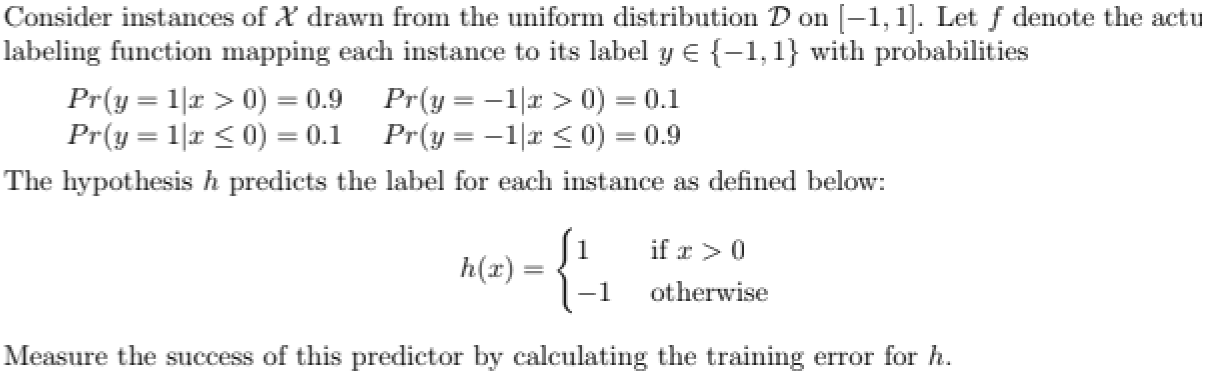
Question 1: A “warm up” problem



Consider instances of $\mathcal{X}$ drawn from the uniform distribution $\mathcal{D}$ on $[-1,1]$. Let f denote the actu labeling function mapping each instance to its label $y \in \{-1,1\}$ with probabilities

$$Pr(y=1|x>0)=0.9 \quad Pr(y=-1|x>0)=0.1$$
$$Pr(y=1|x\leq 0)=0.1 \quad Pr(y=-1|x\leq 0)=0.9$$

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

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

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

SOLUTION:

$h(x) = 1 \ if \ (x \leq 0)$ and $h(x) = 0 \ if \ (x > 0)$ are possible misclassifications

Training Error :

$$Training\ Error\ = \frac{\#\ of\ Wrong\ Predictions}{\#\ of\ Total\ Predictions}$$

$$= \frac{[i\epsilon[m]:h(x_i)\neq y_i]}{m}$$

$$= \frac{P(y=1|x\leq 0)+P(y=-1|x>0)}{P(y=1|x>0)+P(y=-1|x>0)+P(y=1|x\leq 0)+P(y=-1|x\leq 0)}$$

$$= \frac{0.1+0.1}{0.9+0.1+0.1+0.9}$$

$$= 0.1$$

Accuarcy = 1- Error = 1-0.1 = 0.9

Question 2: 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 any classifier $g\colon \mathrm{X} \rightarrow (0{,}1)$

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

Solution:

Bayes Optimal Predictor:

$$\begin{aligned} f_D(x) &= 1 \quad if \;\; P\left([y = 1\big|x] \geq \frac{1}{2}\right) \\ &= 0 \qquad\qquad otherwise \end{aligned}$$

Error is given by, $L_D(f_D)$= $P(f_D(x) \neq y|x)$

Let $P(y = 1|x) = t$

$$\begin{aligned} L_D(f_D) = \; P(f_D(x) \neq y|x) &= P(y = 0|x) \cdot I\left(t \geq \frac{1}{2}\right) + P(y = 1|x) \cdot I\left(t \leq \frac{1}{2}\right) \\ &= (1 - t) \cdot I\left(t \geq \tfrac{1}{2}\right) + t \cdot I\left(t \leq \tfrac{1}{2}\right) \\ &= min(t, 1 - t). \end{aligned}$$

