, an axis-aligned rectangle can shatter 4 points by using the axis-aligned algorithm. Now we need to prove that this rectangle can shatter points in 2- dimension and thus $\mathrm{VCdim}\left(H^{d}\right) = 2\mathrm{d}$.

Lets assume all 2d points in a plane are enclosed in a rectangle.

Two cases can arise:

Case 1: Let us take a set of 2d points where each point only has one of the d dimensions set to either 1 or -1 and 0 for all other dimensions. It is easy to see that any subset of these points can be shattered by an axis-aligned rectangle. Hence the VC-dim is atleast 2d. So we can write, $\mathrm{VCdim}\left(H^{d}\right) \leq 2\mathrm{d}$.

Case 2: Let us take a set of 2d+1 points. Now we find the minimum and maximum of value in each dimension for these set of points and then building a R^{d} rectangle with these bounds. Since there are 2d + 1 points, atleast one point must lie inside this rectangle. Now if we label this interior point as negative then there is no rectangle that can separate this labeling, we can show that $\mathrm{VCdim}\left(H^{d}\right) < 2\mathrm{d}+1$.

Thus combining these two cases from above , we see that $\mathrm{VCdim}\left(H^{d}\right) = 2\mathrm{d}$ when VCdim cannot shatter points more than 2d.

(b) We are given a hypothesis as :

$H = \{x- > *\sin\theta x : \theta \in \mathbb{R}\}$

By showing that the set of points can be shattered, we can prove that $\mathrm{VCdim}\left(H^{d}\right) = \infty$.

Let us assume to be the set of points as $\mathbf{P}\left\{x_1, x_2, \ldots .., x_p\right\}$

Consider the set of points as : $x_i = c^{-\imath}$

Labels are: $\{y_1, y_2, \ldots \ldots, y_p\}$

$\theta = \pi\left(1 + \sum_{i=1}^{p} \frac{(1-y_i)c^{i}}{2}\right)$

The hypothesis is represented as:

$$\begin{aligned} h\left(x_j\right) &= \sin\left(c^{-j} \times \pi\left(x_j\right)\right) \\ \Longrightarrow h\left(x_j\right) &= \sin\left(c^{-j} \times \pi\left(1 + \sum_{i=1}^{p} \frac{(1-y_i)c^{i}}{2}\right)\right) \\ \Longrightarrow h\left(x_j\right) &= \sin\left(c^{-j}\pi + \sum_{i=1}^{p}\left(1-y_i\right)c^{i-j}\frac{\pi}{2}\right) \end{aligned} \tag{7}$$

If the label y_i becomes 1, then the term which has the term of θ becomes 0.

thus,

$$\begin{aligned} \Longrightarrow h\left(x_j\right) &= \sin\left(c^{-j} \times \pi + \left(1-y_j\right)\frac{\pi}{2} + \sum_{i=1}^{p} 2 \times c^{i-j}\frac{\pi}{2}\right) \\ \Longrightarrow h\left(x_j\right) &= \sin\left(\left(1-y_j\right)\frac{\pi}{2} + c^{-j}\pi + \pi\sum_{i=1}^{p} c^{i-j}\right) \end{aligned} \tag{8}$$

Rearranging the above gives us,

$$\Longrightarrow h\left(x_j\right) = \sin\left(\pi\sum_{i=1}^{p} c^{i-j} + \left(1-y_j\right)\frac{\pi}{2} + c^{-j}\pi\right) \tag{9}$$

It can be easily inferred that the equation give negative value, $\sin(-\pi + x) = -\sin x$.

Thus, we can say that :

Case 1: if y_i = 1, then the RHS term would always be less than the second quadrant, which means $h(x) = 1$ as sin function will always be positive. Thus, it matches the answer value.

Case 2: if y_i = -1, then the RHS term would always be less than the second quadrant, which means $h(x) = -1$ as sin function will always be negative. Thus, again it matches the answer value.

So, we can infer from above that $\left\{c^{-1}, c^{-2}, \ldots, c^{-p}\right\}$ can be easily shattered.

$\therefore VCdim(H) = \infty$

Hence Proved.

