Chapter 1

Linear predictors

In this Chapter we will study the family of linear predictors, one of the most useful families of hypothesis classes. Many learning algorithms that are being widely used in practice rely on linear predictors, for their ability to learn them efficiently in many cases and because they are intuitive and easy to interpret.
In particular, we focus on learning linear predictors using the ERM approach. The hypothesis classes can include:

- Halfspaces (classification).
- Linear regression (regression).
- Logistic regression (classification moduled as a regression problem).

The algorithms can include:

- Linear programming (for halfspaces).
- Perceptron (for halfspaces).
- Least squares (for regression).

First, we define the class of **affine functions** as:

$$L_d = \{h_{\mathbf{w},b} \ : \ \mathbf{w} \in \mathbb{R}^d, \ b \in \mathbb{R}\} \tag{1.1}$$

where:

$$h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \left(\sum_{i=1}^{d} w_i x_i \right) + b \tag{1.2}$$

It is convenient to use the following notation for L_d:

$$L_d = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b \ : \ \mathbf{w} \in \mathbb{R}^d, \ b \in \mathbb{R}\} \tag{1.3}$$

which reads as follows: L_d is a set of functions, where each function is parametrized by $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, and each such function takes as input a vector $\mathbf{x}$ and returns as output the scalar $\langle \mathbf{w}, \mathbf{x} \rangle + b$. Therefore the dimension of the parameter space increases from d to $d + 1$.
The different hypothesis classes of linear predictors are composition of function $\varphi : \mathbb{R} \to \mathcal{Y}$ on L_d. For example:

- For binary classification: $\mathcal{Y} = \{-1, 1\} \to \varphi(z) = \text{sign}(z)$
- For regression: $\mathcal{Y} = \mathbb{R} \to \varphi(z) = z$

$$\begin{aligned} \mathbf{w}' &= (b, w_1, w_2, \dots, w_d) \in \mathbb{R}^{d+1} \\ \mathbf{x}' &= (1, x_1, x_2, \dots, x_d) \in \mathbb{R}^{d+1} \end{aligned} \tag{1.4}$$

Therefore $h_{\mathbf{w},b}(\mathbf{x})$ becomes a linear function:

$$h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \langle \mathbf{w}', \mathbf{x}' \rangle \tag{1.5}$$

1.1 Halfspaces

They are exploited for binary classification problems, namely $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. The class of halfspaces is defined as follows:

$$HS_d = \operatorname{sign} \circ L_d = \{\mathbf{x} \mapsto \operatorname{sign}(h_{\mathbf{w},b}(\mathbf{x})) \ : \ h_{\mathbf{w},b} \in L_d\} \tag{1.6}$$

In other words, each halfspace hypothesis in HS_d is parametrized by $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ and upon receiving a vector $\mathbf{x}$ the hypotesis returns the label sign $(\langle \mathbf{w}, \mathbf{x} \rangle + b)$. The realizability can be obtained only if the space is linearly separable.
To illustrate this hypothesis class geometrically, let's consider the case $d = 2$. Each hypothesis forms a hyperplane that is perpendicular to the vector $\mathbf{w}$ and intersects the vertical axis at the point $(0, -\frac{b}{w_2})$. The istances that are "above" the hyperplane, that is, share an acute angle with $\mathbf{w}$, are labeled positively. Istances that are "below" the hyperplane, that is, share an obtuse angle with $\mathbf{w}$, are labeled negatively.

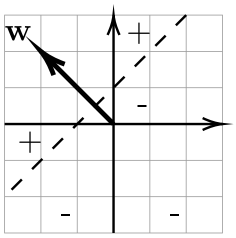

Figure 1.1: Example of halfspace with $d = 2$.

From Fig. 1.1 it is clear the condition needed for the realizability. Implementing the ERM rule in non-separable case (i.e. in the agnostic case) is known to be computationally hard, but there are several approaches to tackle problems with non-separable data.