$L_D(g)$ for the classifier $g\colon X \rightarrow (0{,}1)$

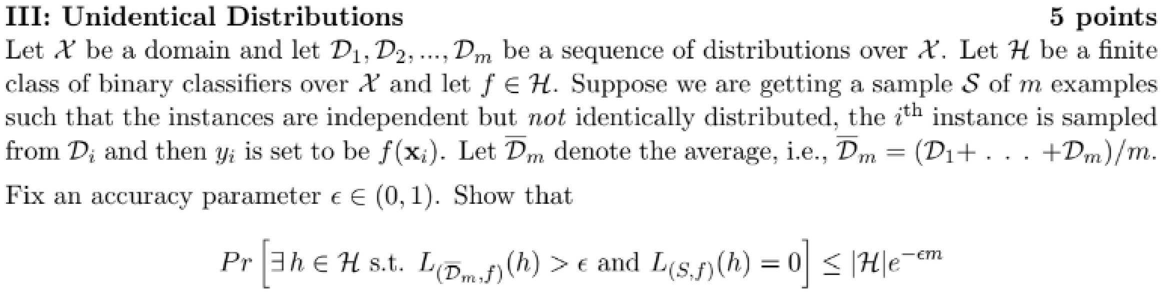
$$\begin{aligned} \mathrm{P}((\mathrm{g}(\mathrm{x}) \neq \; \mathrm{y}\big|\,\mathrm{x})) = \; \mathrm{P}(\mathrm{g}(\mathrm{x}) = 0|x) \cdot \mathrm{P}(\mathrm{y} = 1\,\big|\; \mathrm{x}) \; + \; \mathrm{P}(\mathrm{g}(\mathrm{x}) = 1|x) \cdot \mathrm{P}(\mathrm{y} = 0\,\big|\; \mathrm{x}) \\ = \; P(g(x) = 0|x) \cdot t \; + \; P(g(x) = 1|x) \cdot (1 - t) \end{aligned}$$

$$\begin{aligned} & P(g(x) = 0|x) \cdot t \; + \; P(g(x) = 1|x) \cdot (1 - t) \\ &\qquad \geq P\big(g(x = 0|x)\big) \cdot min(t, 1 - t) + P\big(g(x = 1|x)\big) \cdot min(t, 1 - t) \\ &\qquad \geq min(t, 1 - t) \cdot 1 \\ &\qquad \geq min(t, 1 - t) \\ &\qquad \geq L_D(f_D) \end{aligned}$$

Therefore,

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

Question 3: Unidentical Distributions



III: Unidentical Distributions **5 points**

Let $\mathcal{X}$ be a domain and let $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in \mathcal{H}$. Suppose we are getting a sample $\mathcal{S}$ of m examples such that the instances are independent but *not* identically distributed, the i^{th} instance is sampled from $\mathcal{D}_i$ and then y_i is set to be $f(\mathbf{x}_i)$. Let $\overline{\mathcal{D}}_m$ denote the average, i.e., $\overline{\mathcal{D}}_m = (\mathcal{D}_1 + \ . \ . \ . \ + \mathcal{D}_m)/m$.

Fix an accuracy parameter $\epsilon \in (0,1)$. Show that

$$Pr\left[\exists\, h \in \mathcal{H} \text{ s.t. } L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right] \le |\mathcal{H}|e^{-\epsilon m}$$

SOLUTION:

Let h ϵ H, such that

$$L_{(\overline{D_m}:f)}(h) > \epsilon$$

Where, $D_1, D_2, \cdots, D_m$ are sequence of distributions over X and D_M is the average of sequence. $D_1, D_2, \cdots, D_m$ are independent.

