MLT : Week 9

Perceptron & Logistic Regression

Vivek Sivaramakrishnan

The weight vector for two consecutive iterations of a perceptron model are given below:

The weight vector for two consecutive iterations of a perceptron model are given below:

$$\mathbf{w}^{t} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}, \ \mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

The weight vector for two consecutive iterations of a perceptron model are given below:

$$\mathbf{w}^t = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}, \; \mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

Find all possible datapoints (along with its true label) that could've been used for the update from $\mathbf{w}^t$ to $\mathbf{w}^{t+1}$.

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + y_i \mathbf{x}_i$$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + y_i \mathbf{x}_i$$

Substituting $\mathbf{w}^{t} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + y_i \mathbf{x}_i$$

Substituting $\mathbf{w}^{t} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i \mathbf{x}_i$$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + y_i\mathbf{x}_i$$

Substituting $\mathbf{w}^t = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i\mathbf{x}_i$$

$$\implies y_i\mathbf{x}_i = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + y_i \mathbf{x}_i$$

Substituting $\mathbf{w}^{t} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i \mathbf{x}_i$$

$$\implies y_i \mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + y_i \mathbf{x}_i$$

Substituting $\mathbf{w}^t = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i \mathbf{x}_i$$

$$\implies y_i \mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Case 1: $y_i = 1$

Case 2: $y_i = -1$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + y_i\mathbf{x}_i$$

Substituting $\mathbf{w}^t = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i\mathbf{x}_i$$

$$\implies y_i\mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Case 1: $y_i = 1$

$$\implies \mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Case 2: $y_i = -1$

Recall the update rule for the Perceptron Learning Algorithm:

if mistake:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + y_i\mathbf{x}_i$$

Substituting $\mathbf{w}^t = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}$, $\mathbf{w}^{t+1} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$ from the question, we get:

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} + y_i\mathbf{x}_i$$

$$\implies y_i\mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Case 1: $y_i = 1$

$$\implies \mathbf{x}_i = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Case 2: $y_i = -1$

$$\implies \mathbf{x}_i = -\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$$

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| \;=\; 1,\; y_i\left(\mathbf{w}^t\mathbf{x}_i\right) \geqslant 0.2 \;\forall i$$

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| \;=\; 1,\; y_i\left(\mathbf{w}^t\mathbf{x}_i\right) \geqslant 0.2 \;\forall i$$

Also assume that $||\mathbf{x}||_2 \leqslant 1 \;\forall i$.

Consider a linearly seperable binary classification data set with 1000 data points and 100 features. Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| \;=\; 1,\; y_i\left(\mathbf{w}^t\mathbf{x}_i\right) \geqslant 0.2 \;\forall i$$

Also assume that $||\mathbf{x}||_2 \leqslant 1 \;\forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Consider a linearly seperable binary classification data set with 1000 data points and 100 features. Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| = 1,\ y_i\left(\mathbf{w}^t\mathbf{x}_i\right) \geqslant 0.2\ \forall i$$

Margin γ

Also assume that $||\mathbf{x}||_2 \leqslant 1\ \forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Consider a linearly seperable binary classification data set with 1000 data points and 100 features. Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| = 1, \ y_i\left(\mathbf{w}^t \mathbf{x}_i\right) \geqslant \underbrace{0.2}_{\text{Margin } \gamma} \ \forall i$$

Also assume that $||\mathbf{x}||_2 \leqslant \underbrace{1}_{\text{Radius } R} \ \forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| \;=\; 1,\; y_i\big(\mathbf{w}^t\mathbf{x}_i\big) \geqslant \underbrace{0.2}_{\text{Margin } \gamma} \;\forall i$$

Also assume that $||\mathbf{x}||_2 \leqslant \underbrace{1}_{\text{Radius } R} \;\forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Recall the Radius-Margin Bound:

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| \;=\; 1,\; y_i\Big(\mathbf{w}^t\mathbf{x}_i\Big) \geqslant 0.2 \;\forall i$$

