

Machine Learning

Linear Models

Fabio Vandin

October 18th, 2022

Linear Predictors and Affine Functions

Consider $\mathcal{X} = \mathbb{R}^d$

“Linear” (affine) functions:

$$L_d = \{h_{\mathbf{w},b} : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

where

$$h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \left(\sum_{i=1}^d w_i x_i \right) + b$$

Note:

- each member of L_d is a function $\mathbf{x} \rightarrow \langle \mathbf{w}, \mathbf{x} \rangle + b$
- b : *bias*

Linear Models

Hypothesis class \mathcal{H} : $\phi \circ L_d$, where $\phi : \mathbb{R} \rightarrow \mathcal{Y}$

- $h \in \mathcal{H}$ is $h : \mathbb{R}^d \rightarrow \mathcal{Y}$

ϕ depends on the learning problem

Example

- binary classification, $\mathcal{Y} = \{-1, 1\} \Rightarrow \phi(z) = \text{sign}(z)$
- regression, $\mathcal{Y} = \mathbb{R} \Rightarrow \phi(z) = z$

Equivalent Notation

Given $\mathbf{x} \in \mathcal{X}$, $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, define:

- $\mathbf{w}' = (b, w_1, w_2, \dots, w_d) \in \mathbb{R}^{d+1}$
- $\mathbf{x}' = (1, x_1, x_2, \dots, x_d) \in \mathbb{R}^{d+1}$

Then:

$$h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \langle \mathbf{w}', \mathbf{x}' \rangle \quad (1)$$

\Rightarrow we will consider bias term as part of \mathbf{w} and assume $\mathbf{x} = (1, x_1, x_2, \dots, x_d)$ when needed, with $h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$

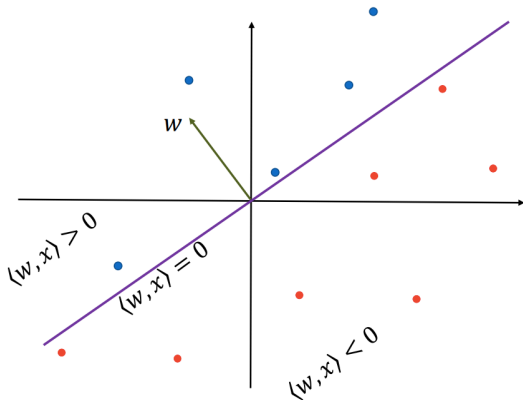
Linear Classification

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$, 0-1 loss

Hypothesis class = *halfspaces*

$$HS_d = \text{sign} \circ L_d = \{\mathbf{x} \rightarrow \text{sign}(h_{\mathbf{w},b}(\mathbf{x})) : h_{\mathbf{w},b} \in L_d\}$$

Example: $\mathcal{X} = \mathbb{R}^2$



Finding a Good Hypothesis

Linear classification with hypothesis set \mathcal{H} = halfspaces.

How do we find a good hypothesis?

Good = minimizes the training error (ERM)

\Rightarrow Perceptron Algorithm (Rosenblatt, 1958)

Note:

if $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$ for all $i = 1, \dots, m \Rightarrow$ all points are classified correctly by model $\mathbf{w} \Rightarrow$ *realizability assumption* for training set

Linearly separable data: there exists \mathbf{w} such that: $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$