S is sample of m points with each point drawn D_i such that Loss on sample S is 0. ($L_S(h) = 0$)

The Probability of predictions over these distributions is given as:

$$\frac{P_{X\sim D_1}[h(X) = f(X)] + \cdots + P_{X\sim D_m}[h(X) = f(X)]}{\mathrm{m}}$$

Training error is given by,

$$\frac{1}{m}\sum_{i=1}^{m} P_{X\sim Di}[h(X) = f(X)] < 1 - \epsilon$$

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

Now ,

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

$$P\left[h(X) = f\left(X_{D_1}\right)\right] * P\left[h(X) = f\left(X_{D_2}\right)\right] \cdots P\left[h(X) = f\left(X_{D_m}\right)\right]$$

$$= \left(\left(\prod_{i=1}^{m} P_{S\sim Di}[h(X) = f(X)]\right)^{\frac{1}{m}}\right)^{m}$$

We Know that, $(a+b)/2 \ge \sqrt{(a,b)}$

$$\le \left(\frac{(\sum_{i=1}^{m} P_{X\sim Di}[h(X) = f(X)])}{m}\right)^{m}$$

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

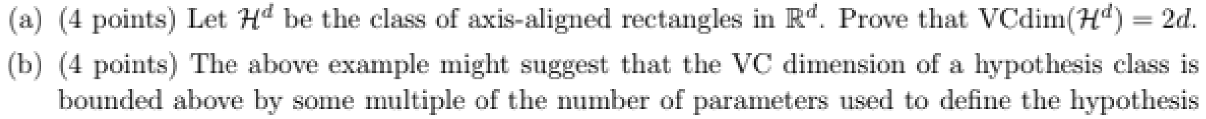
$$\le\ e^{-\epsilon m}$$

As we know, $P(AUB) \le P(A) + P(B)$. We can say that the probability exists where $h\ \mathcal{E}\ H$ such that $L_{(\overline{D_m}:f)}(h) >\ \epsilon$ and $P_{S\sim\prod_{i=1}^{m} Di}[L_S(h)] = 0\ \le\ |H|e^{-\epsilon m}$.

Therefore,

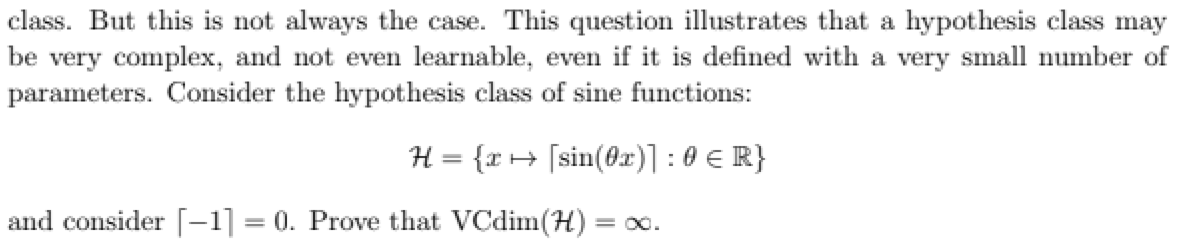
$$P\big[\exists\ \ h\ \varepsilon\ H\ s.t.\ L_{\overline{D_m},f}(h) > \epsilon\ and\ L_{(f)}(h=0)\big]\ \le\ |H|e^{-\epsilon m}$$

Question 4: Vapnik-Chervonenkis (VC) Dimension



(a) (4 points) Let $\mathcal{H}^d$ be the class of axis-aligned rectangles in $\mathbb{R}^d$. Prove that $\text{VCdim}(\mathcal{H}^d) = 2d$.

(b) (4 points) The above example might suggest that the VC 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:

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

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

(a)

Let $h\ \varepsilon\ H,$ given x; h maps x to 1 if x lies inside the rectangle else it maps to 0.

$$k(x) = 1\ \ if\ x\ lies\ inside\ rectangle$$

$$k(x) = 0\ otherwise$$

When the points are on axes they will be covered by the rectangle. The VC Dimension will be at least d=4. But if we take 5 points such the one point lies on the origin and remaining 4 on each side of axes. When the point on origin becomes mid point of two points no axis aligned rectangle can shatter these three points considering the origin point has opposite label to points on axes. Hence the VC dimension will be 4. As H-class shatters 4 points but not 5.

