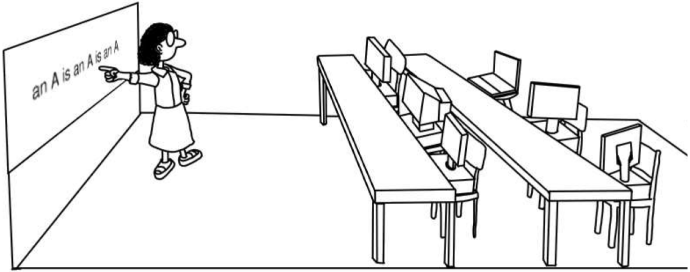
Machine Learning

Krystian Mikolajczyk & Deniz Gunduz

Department of Electrical and Electronic Engineering

Imperial College London



Machine Learning - Part 1.1 Summary

Department of Electrical and Electronic Engineering

Imperial College London

- A simple hypothesis class

A Simple Hypothesis Class – Linear predictor

For input $\mathbf{x} = (x_1, \ldots, x_d)$ (numerical representation of data),

and hypothesis $\mathbf{w} = (w_1, \ldots, w_d)$ (model parameters),

the linear predictor is $h(\mathbf{x}, \mathbf{w}) \to y$, with $\sum_{i=1}^{d} w_i x_i = \mathbf{w}^T \mathbf{x}$

Classification (binary)

Label: $y \in \mathcal{Y} = \{-1, +1\}$

$$h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} < t : \quad h(\mathbf{x}, \mathbf{w}) = -1$$

$$h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \geqslant t : \quad h(\mathbf{x}, \mathbf{w}) = +1$$

$$\sum_{i=1}^{d} w_i x_i - t \geqslant 0$$

$$\sum_{i=1}^{d} w_i x_i - w_0 x_0 \geqslant 0, \text{ with } x_0 = 1$$

$$\sum_{i=0}^{d} w_i x_i \geqslant 0 \to \text{sign}(\textstyle\sum_{i=0}^{d} w_i x_i)$$

Predictor: $h(\mathbf{x}, \mathbf{w}) = \text{sign}(\mathbf{w}^T \mathbf{x})$

Regression

Label: $y \in \mathcal{Y} \subset \mathbb{R}$

Predictor: $h(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

Perceptron Learning Algorithm

Perceptron Learning Algorithm (PLA)

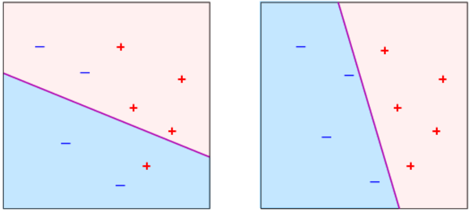
1. While there exists a missclassified data point with

   $\text{sign}(\mathbf{w}^\top \mathbf{x}_i) \neq y_i \qquad\qquad y \in \{-1, +1\}$

2. update $\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$

Intuition: $y_i h(\mathbf{x}_i) > 0$ if $\mathbf{x}_i$ is correctly classified and $y_i h(\mathbf{x}_i) < 0$ if incorrectly.

$$\begin{aligned} y_i \cdot \mathbf{w}'^\top \mathbf{x}_i &= y_i \cdot (\mathbf{w} + y_i \mathbf{x}_i)^\top \mathbf{x}_i \\ &= y_i \cdot \mathbf{w}^\top \mathbf{x}_i + y_i^2 \cdot \mathbf{x}_i^\top \mathbf{x}_i \\ &= y_i \cdot \mathbf{w}^\top \mathbf{x}_i + \|\mathbf{x}_i\|^2 \end{aligned}$$



linearly separable data

Remark: Algorithm stops after a finite number of steps if the data is separable.

Simple Regressor – Closed form solution

Regression error $\widehat{R}_n(h)$

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=0}^{n}(h(\mathbf{w},\mathbf{x}_i) - y_i)^2$$
$$\widehat{R}_n(\mathbf{w}) = \frac{1}{n}\sum_{i=0}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$$
$$\widehat{R}_n(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

with data matrix X and label vector $\mathbf{y}$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Solution

find $\mathbf{w}$ that minimizes error $\widehat{R}(\mathbf{w})$ i.e. $\dot{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \widehat{R}(\mathbf{w})$

$$\nabla \widehat{R}_n(\mathbf{w}) = 0, \qquad \frac{2}{n}X^T(X\mathbf{w} - y) = 0$$
$$X^TX\mathbf{w} = X^T\mathbf{y}$$
$$\mathbf{w} = (X^TX)^{-1}X^T\mathbf{y} = \dot{X}\mathbf{y}$$

where $\dot{X}$ is Moore-Penrose pseudoinverse

Classification/regression example

$x_1 = 1, x_2 = 2, x_3 = 7, x_4 = 8$

$y_1 = -1, y_2 = -1, y_3 = +1, y_4 = +1$

$\mathbf{w} = ?$

$\mathbf{x}_1 = (1,1), \mathbf{x}_2 = (1,2), \mathbf{x}_3 = (1,7), \mathbf{x}_4 = (1,8)$

