UNIVERSITY OF LEEDS

Class: Machine Learning

Neural Networks: Perceptron

Instructor: Matteo Leonetti

Learning outcomes

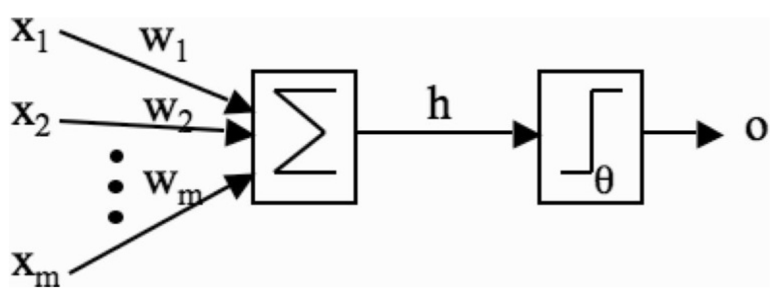
- Define an appropriate error function for the perceptron.
- Derive the corresponding update algorithm.
- Describe the difference between gradient descent and stochastic gradient descent.

Recap

We want to apply gradient descent:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

To the parameters of a perceptron:



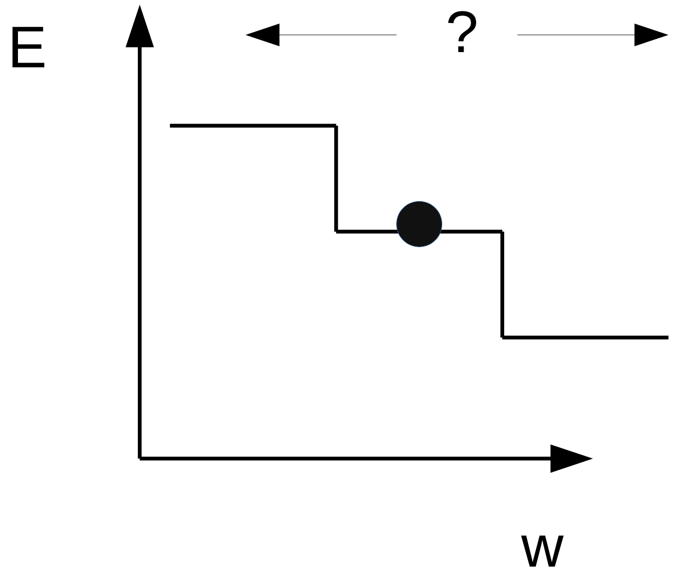
So as to minimise an error (or loss) function, such as:

$$E(\boldsymbol{X}) = \sum_{\vec{x}_n \in \boldsymbol{X}} |y_n - t_n|$$

Number of mistakes

$$E(\boldsymbol{X}) = \sum_{\vec{x}_n \in \boldsymbol{X}} |y_n - t_n|$$

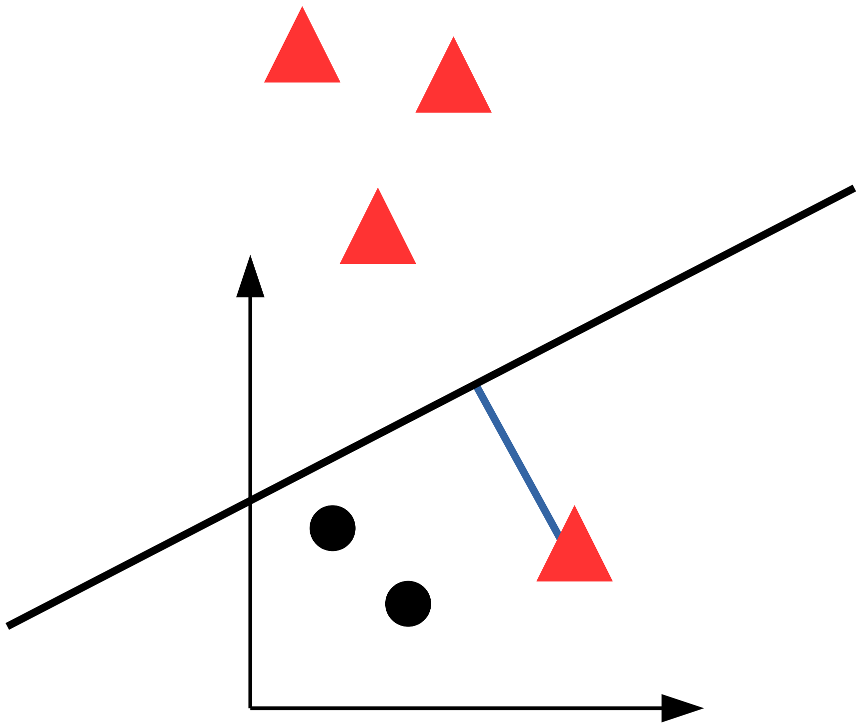
Number of mistakes on the dataset. Piecewise constant → no gradient.