Margin γ

Also assume that $||\mathbf{x}||_2 \leqslant 1 \;\forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Radius R

Recall the Radius-Margin Bound:

$$\#\text{mistakes} \;\leqslant\; \frac{R^2}{\gamma^2}$$

Consider a linearly seperable binary classification data set with 1000 data points and 100 features. Assume that there exists a $\mathbf{w}$ such that

$$||\mathbf{w}|| = 1,\ y_i\left(\mathbf{w}^t\mathbf{x}_i\right) \geqslant 0.2\ \forall i$$

Margin γ

Also assume that $||\mathbf{x}||_2 \leqslant 1\ \forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Radius R

Recall the Radius-Margin Bound:

$$\begin{aligned}\#\text{mistakes} &\leqslant \frac{R^2}{\gamma^2} \\ &\leqslant \frac{1^2}{0.2^2}\end{aligned}$$

Consider a linearly seperable binary classification data set with 1000 data points and 100 features.

Assume that there exists a $\mathbf{w}$ such that

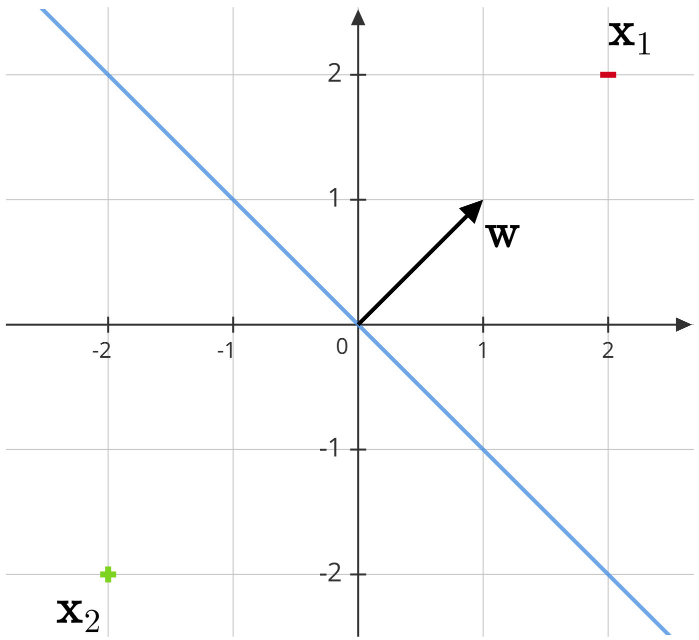
$$||\mathbf{w}|| \;=\; 1,\; y_i\Big(\mathbf{w}^t\mathbf{x}_i\Big) \geqslant \underbrace{0.2}_{\text{Margin } \gamma}\ \forall i$$

Also assume that $||\mathbf{x}||_2 \leqslant \underbrace{1}_{\text{Radius } R}\ \forall i$. What is the maximum number of mistakes that Perceptron algorithm can make on this dataset?

Recall the Radius-Margin Bound:

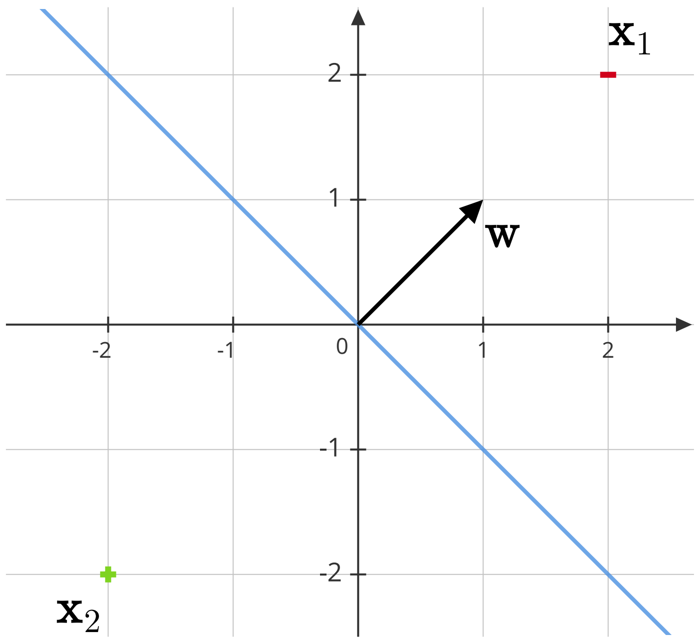
$$\begin{aligned}\#\text{mistakes} \;&\leqslant\; \frac{R^2}{\gamma^2}\\ &\leqslant\; 25\end{aligned}$$