Question 5

Solution:

We need to prove that, $\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = 1/2$

The distribution vector for $(t+1)^{th}$ iteration can be represented in terms of D in $(t)^{th}$ iteration as,

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{\sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_i h_t(x_j)}} \tag{10}$$

Now, the denominator can be broken down as,

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{\sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)} \mathbb{I}_{[y_i \neq h_t(x_t)]}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_i h_t(x_j)} \mathbb{I}_{[y_i = h_t(x_t)]} + \sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_i h_t(x_j)} \mathbb{I}_{[y_i \neq h_t(x_t)]}} \tag{11}$$

Now, we know that $y_i h_t\left(x_i\right) = -1$, if $y_i \neq h_t\left(x_i\right)$ and $y_i h_t\left(x_i\right) = 1$, if $y_i = h_t\left(x_i\right)$

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{\sum_{i=1}^{m} D_i^{(t)} e^{w_t} \mathbb{I}_{[y_i \neq h_t(x_t)]}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t} \mathbb{I}_{[y_i = h_t(x_t)]} + \sum_{j=1}^{m} D_j^{(t)} e^{w_t} \mathbb{I}_{[y_i \neq h_t(x_t)]}} \tag{12}$$

Also, substituting error ϵ_t computed as, $\epsilon_t = \sum_{i=1}^{m} D_i^{(t)} \mathbb{I}_{[y_i \neq h_t(x_t)]}$, we get,

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{e^{w_t}}{e^{-w_t}\left(1-\epsilon_t\right) + e^{w_t}\epsilon_t} \tag{13}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{e^{w_t}\epsilon_t}{e^{-w_t}\left(1-\epsilon_t\right) + e^{w_t}\epsilon_t} \tag{14}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{e^{2w_t}\epsilon_t}{\left(1-\epsilon_t\right) + e^{2w_t\epsilon_t}} \tag{15}$$

Also, we can say $e^{2w_t} = \frac{1-\epsilon_t}{\epsilon_t}$ Thus, substituting,

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{\frac{1-\epsilon_t}{\epsilon_t}\epsilon_t}{\left(1-\epsilon_t\right) + \frac{1-\epsilon_t}{\epsilon_t}\epsilon_t} \sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[y_i \neq h_t(x_t)]} = \frac{1-\epsilon_t}{2\left(1-\epsilon_t\right)} = 1/2 \tag{16}$$

Hence, proved.

Question 6

Solution:

The hypothesis given is a convex set and $l(w,(x,y)) = \log[1+\exp(-y\langle w,x\rangle)]$ can be re-written as, $l(u) = \log[1+\exp(u)]$

Also,

$$\begin{aligned} l'(u) = \tfrac{1}{1+\exp(u)} \cdot \exp(u) \implies l'(u) = \tfrac{1}{1+\exp(-u)} \leq 1(** \text{ when } u = \infty) \longrightarrow \text{(I)} \\ l''(u) = \tfrac{\exp(-u)}{(1+\exp(-u))^2} \implies l''(u) = \tfrac{1}{(1+\exp(-u))(1+\exp(u))} \leq \tfrac{1}{4}(** \text{ when u} = 0) \longrightarrow \text{(II)} \end{aligned} \tag{17}$$

Also, $l''(u) \geq 0$ always, which means that the function is convex by non-negative double derivative property.

From (I), we can say that $l(u)$ is 1-Lipschitz. By the definition of Convex-Lipschitz-Bounded Learning Problem, $\rho = \mathrm{B}$ and B=B from the problem, i.e; boundedness, we can say that problem is B-convex-Lipschitz bounded with parameter B, B. (where B is a positive scalar as mentioned)

From (II), we can say that $l(u)$ is $\frac{1}{4}$ -Smooth bounded, as the function is non negative. By the definition of Convex-Smooth-Bounded Learning Problem, $\rho = \mathrm{B}^2$ and B=B from the problem, i.e; boundedness, we can say that the problem is convex-Smooth bounded with parameters $\frac{\mathrm{B}^2}{4}$ and B.

Thus the we can say l is B-Lipschitz and $\frac{B^2}{4}$ smooth.

Question 7

Solution:

As Claim 12.5 in the textbook states that the maximum function of convex functions is also convex we can say the given loss function is also convex since the loss function $l(w,(x,y)) = \max\{0, 1-y\langle w,x\rangle\}$ can be considered as maximum of two convex functions because $1-y\langle w,x\rangle$, can be considered as convex since it is a composition of 1-f(x) into a linear convex function.