Consider $a_1 \le b1, a_2 \le b2, \cdots, a_d \le b_d$, and $h_{(a_1,b_1,\cdots a_d,b_d)}(x_1,\cdots x_d) = \prod_{i=1}^{d} I_{[x_i \varepsilon|(a_i,b_i)]}$

$h_{(a_1,b_1,\cdots a_d,b_d)}(x_1,\cdots x_d)\ \forall\ \ i\ \varepsilon\ [d]$ and $a_i \le b_i = H^d$

Consider $(x_1,\cdots x_{2d})$ where $x_i = e_i\ if\ i\ \varepsilon\ d$ or $x_i = e_{i-d}\ if\ i > d$

Since this will be shattered, therefore, the VC dimension is at least $2d$.

Considering example $a_i = -4$ if $y_{i+d} = 1$ and $a_i = 0$ otherwise.

$b_i = 4$ if $y_i = 1$ and $b_i = 0$ otherwise.

Then $h_{(a_1,b_1,\cdots a_d,b_d)}(x_i) = y_i$ for $i \, \varepsilon \, 2d$

If K is a set of size $(2d+1)$, as per pigeonhole principle, x ε K such that,

For every j in |d| and x' exists in K when $x'_j \leq x_j$. Now the points are 2d+1 then minimum 1 point must lie inside the rectangle. And if we label this point as negative then this labelling cannot be separated by any rectangle. This proves $VC\,dim\left(H^d\right) < 2d+1$.

If we consider positive labelling points, we can say that, $VC\,dim\left(H^d\right) \geq 2d$

Therefore, it is proved that

$$VC\,dim\left(H^d\right) = 2d$$

(b)

It is given that $H = x \mapsto [sin(\Theta_x)] : \Theta \, \varepsilon \, R$ where $\Theta = 2^{\mathrm{m}} \pi$

$$Sin(2^m \pi) = sin\left(2^m \pi(0, x_1 x_2 \cdots)\right)$$

Upon shifting 2^m shifting

$$= sin\left(2^m \Pi(x_1 x_2 \cdots x_{m-1} x_m x_{m+1} \cdots)\right)$$

$$= sin\left(2^m \Pi(x_1 x_2 \cdots x_{m-1} x_m x_{m+1} \cdots) - 2\Pi(x_1 x_2 \cdots x_{m-1}.0)\right)$$

$$\because Sin(2n\Pi) = Sin(2\Pi)$$

$$= sin\left(2^m \Pi(0.x_m x_{m+1} \cdots)\right).$$

When x_m= 0, then the above equation ε $(0,\Pi)$ i.e, positive and $[Sin(2^m \Pi x)] = 1$.

When x_m= 1, then the above equation $\epsilon(\Pi, 2\Pi)$ i.e, negative and $[Sin(2^m \Pi x)] = 0$

We can see that $2^m(0.x_m x_{m+1})\varepsilon[\Pi, 2\Pi]$

Therefore, $[Sin(2^m \Pi x)] = 1 - x_m$

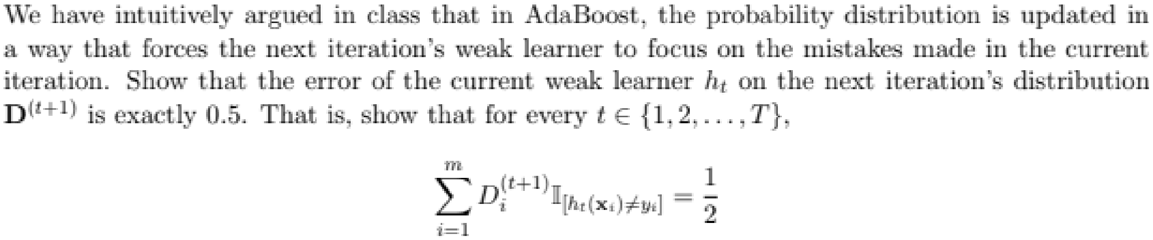
$x_1, x_2, \cdots, x_n$ ε $[0,1]$ be points that can be shattered by H and m ranges from $1 \; to \; 2^n$ and over all the values of $x_1, \ldots x_n$

Now to set label 1 for all the instances $h(x) = [sin(2^1 x)]$ returns first bit column in binary expansion.

Similarly, $h(x) = [sin(2^2 x)]$ returns second bit column in binary expansion.