1.1.1 Linear programming for the class of halfspaces

Linear programs (LP) are problems that can be expressed as maximizing a linear function subject to linear inequalities (constraints). That is:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \langle \mathbf{u}, \mathbf{w} \rangle \qquad \text{subject to } A\mathbf{w} \geq \mathbf{v} \tag{1.7}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the vector of variables we wish to determine, A is an $m \times d$ matrix, and $\mathbf{v} \in \mathbb{R}^m$, $\mathbf{u} \in \mathbb{R}^d$ are vectors. Linear programming can be solved efficiently[1], and furthermore, there are publicly implementations of LP solvers.
The ERM problem for halfspaces in the realizable case can be expressed as a linear program. To show this, we assume for simplicity the homogeneous case. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a training set of size m. We can say:

[1] Namely, in time polynomial in m, d, and in the representation size of real numbers.

- For the realizable assumption, an ERM predictor should have zero errors on the training set. That is, we are looking for some vector $\mathbf{w} \in \mathbb{R}^d$ for which:

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \qquad \forall i = 1, \ldots, m \tag{1.8}$$

- There exists also a vector $\mathbf{w}$ that satisfy

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \qquad \forall i = 1, \ldots, m \tag{1.9}$$

Proof. Let $\mathbf{w}^*$ be a vector that satisfies $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \ \forall i$. We define $\gamma = \min_i(y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle)$ and let $\bar{\mathbf{w}} = \frac{\mathbf{w}^*}{\gamma}$. Therefore, $\forall i$ we have:

$$y_i \langle \bar{\mathbf{w}}, \mathbf{x}_i \rangle = \frac{1}{\gamma} y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$$

We have thus shown that there exists a vector that satisfies Eq. 1.9. □

- We want to express the ERM model with linear programming. Let A be a $m \times d$ matrix such that $A_{ij} = y_i x_{ij}$, with x_ij the j^{th} element of the vector $\mathbf{x}_i$. Let $\mathbf{v}$ be the vector $(1, \ldots, 1) \in \mathbb{R}^m$. Then, Eq. 1.9 can be rewritten as:

$$A\mathbf{w} \geq \mathbf{v}$$

The LP form requires a maximization objective, yet all the $\mathbf{w}$ that satisfy the constraints are equal candidates as output hypotheses. Thus, we set a "dummy" objective $\mathbf{u} = (0, \ldots, 0) \in \mathbb{R}^d$.

1.1.2 Perceptron for halfspaces

It is a different implementation of the ERM rule by Rosenblatt (1958). The perceptron is an iterative algorithm that constructs a sequence of vectors $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)}$, ... Initially, $\mathbf{w}^{(1)}$ is set to be the all-zeros vector. At iteration time t, the perceptron finds an example i that is mislabeled by $\mathbf{w}^{(t)}$, namely, an example for which $\text{sign}(\langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle) \neq y_i$. Then, the perceptron updates $\mathbf{w}^{(t)}$ by adding to it the istance $\mathbf{x}_i$ scaled by the label y_i. That is:

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

Our goal is to have $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \ \forall i$ and note that:

$$y_i \left\langle \mathbf{w}^{(t+1)}, \mathbf{x}_i \right\rangle = y_i \left\langle \mathbf{w}^{(t)} + y_i \mathbf{x}_i, \mathbf{x}_i \right\rangle = y_i \left\langle \mathbf{w}^{(t)}, \mathbf{x}_i \right\rangle + \|\mathbf{x}_i\|^2$$

Hence, the update of the perceptron guides the solution to be "more correct" on the i^{th} example.

Algorithm 1: Batch Perceptron.

```
input      : A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
output     : $\mathbf{w}^{(t)}$
initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$
for $t = 1, 2, \ldots$ do
    if $\exists i$ s.t. $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$ then
        $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
    else
        return $\mathbf{w}^{(t)}$
```

The following theorem guarantees that in the realizable case, the algorithm stops with all sample points correctly classified.