Consider a logistic regression model that has been trained for a binary dassification problem on a dataset in $\mathbb{R}^2$. two data-points from the training set: $(\mathbf{x}_1,\ y_1)$ and $(\mathbf{x}_2,\ y_2)$. The models' weight vector and these two data-points (along with their labels) are drawn in the diagram given below. The symbol $+$ corresponds to label 1 and — corresponds to label 0.



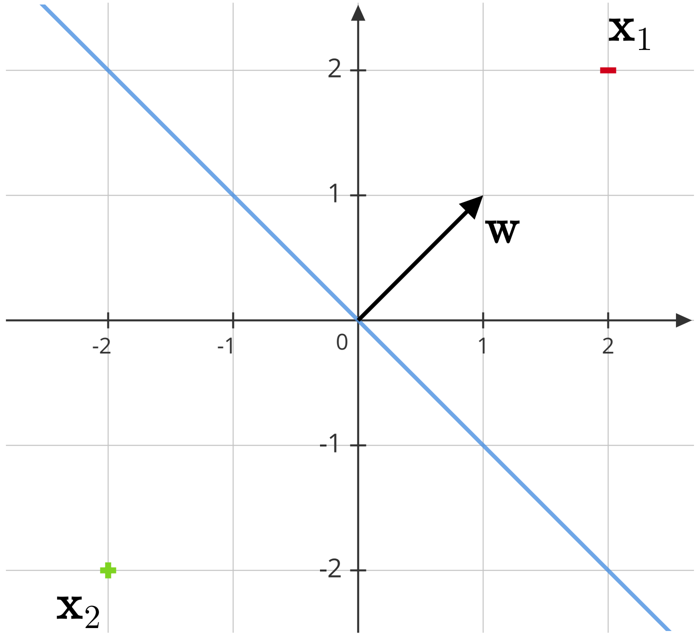
Consider a logistic regression model that has been trained for a binary dassification problem on a dataset in $\mathbb{R}^2$. two data-points from the training set: $(\mathbf{x}_1,\ y_1)$ and $(\mathbf{x}_2,\ y_2)$. The models' weight vector and these two data-points (along with their labels) are drawn in the diagram given below. The symbol $+$ corresponds to label 1 and — corresponds to label 0.

Let σ_1,σ_2 be the probabilities output by the model for the two data-points.



Consider a logistic regression model that has been trained for a binary dassification problem on a dataset in $\mathbb{R}^2$. two data-points from the training set: $(\mathbf{x}_1,\ y_1)$ and $(\mathbf{x}_2,\ y_2)$. The models' weight vector and these two data-points (along with their labels) are drawn in the diagram given below. The symbol $+$ corresponds to label 1 and — corresponds to label 0.