$\dot{\mathbf{w}} = (w_0, w_1) = ?$

$$\text{find } \operatorname{sign}(\mathbf{w}^\top \mathbf{x}_i) \neq y_i \qquad \text{update } \mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$$

$$\text{initial } \mathbf{w}_{(1)} = (0,0) \qquad \text{iteration 1: } \mathbf{w}_{(1)}^T \mathbf{x}_1 = 0 \Rightarrow +1 \neq y_1, \Rightarrow \mathbf{w} = y_1 \mathbf{x}_1$$

$$\mathbf{w}_{(2)} = (-1,-1) \qquad \text{iteration 2: } \mathbf{w}_{(2)}^T \mathbf{x}_1 = -2 \Rightarrow -1 = y_1$$

$$\mathbf{w}_{(3)} = (-1,-1) \qquad \mathbf{w}_{(3)}^T \mathbf{x}_2 = -3 \Rightarrow -1 = y_2$$

$$\mathbf{w}_{(4)} = (-1,-1) \qquad \mathbf{w}_{(4)}^T \mathbf{x}_3 = -8 \Rightarrow -1 \neq y_3 \Rightarrow \mathbf{w}_{(5)} = \mathbf{w}_{(4)} + y_3 \mathbf{x}_3$$

$$\mathbf{w}_{(5)} = (0,6) \qquad \mathbf{w}_{(5)}^T \mathbf{x}_1 = 6 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(6)} = \mathbf{w}_{(5)} + y_1 \mathbf{x}_1$$

$$\mathbf{w}_{(6)} = (-1,5) \qquad \mathbf{w}_{(6)}^T \mathbf{x}_1 = 4 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(7)} = \mathbf{w}_{(6)} + y_1 \mathbf{x}_1$$

Classification/regression example

$\mathbf{x}_1 = (1,1), \mathbf{x}_2 = (1,2), \mathbf{x}_3 = (1,7), \mathbf{x}_4 = (1,8)$

$y_1 = -1, y_2 = -1, y_3 = +1, y_4 = +1$

$\mathbf{w}_{(7)} = (-2,4) \qquad \mathbf{w}_{(7)}^T\mathbf{x}_1 = 2 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(8)} = \mathbf{w}_{(7)} + y_1\mathbf{x}_1$

$\mathbf{w}_{(8)} = (-3,3) \qquad \mathbf{w}_{(8)}^T\mathbf{x}_1 = 0 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(9)} = \mathbf{w}_{(8)} + y_1\mathbf{x}_1$

$\mathbf{w}_{(9)} = (-4,2) \qquad \mathbf{w}_{(9)}^T\mathbf{x}_1 = -2 \Rightarrow -1 = y_1 \Rightarrow \mathbf{w}_{(9)}$

$\mathbf{w}_{(9)} = (-4,2) \qquad \mathbf{w}_{(9)}^T\mathbf{x}_2 = 0 \Rightarrow +1 \neq y_2 \Rightarrow \mathbf{w}_{(10)} = \mathbf{w}_{(9)} + y_2\mathbf{x}_2$

$\mathbf{w}_{(10)} = (-5,0) \qquad \mathbf{w}_{(10)}^T\mathbf{x}_3 = -5 \Rightarrow -1 \neq y_3 \Rightarrow \mathbf{w}_{(11)} = \mathbf{w}_{(10)} + y_3\mathbf{x}_3$

$\mathbf{w}_{(11)} = (-4,7) \qquad \mathbf{w}_{(10)}^T\mathbf{x}_1 = 3 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(12)} = \mathbf{w}_{(11)} + y_1\mathbf{x}_1$

$\mathbf{w}_{(12)} = (-5,6) \qquad \mathbf{w}_{(12)}^T\mathbf{x}_1 = 1 \Rightarrow +1 \neq y_1 \Rightarrow \mathbf{w}_{(13)} = \mathbf{w}_{(12)} + y_1\mathbf{x}_1$

$\mathbf{w}_{(13)} = (-6,5) \qquad \mathbf{w}_{(13)}^T\mathbf{x}_2 = 4 \Rightarrow +1 \neq y_2 \Rightarrow \mathbf{w}_{(14)} = \mathbf{w}_{(13)} + y_2\mathbf{x}_2$