There is no local information on the direction of improvement

Towards a better error function

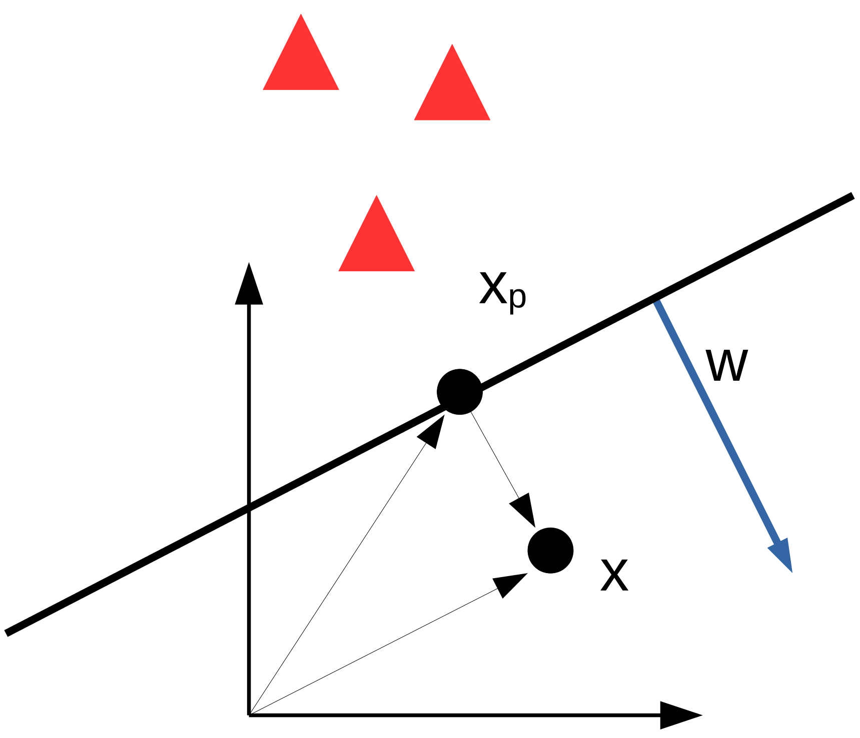
UNIVERSITY OF LEEDS



For each misclassified point, we would like to know not only that they are on the wrong side, but also **by how much**.

Towards a better error function

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = 0$$



Distance to the hyperplane

$$\boldsymbol{x} = \boldsymbol{x}_p + d\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}\left(\boldsymbol{x}_p + d\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0$$

$$= \cancel{\boldsymbol{w}\,\boldsymbol{x}_p + w_0} + d\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|} = d\|\boldsymbol{w}\|$$

Recall that:

$$\boldsymbol{w}^T\boldsymbol{w} = w_1^2 + w_2^2 + \cdots + w_n^2 = \|\boldsymbol{w}\|^2$$