Let σ_1,σ_2 be the probabilities output by the model for the two data-points. Compute the following quantity: $\dfrac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|}$

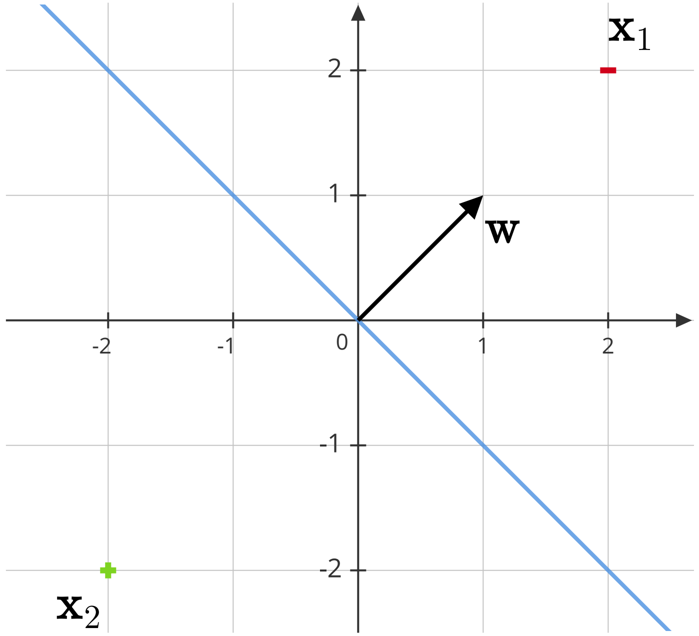


Consider a logistic regression model that has been trained for a binary dassification problem on a dataset in $\mathbb{R}^2$. two data-points from the training set: $(\mathbf{x}_1,\ y_1)$ and $(\mathbf{x}_2,\ y_2)$. The models' weight vector and these two data-points (along with their labels) are drawn in the diagram given below. The symbol $+$ corresponds to label 1 and — corresponds to label 0.

Let σ_1,σ_2 be the probabilities output by the model for the two data-points. Compute the following quantity: $\dfrac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|}$

Recall that the $\sigma_1 \ = \ \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_1\right) \ = \ \dfrac{1}{1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}}$

(sigmoid function)



It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \;=\; 1 - \sigma(-a)$$

It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \;=\; 1 - \sigma(-a)$$

Proof is straightforward; just evaluate the right hand side:

$$1 - \sigma(-a) = 1 - \frac{1}{1 + e^{-(-a)}}$$

It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \ = \ 1 - \sigma(-a)$$

Proof is straightforward; just evaluate the right hand side:

$$\begin{aligned} 1 - \sigma(-a) &= 1 - \frac{1}{1 + e^{-(-a)}} \\ &= \ 1 - \frac{1}{1 + e^{a}} \end{aligned}$$

It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \;=\; 1 - \sigma(-a)$$

Proof is straightforward; just evaluate the right hand side:

$$\begin{aligned}
1 - \sigma(-a) &= 1 - \frac{1}{1 + e^{-(-a)}} \\
&= 1 - \frac{1}{1 + e^{a}} \\
&= \frac{1 + e^{a}}{1 + e^{a}} - \frac{1}{1 + e^{a}}
\end{aligned}$$

It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \;=\; 1 - \sigma(-a)$$

Proof is straightforward; just evaluate the right hand side:

$$\begin{aligned} 1 - \sigma(-a) &= 1 - \frac{1}{1 + e^{-(-a)}} \\ &= 1 - \frac{1}{1 + e^{a}} \\ &= \frac{1 + e^{a}}{1 + e^{a}} - \frac{1}{1 + e^{a}} \\ &= \frac{e^{a}}{1 + e^{a}} \end{aligned}$$

It's worth to prove the following result before proceeding with the question:

$$\sigma(a) \; = \; 1 - \sigma(-a)$$

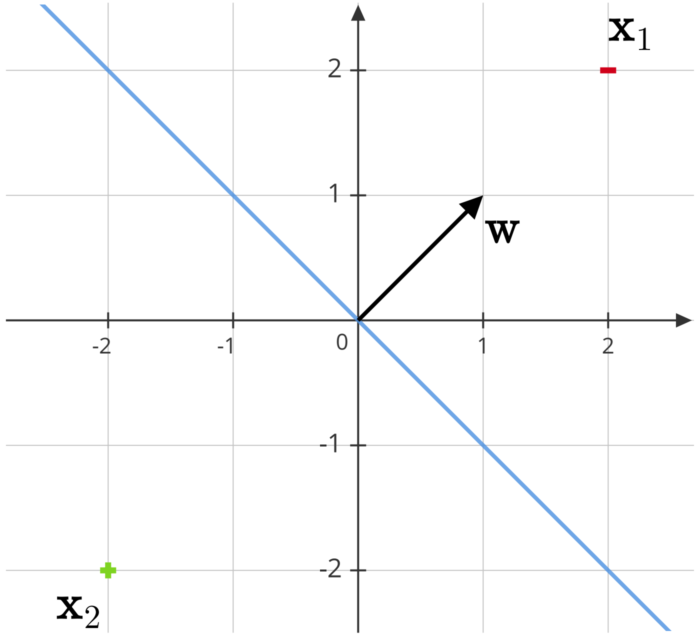
Proof is straightforward; just evaluate the right hand side:

$$\begin{aligned}
1 - \sigma(-a) &= 1 - \frac{1}{1 + e^{-(-a)}} \\
&= \; 1 - \frac{1}{1 + e^{a}} \\
&= \; \frac{1 + e^{a}}{1 + e^{a}} - \frac{1}{1 + e^{a}} \\
&= \frac{e^{a}}{1 + e^{a}} = \frac{1}{1 + e^{-a}} \; = \; \sigma(a)
\end{aligned}$$

We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 \ =$$

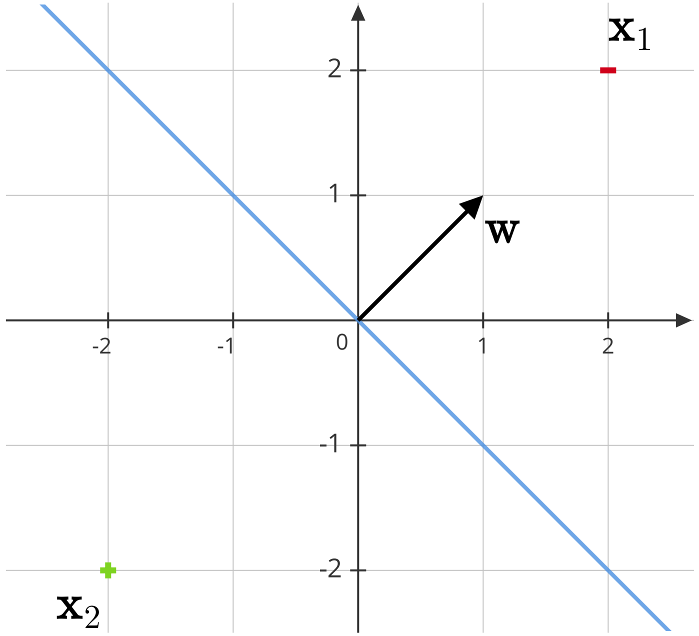
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 \ =$$



We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

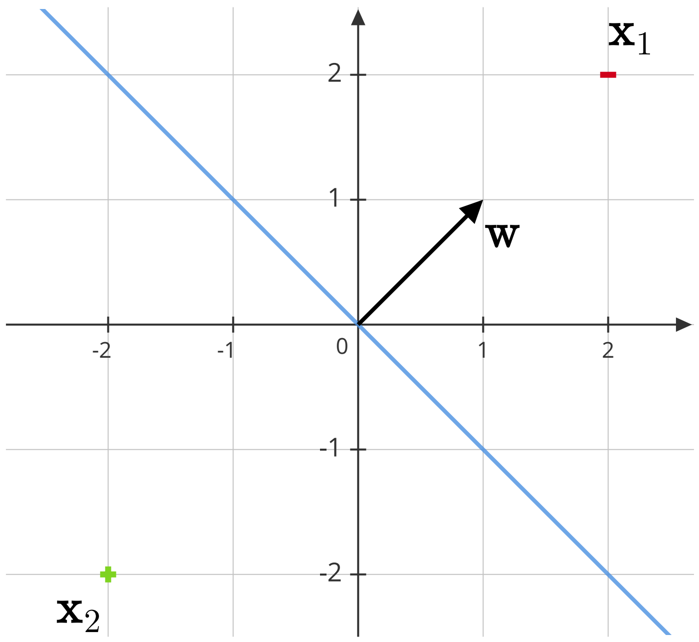
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$