If we continue the same till n and it can be shattered. And conclude that $VCdim(H) = \infty$

Question 5: Boosting



We have intuitively argued 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 $\mathbf{D}^{(t+1)}$ is exactly 0.5. That is, show that for every $t \in \{1, 2, \dots, T\}$,

$$\sum_{i=1}^{m} D_i^{(t+1)} \mathbb{I}_{[h_t(\mathbf{x}_i) \neq y_i]} = \frac{1}{2}$$

SOLUTION:

we know that,

$$\epsilon_t = \sum_{i-1}^{m} D_i^t I_{(y_i)\neq[h_t](x_i)}$$

So, the error on the iteration's will be and the error of the weak learner h_t on the next iteration's

$$= \sum_{i-1}^{m} D_i^{t+1} I_{(y_i)\neq[h_t](x_i)}$$

$$\omega_t = \frac{1}{2} log\left(\frac{1}{\epsilon_t} - 1\right)$$

$$e^{\omega_t} = \sqrt{\left(\frac{1}{\epsilon_t} - 1\right)}$$

Also,$\sum_{i-1}^{m} D_i^t\, e^{-y_i\omega_i h_i(x_i)} I_{(y_i)\neq[h_t](x_i)} = \epsilon_t \sqrt{\left(\frac{1}{\epsilon_t} - 1\right)}$

Also,$\sum_{j}^{m} D_j^t\, e^{-y_j\omega_j h_j(x_j)} = \epsilon_t \sqrt{\left(\frac{1}{\epsilon_t} - 1\right)} + (1 - \epsilon_t)\sqrt{\left(\frac{\epsilon_t}{1-\epsilon_t}\right)}$

So, now dividing the above two equations:

$$Error = \frac{\sum_{i-1}^{m} D_i^t\, e^{-y_i\omega_i h_i(x_i)} I_{(y_i)\neq[h_t](x_i)}}{\sum_{j}^{m} D_j^t\, e^{-y_j\omega_j h_j(x_j)}}$$

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

$$= \frac{1}{2}$$

Therefore, $\sum_{i-1}^{m} D_i^{t+1} I_{(y_i)\neq[h_t](x_i)} = \frac{1}{2}$

Question 6: Learnability of logistic regression

Consider a function, g(a) = log (1 + e^{a}).
the function is convex as there is only one minima.

We know that if $f(w) = g(\langle w, x\rangle + y)$, then if g is convex then f is convex.

$$\begin{aligned} f(\alpha w_1 + (1-\alpha)w_2) &= g(\langle \alpha w_1 + (1-\alpha)w_2, x\rangle + y) \\ &= g(\alpha\langle w_1, x\rangle + (1-\alpha)\langle w_2, x\rangle + y) \\ &= g\big(\alpha(\langle w_1, x\rangle + y) + (1-\alpha)(\langle w_2, x\rangle + y)\big) \\ &\leq \alpha g\ (\langle w_1, x\rangle + y) + (1-\alpha) g(\langle w_2, x\rangle + y) \end{aligned}$$

We know that the function is learnable if it is convex-smooth or convex-lipschitzness with bound on parameters and domain.

We will prove g is learnable by proving these

- Lipschitzness
- Smoothness
- Boundedness

Lipschitzness:

$$|g'(a)| = \frac{\mathrm{e^a}}{1 + \mathrm{e^a}}$$
$$= \frac{1}{\mathrm{e^{-a}} + 1} \leq 1.$$

Here slope for any ‘a’ is <=1.
Therefore, It is 1- Lipschitz.

Consider $f(x) = g_1(g_2(x))$ where g_1 is ρ_1-Lipschitz and $g_2$is ρ_2-Lipschitz.
Then we know that f is ($\rho_1\rho_2$)-Lipschitz. If g_2 is linear function $g_2 = \langle v, x\rangle + b$ for some $v\ \varepsilon\ R^d, b\ \varepsilon\ R$ then f is ($\rho_1|v|$) - Lipschitz

$$\begin{aligned} |f(w_1) - f(w_2)| &= |g_1(g_2(w)) - g_1(g_2(w2))| \\ &\leq \rho_1 |g_2(w1) - g_2(w_2)| \\ &\leq \rho_1\rho_2 |w_1 - w_2| \end{aligned}$$

Therefore, the function is B-Lipschitz.