Towards a better error function

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = d \|\boldsymbol{w}\|$$

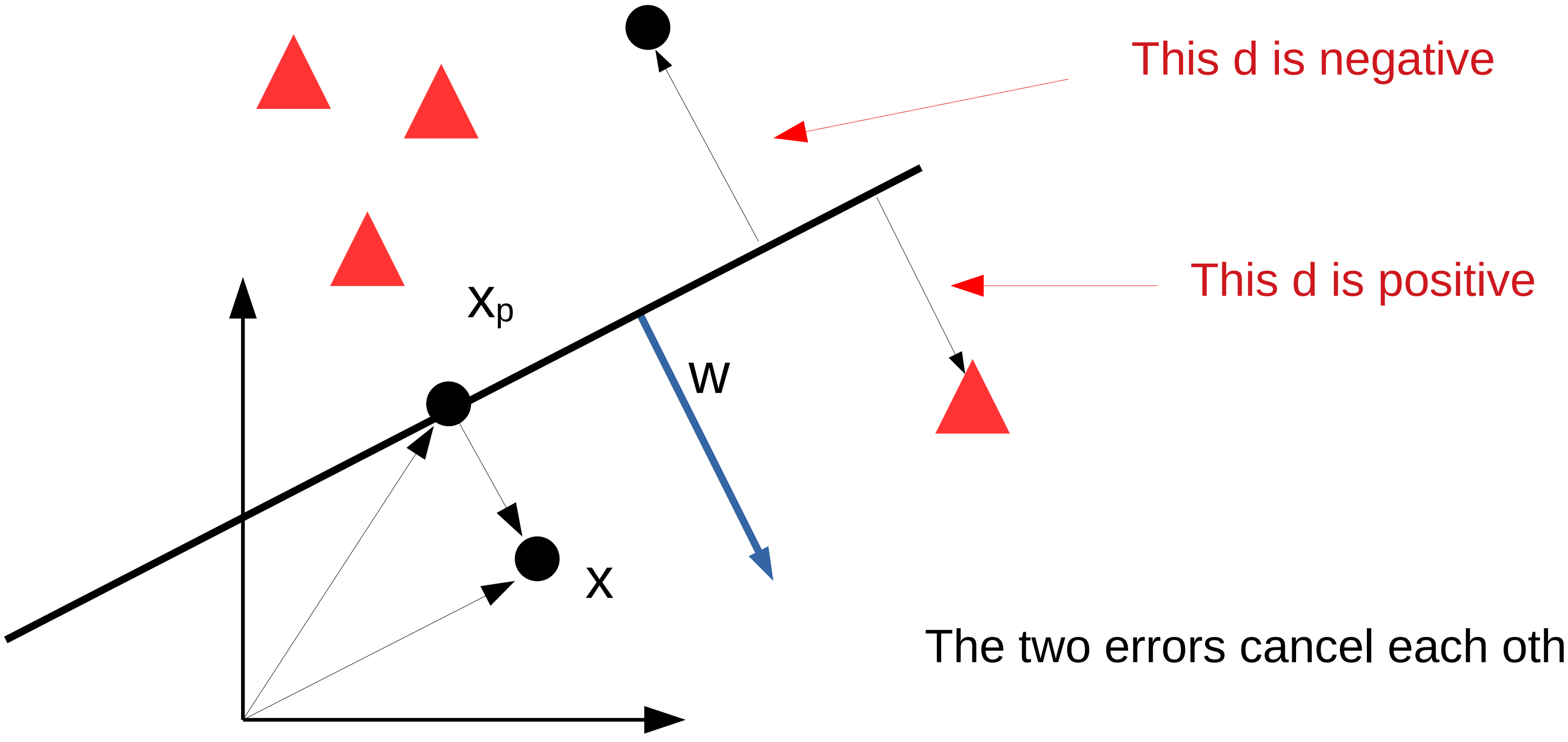
$$E(\boldsymbol{X}) = \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} (\boldsymbol{w}^T \boldsymbol{x}_n + w_0)$$

Is this a good error?

Towards a better error function

UNIVERSITY OF LEEDS

$$x = x_p + d\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$



The two errors cancel each other out!