We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2 + 2 = 4$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -2 \\ -2 \end{bmatrix} = -2 - 2 = -4$$

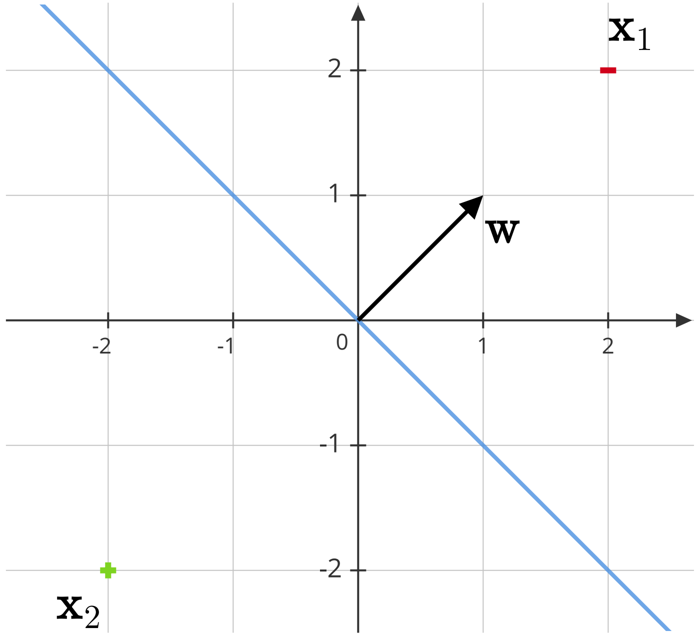


We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = \begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2\\2\end{bmatrix} = 2+2 = 4$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 = \begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}-2\\-2\end{bmatrix} = -2-2 = -4$$

Note that $\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -\mathbf{w}^{\mathrm{T}}\mathbf{x}_2$



We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2 + 2 = 4$$

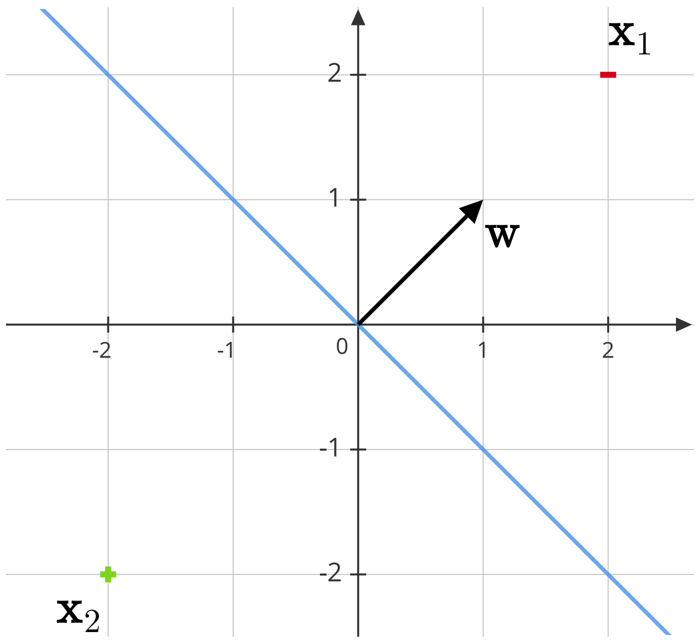
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -2 \\ -2 \end{bmatrix} = -2 - 2 = -4$$