Now, to prove that the loss function is R-Lipschitz, we need to show that, $|l_1 - l_2| \leq R\,\|w_1 - w_2\|$ where l_1 and l_2 are the loss functions for weights w_1 and w_2 respectively.

Now, to prove the above let us consider these:

CASE I: Both $y\,\langle w_1, x\rangle \geq 1$ and $y\,\langle w_2, x\rangle \geq 1$

Then, $l_1 - l_2 = 0$
$\implies |l_1 - l_2| = 0$
$\implies |l_1 - l_2| \leq R\,\|w_1 - w_2\|$

CASE II: $y\,\langle w_1, x\rangle < 1$ and $y\,\langle w_2, x\rangle \geq 1$

Then, $l_1 - l_2 = 1 - y\,\langle w_1, x\rangle - 1$
$\implies |l_1 - l_2| = y\langle w, x\rangle$
$\implies |l_1 - l_2| \leq \|w_1\|\,\|x\|$
$\implies |l_1 - l_2| \leq R\,\|w_1 - w_2\|$

CASE III: Both $y\,\langle w_1, x\rangle < 1$ and $y\,\langle w_2, x\rangle < 1$

Then, $l_1 - l_2 = 1 - y\,\langle w_1, x\rangle - (1 - y\,\langle w_2, x\rangle)$
$\implies |l_1 - l_2| = y\langle w_2 - w_1, x\rangle$
$\implies |l_1 - l_2| \leq \|w_1 - w_2\|\,\|x\|$
$\implies |l_1 - l_2| \leq R\,\|w_1 - w_2\|$

Thus we can say that the loss function is convex and R-Lipschitz.

Question 8

Solution:

It is given that the hypothesis h(x) = 1 on odd number of ones in training labels and h(x) = 0 when the number of ones in training labels is even. Lets assume a data sample S and the data generation process is uniformly distributed with P[y=1]=P[y=0]= 0.5 and hence the true error will be = 1/2.

Thus, two cases arise:

Case I: When S has even number of 1s as training labels then h(x)= 0. But the leave one out estimate will be y=1. Because, if we take 1 as the leave one estimate, then prediction is 1 for training set because in training set number of ones will be odd. But the prediction for leave one out label is 0, which leads to leave one out estimate being 1. Similarly, if 0 is picked same will be the case for leave out one estimate.

Case II: When S has odd number of 1s as training labels then h(x) =1. The leave one out estimate will again be 1. This is because, if we take 1 as the leave one out label, the training set will predict h(x) = 0 and if we take 0 as the leave one out label, the training set will predict h(x) = 1, because of which the prediction will always come out as wrong. Thus, leave one out estimate will be 1.

Since the true error is 1/2 and leave one out estimate is 1, we can conclude their difference will always be 1/2.

Question 9

Solution:

Lets take a sample $S = (x, y) = (e_1, 1)$, where e_1 is the standard unit vector ([1,0,0,...]). Also, let $w = -e_1$. Then clearly, $\langle w, x\rangle$ = -1 (as their inner product will be the sum of their element wise products where the first term would be -1 and all others 0)

Now, $L_S(w) = 1$ since it is a 0-1 Loss Function.

We need to prove $L_S(w) \leq L_S\left(w'\right), \forall w'$ to show w is a local minimum;

Let us take $\epsilon \in (0, 1)$ and make sure $\epsilon > 0$

For every w' having $\|w' - w\| \leq \epsilon$, we can write,

$\langle w', x\rangle = -1 - \langle w' - w, x\rangle$ (By Cauchy Schwartz inequality)
$\implies \langle w', x\rangle \leq -1 + \|w' - w\|_2 \, \|x\|_2$
$\implies \langle w', x\rangle < -1 - (-1)$
$\implies \langle w', x\rangle < 0$

This implies that, $L_S\left(w'\right) = 1$ and makes w a local minimum since the above stated inequality becomes true.

Again, w cannot be a global minimum because the lowest value of this function is 0 which means there exists a w^* such that $L_S\left(w^*\right) = 0$.

So we can say, $L_S\left(w^*\right) < L_S\left(w'\right)$

References

http://www.stat.cmu.edu/ larry/=sml2008/hw5_solution.pdf
http://www.stat.cmu.edu/Alon Gonen_Dana Rubinstein