Smoothness -

We show that g(a) is 1/4-smooth.

$$g''(a) = \frac{\mathrm{e^{-a}}}{(e^{-a} + 1)^2}$$
$$= \left((e^a)(e^{-a} + 1)^2\right)^{-1}$$
$$= \frac{1}{2 + e^a + e^{-a}} \leq 1/4$$

We know that for $f(w) = g(\langle w, x\rangle + b)$ where g is a β-smooth function , then f is (β$|x|^2$)-smooth
Using the mean value theorem, we conclude that g' is 1/4 Lipschitz. And thus, g is $B^2/4$ – smooth.

Parameters for Convex-smooth-bounded is $\beta^2/4$ and Convex-Lipschitz-bounded is B.

Question 7: Learnability of Halfspaces with hinge loss

We consider $w_1, w_2 \; \varepsilon \; R^d$, and $(x, y) \; \varepsilon \; \{x \; \varepsilon R^d : |x'|_2 \leq R \} \times \{-1,1\}$
Let $l_i = max\{0,1 - y\langle w_i, x\rangle\}$ for $i \; \varepsilon \; [2]$.

To prove that $|l_1 - l_2| \leq R|w_1 - w2|_2$, considering $y\langle w_1, x\rangle \geq 1 \; and \; y\langle w_2, x\rangle \geq 1$
we have $|l_1 - l_2| = 0 \leq R|w_1 - w2|_2$
Assume that $|\{ i: y\langle w\_i, x \rangle < 1\}| \geq 1$
w.l.o.g , $1 - y\langle w_1, x\rangle \geq 1 - y\langle w_2, x\rangle$

$$\begin{aligned} |l_1 - l_2| &= l_1 - l_2 \\ &= 1 - y\langle w_1, x\rangle - max\langle 0,1 - y\langle w_2, x\rangle \\ &\leq 1 - y\langle w_1, x\rangle - (1 - y\langle w_2, x\rangle) \\ &= y\langle w_2 - w_1 . x\rangle \\ &\leq ||w_1 - w_2|||x| \\ &\leq R||w_1 - w2|| \end{aligned}$$

Question 8: Cross – validation

Given the labels (0,1) are chosen randomly with probability 0.5

On the Distribution D the labeling function is going to assign labels independent of input with probability 0.5. Therefore in total half the points will have 0 and half the points will have 1 as label.
Our output Hypothesis h is going to assign 0 to all inputs if the number of 1s in the training data are even. Else it assigns 1.

$$L_D(h) = 1/2$$

For parity of S =1 and fold set(x,y) in S,
When the parity of S/x is 1, then y = 0. The output of h(x) of training using S/x will be 1 and the Leave on out using this fold is 1.

when the parity of S/x is 0, then y = 1. The constant prediction output h(x) of training using S/x will be 0 and the Leave on out using this fold is again 1.

The estimate of error of h is 1 when we calculate the average over folds.

Similarly, for parity of S=0, the estimate of error of h averaging over folds is 1.

the difference between error and estimate is:

$$1 - 1/2 = 1/2$$

Question 9: Local minimum

Let H be the Hypothesis class of homogenous halfspaces in R_d
Let $x = e_1, y = 1$, sample $S = \{(x, y)\}$, $w = -e_1$
Then $\langle w, x\rangle = -1$ and $L_s(w) = 1$
Still, w is a local minima. Let ϵ ε $(0,1)$.
By Cauchy–Schwarz inequality of inner products for all vectors, for every w’ with $|w' - w| \leq \epsilon$

$$\begin{aligned} < w', x > &= < w, x \succ< w' - w, x > \\ &= -1 \prec w' - w, x > \\ &\leq -1 + |w' - w|_2 |x|_2 \\ &< -1 + 1 \\ &= 0 \end{aligned}$$

Therefore, we get $L_s(w') = 1$