Note that $\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -\mathbf{w}^{\mathrm{T}}\mathbf{x}_2$

$$\implies \boxed{\sigma_2 = 1 - \sigma_1}$$

$$\sigma_1 = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_1\right), \quad \sigma_2 = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_2\right)$$

Also note $y_1 = 0, \; y_2 = +1$



We begin with the question. We will first compute

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = \begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2\\2\end{bmatrix} = 2+2=4$$

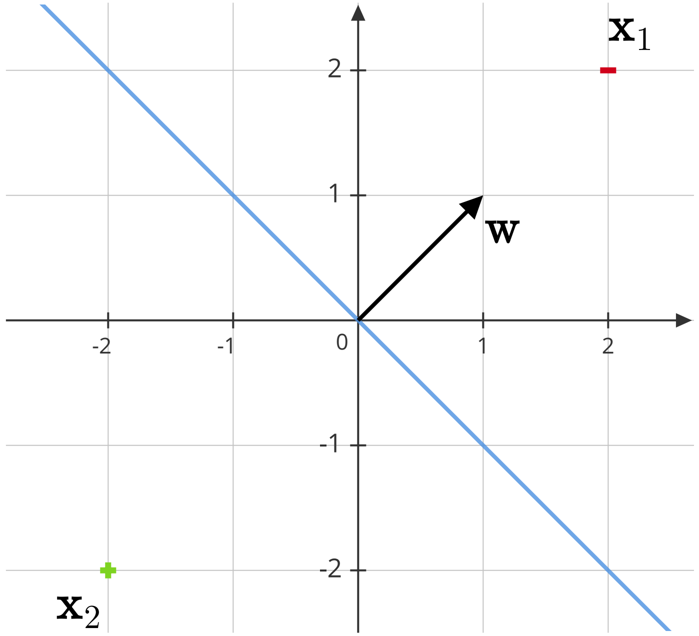
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 = \begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}-2\\-2\end{bmatrix} = -2-2=-4$$

Note that $\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -\mathbf{w}^{\mathrm{T}}\mathbf{x}_2$

$$\implies \boxed{\sigma_2 = 1-\sigma_1}$$

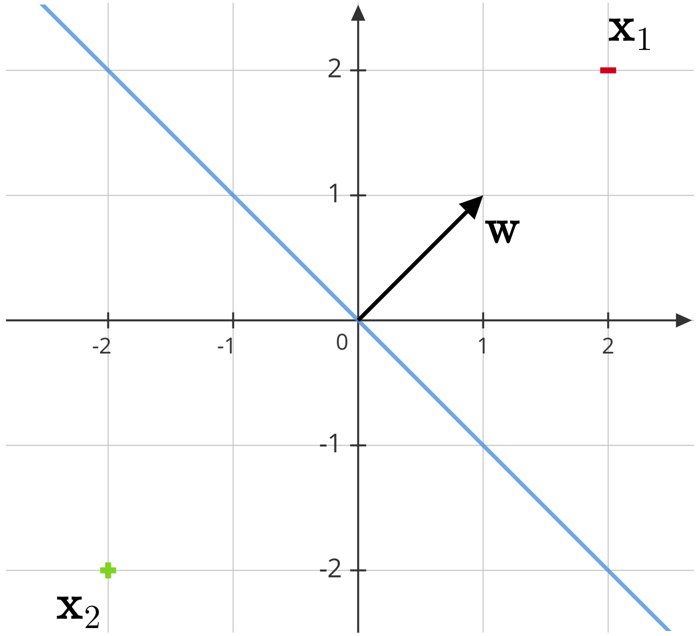
Also note $y_1 = 0, \;\; y_2 = +1$

We are now ready to compute $\dfrac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|}$