The perceptron criterion

UNIVERSITY OF LEEDS

$h_w(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = 0$ apply the bias input

if $\boldsymbol{w}^T \boldsymbol{x} > 0$ then $y = 1$ In case of mistake: $t = 0$ $(y - t) = 1$

if $\boldsymbol{w}^T \boldsymbol{x} \leq 0$ then $y = 0$ In case of mistake: $t = 1$ $(y - t) = -1$

Therefore, if mistake: $\boldsymbol{w}^T \boldsymbol{x}(y - t) > 0$

$$E(\boldsymbol{X}) = \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} |y_n - t_n|$$

Number of mistakes on the dataset. Piecewise constant → gradient useless.

$$E_p(\boldsymbol{X}) = \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} \boldsymbol{w}^T \boldsymbol{x}_n (y_n - t_n)$$

Proportional to distance of misclassified points from surface. → gradient ok.

Derivatives

What is the derivative of

$$y = 2x \quad ?$$

Question

Given the perceptron error (below), what is the gradient with respect to **w**?

$$E_p(\boldsymbol{X}) = \boldsymbol{w}^T \boldsymbol{x}(y-t) = (w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n)(y-t)$$

Solution

UNIVERSITY OF LEEDS

$$E_p(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}(y-t) = w_0 x_0 (y-t) + w_1 x_1 (y-t) + \cdots + w_m x_m (y-t)$$

$$\nabla E_p(x) = \begin{bmatrix} \frac{\partial}{\partial w_0} E_p(x) \\ \frac{\partial}{\partial w_1} E_p(x) \\ \frac{\partial}{\partial w_2} E_p(x) \\ \ldots \\ \frac{\partial}{\partial w_n} E_p(x) \end{bmatrix} = \begin{bmatrix} x_0(y-t) \\ x_1(y-t) \\ x_2(y-t) \\ \ldots \\ x_n(y-t) \end{bmatrix}$$

Gradient descent

UNIVERSITY OF LEEDS

$$\nabla E_p(\boldsymbol{X}) = \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} \boldsymbol{x}_n (y_n - t_n)$$

Recall that gradient descent does the following update:

$$w_{k+1} = w_k - \eta \nabla f(w_k)$$

Which leads us to the update rule for the perceptron:

$$w_{k+1} = w_k - \eta \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} \boldsymbol{x}_n (y_n - t_n)$$

Stochastic gradient descent

$$E_p(\boldsymbol{X}) = \frac{1}{N} \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} \boldsymbol{w}^T \boldsymbol{x}_n (y_n - t_n) = \boldsymbol{E}[\boldsymbol{w}^T \boldsymbol{x}_n (y_n - t_n)]$$