Theorem 1.1.1. *Assume that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ is separable, let:*

$$\begin{aligned}&B = \min\left\{\|\mathbf{w}\| \ : \ \forall i \in [m], \ y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1\right\}\\ &R = \max{}_i \|\mathbf{x}_i\|\end{aligned}$$

Then, the perceptron algorithm stops after at most $(RB)^2$ iterations, and when it stops it holds that $\forall i \in [m], \ y_i \left\langle \mathbf{w}^{(t)}, \mathbf{x}_i \right\rangle > 0$.

Proof. We give the proof in steps:

▷ Let's define a vector $\mathbf{w}^*$ achieving the minimum in B, and T the number of iterations before stopping.

▷ Consider:

$$\frac{\left\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \right\rangle}{\|\mathbf{w}^*\| \left\|\mathbf{w}^{(t+1)}\right\|} \leq 1 \tag{1.10}$$

▷ We need to demonstate:

$$\frac{\sqrt{T}}{RB} \leq \frac{\left\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \right\rangle}{\|\mathbf{w}^*\| \left\|\mathbf{w}^{(T+1)}\right\|} \leq 1 \Longrightarrow T < (RB)^2 \tag{1.11}$$

▷ We divide this step into two parts:

a) Numerator:
We want to demonstate that: $\left\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \right\rangle \geq T$.

- First iteration: $\mathbf{w}^{(1)} = (0, \ldots, 0) \Longrightarrow \left\langle \mathbf{w}^*, \mathbf{w}^{(1)} \right\rangle = 0$
- At each step: $\left\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \right\rangle - \left\langle \mathbf{w}^*, \mathbf{w}^{(t)} \right\rangle \geq 1$
- After T iterations: $\left\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \right\rangle \geq T$
  In fact:

$$\begin{aligned}\left\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \right\rangle &= \sum_{t=1}^{T} \left(\left\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \right\rangle - \left\langle \mathbf{w}^*, \mathbf{w}^{(t)} \right\rangle \right)\\ &= \sum_{t=1}^{T} \langle \mathbf{w}^*, y_i \mathbf{x}_i \rangle \geq T \end{aligned} \tag{1.12}$$

b) Denominator:
We want to demonstate that: $\|\mathbf{w}^*\| \left\|\mathbf{w}^{(T+1)}\right\| \leq \sqrt{T} RB$.

- We have:

$$\begin{aligned}\left\|\mathbf{w}^{t+1}\right\| = \left\|\mathbf{w}^{(t)} + y_i \mathbf{x}_i\right\|^2 &= \left\|\mathbf{w}^{(t)}\right\|^2 + 2y_i \left\langle \mathbf{w}^{(t)}, \mathbf{x}_i \right\rangle + y_i^2 \|\mathbf{x}_i\|^2\\ &\leq \left\|\mathbf{w}^{(t)}\right\|^2 + R^2\end{aligned} \tag{1.13}$$

Therefore:

$$\begin{aligned}\left\|\mathbf{w}^{(T+1)}\right\| &= \sum_{t=1}^{T}\left(\left\|\mathbf{w}^{(t+1)}\right\|^2 - \left\|\mathbf{w}^{(t)}\right\|^2\right) \\ &= \sum_{t=1}^{T}\left(\left\|\mathbf{w}^{(t)} + y_i\mathbf{x}_i\right\|^2 - \left\|\mathbf{w}^{(t)}\right\|^2\right) \\ &= \sum_{t=1}^{T}\left(2y_i\left\langle\mathbf{w}^{(t)}, \mathbf{x}_i\right\rangle + \|\mathbf{x}_i\|^2\right) \leq TR^2 \qquad (1.14)\end{aligned}$$

▷ Combining the two previous results we obtain:

$$\frac{\left\langle\mathbf{w}^*, \mathbf{w}^{(T+1)}\right\rangle}{\|\mathbf{w}^*\|\left\|\mathbf{w}^{(T+1)}\right\|} \geq \frac{T}{BR\sqrt{T}} = \frac{\sqrt{T}}{BR} \qquad (1.15)$$

□

Remark 1. Convergence is guaranteed and it depends on B, which can be exponential in d.

1.2 Linear regression