$\mathbf{w}_{(14)} = (-7,3) \qquad \mathbf{w}_{(14)}^T\mathbf{x}_4 = 17 \Rightarrow +1 = y_4 \Rightarrow \mathbf{w}_{(14)} = \mathbf{w}_{(14)}$

SOLVED! $\dot{\mathbf{w}} = (-7,3)$

Classification/regression example

$\mathbf{x}_1 = (1,1), \mathbf{x}_2 = (1,2), \mathbf{x}_3 = (1,7), \mathbf{x}_4 = (1,8)$

$y_1 = -1, y_2 = -1, y_3 = +1, y_4 = +1$

$\widehat{R}_n(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$

$\nabla \widehat{R}_n(\mathbf{w}) = 0, \qquad \frac{2}{n}X^T(X\mathbf{w} - y) = 0$

$X^TX\mathbf{w} = X^T\mathbf{y}$

$\mathbf{w} = (X^TX)^{-1}X^T\mathbf{y} = \dot{X}\mathbf{y}$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1,1 \\ 1,2 \\ 1,7 \\ 1,8 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

Classification/regression example

$$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y} = \dot{X}\mathbf{y}$$

$$X^T X = \begin{bmatrix} 1,1,1,1 \\ 1,2,7,8 \end{bmatrix} \begin{bmatrix} 1,1 \\ 1,2 \\ 1,7 \\ 1,8 \end{bmatrix} = \begin{bmatrix} 4,18 \\ 18,118 \end{bmatrix}$$

$$\dot{X} = (X^T X)^{-1} X^T = \begin{bmatrix} 0.7973 - 0.1216 \\ -0.12160.0270 \end{bmatrix} \begin{bmatrix} 1,1,1,1 \\ 1,2,7,8 \end{bmatrix} = \begin{bmatrix} 0.676, 0.554, -0.054, -0.176 \\ -0.095, -0.068, 0.068, 0.095 \end{bmatrix}$$

$$\mathbf{w} = \dot{X}\mathbf{y} = \begin{bmatrix} 0.676, 0.554, -0.054, -0.176 \\ -0.095, -0.068, 0.068, 0.095 \end{bmatrix} \begin{bmatrix} -1, -1, +1, +1 \end{bmatrix}^T = \begin{bmatrix} -1.5 \\ 0.32 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Classification/regression example

$\mathbf{x}_1 = (1, 1), \mathbf{x}_2 = (1, 2), \mathbf{x}_3 = (1, 7), \mathbf{x}_4 = (1, 8)$

$y_1 = -1, y_2 = -1, y_3 = +1, y_4 = +1$

Regression solution used for classification:

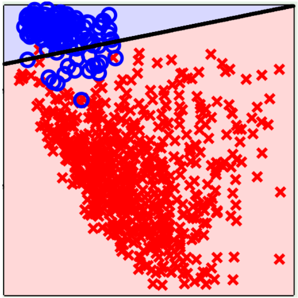
$\mathbf{w}_{(1)} = (-1.5, 0.32) \qquad \mathbf{w}_{(1)}^T \mathbf{x}_1 = -1.18 \Rightarrow -1 = y_1$

$\mathbf{w}_{(1)} = (-1.5, 0.32) \qquad \mathbf{w}_{(1)}^T \mathbf{x}_2 = -0.86 \Rightarrow -1 = y_2$

$\mathbf{w}_{(1)} = (-1.5, 0.32) \qquad \mathbf{w}_{(1)}^T \mathbf{x}_3 = 0.74 \Rightarrow +1 = y_3$

$\mathbf{w}_{(1)} = (-1.5, 0.32) \qquad \mathbf{w}_{(1)}^T \mathbf{x}_4 = 1.06 \Rightarrow +1 = y_4$

Linear Regression for Classification



- Linear regression learns a real-valued function $y = f(\mathbf{x}) \in \mathbb{R}$.
- Binary valued functions are also real valued: $\pm 1 \in \mathbb{R}$.
- Use linear regression to get $\dot{\mathbf{w}}$ such that $\mathbf{w}^\top \mathbf{x}_i \approx y_i \in \{+1, -1\}$
- Then it is likely that $\text{sign}(\mathbf{w}^\top \mathbf{x}_i) = y_i$.
- Good initial weights for classification
- Error function suboptimal

$$\widehat{R}_{regression}(\mathbf{w}) = \frac{1}{n}\sum_i (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \qquad \widehat{R}_{classification}(\mathbf{w}) = \frac{1}{n}\sum_i \mathbb{I}[\text{sign}(\mathbf{w}^T \mathbf{x}_i) \neq y_i] \tag{1}$$