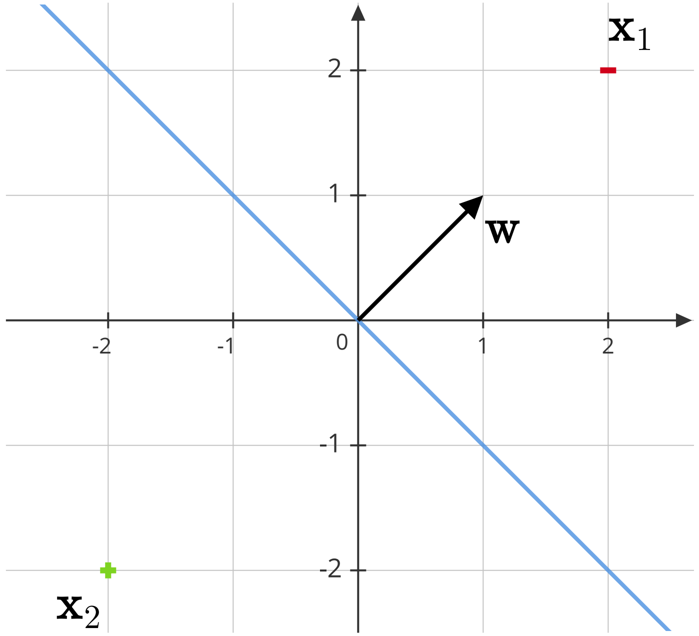
$$\boxed{\begin{aligned} \sigma_2 &= 1-\sigma_1 \\ y_1 = 0,\ y_2 &= +1 \end{aligned}}$$

$$\frac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|}$$



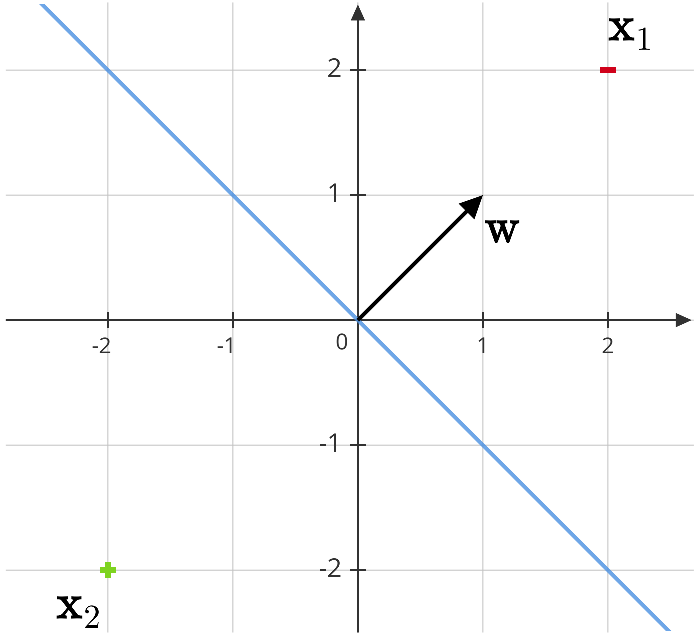
$$\boxed{\begin{aligned} \sigma_2 &= 1 - \sigma_1 \\ y_1 = 0,\ y_2 &= +1 \end{aligned}}$$

$$\frac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|} = \frac{|\sigma_1 - 0|}{|1 - \sigma_1 - 1|}$$



$$\boxed{\begin{aligned}\sigma_2 &= 1-\sigma_1 \\ y_1 = 0,\ y_2 &= +1\end{aligned}}$$

$$\frac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|} = \frac{|\sigma_1 - 0|}{|1-\sigma_1-1|} = \frac{|\sigma_1|}{|-\sigma_1|}$$



$$\boxed{\begin{aligned} \sigma_2 &= 1 - \sigma_1 \\ y_1 = 0,\ y_2 &= +1 \end{aligned}}$$

$$\frac{|\sigma_1 - y_1|}{|\sigma_2 - y_2|} = \frac{|\sigma_1 - 0|}{|1 - \sigma_1 - 1|} = \frac{|\sigma_1|}{|-\sigma_1|} = 1$$

Note that this did not require an explicit computation of the output of the logistic function.