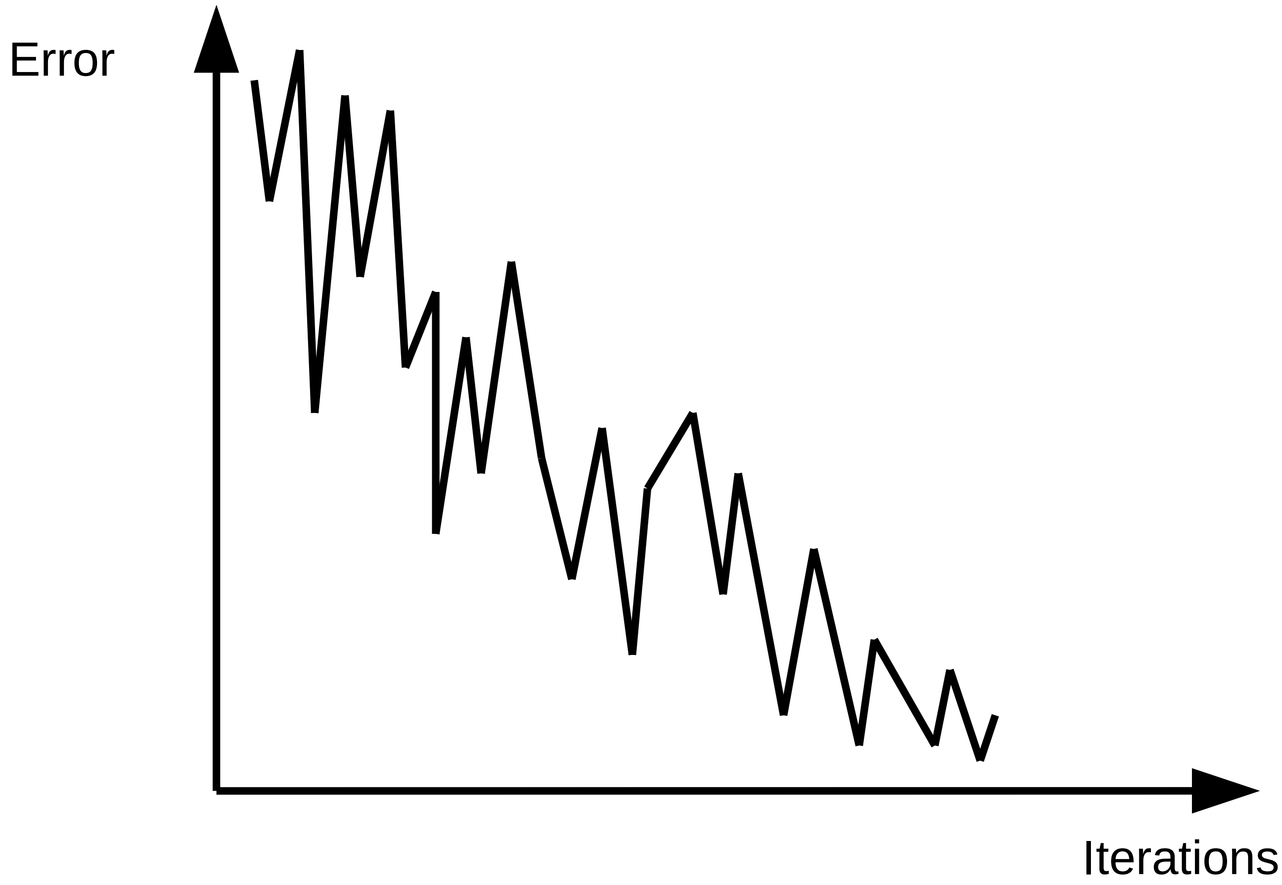
Gradient:

$$\boldsymbol{w}_{\boldsymbol{k+1}} = \boldsymbol{w}_{\boldsymbol{k}} - \eta \frac{1}{N} \sum_{\boldsymbol{x}_n \in \boldsymbol{X}} \boldsymbol{x}_n (y_n - t_n)$$

Stochastic gradient descent:

$$\boldsymbol{w}_{\boldsymbol{k+1}} = \boldsymbol{w}_{\boldsymbol{k}} - \eta \, \boldsymbol{x} (y - t)$$

Stochastic gradient descent

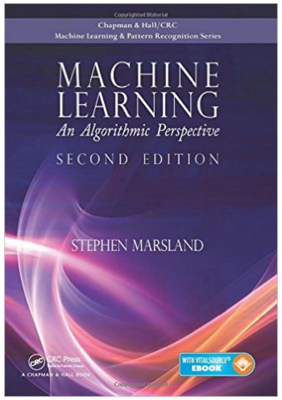


Conclusion

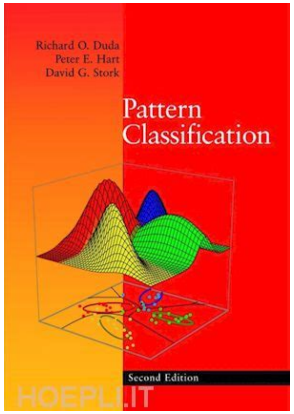
Learning outcomes

UNIVERSITY OF LEEDS

- Define an appropriate error function for the perceptron.
- Derive the corresponding update algorithm.
- Describe the difference between gradient descent and stochastic gradient descent.



Section 3.4



Book in Minerva
in " Online Course Readings Folder"

Section 5.2.1, 5.4. and 5.5
(without convergence proof)