Part VII: Learning from Data

VII.1 The Construction of Deep Neural Networks

A <u>neural network</u> is a model $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$ that makes a <u>prediction</u> $\hat{y} \in \mathbb{R}^k$ based on the <u>parameters</u> $x \in \mathbb{R}^n$ and the <u>feature vector</u> $v \in \mathbb{R}^m$:

$$\hat{y} = F(x, v).$$

Given the <u>training data</u>,

$$(v_i, y_i) \in \mathbb{R}^m \times \mathbb{R}^k, \quad i = 1, \ldots, N,$$

where v_i are <u>features</u> and y_i are <u>labels</u>, we optimize the parameters x so that the prediction of the model satisfies

$$y_i \approx F(x, v_i), \quad i = 1, \ldots, N.$$

In particular, neural networks are constructed from the following components:

① <u>Affine functions</u>: $v \mapsto Av + b$

② <u>Activation functions</u>: $w \mapsto \sigma(w)$

③ <u>Composition</u>: $F = F_L \circ \cdots \circ F_2 \circ F_1$

Note that the composition of affine functions is affine:

$$F_1(v) = A_1 v + b_1, \quad F_2(v) = A_2 v + b_2$$

$$\begin{aligned}
\Rightarrow \quad (F_2 \circ F_1)(v) &= F_2(F_1(v)) \\
&= F_2(A_1 v + b_1) \\
&= A_2(A_1 v + b_1) + b_2 \\
&= A_2 A_1 v + (A_2 b_1 + b_2) \\
&= A v + b,
\end{aligned}$$

where $A = A_2 A_1$ and $b = A_2 b_1 + b_2$.

The key to the modeling power of neural networks is the <u>activation function</u>.

The most popular activation function is the <u>rectified linear unit</u>:

$$\boxed{\mathrm{ReLU}(w) = [w]_+ = \begin{cases} 0 & \text{if } w \le 0 \\ w & \text{if } w > 0 \end{cases}}$$

Some other activation functions are:

(1) <u>Linear</u>:

$$\sigma(w) = a^T w + b,$$

(2) <u>Hyperbolic tangent</u>:

$$\sigma(w) = \tanh(w) \in (-1, 1),$$

(3) <u>Logistic</u> / <u>Sigmoid</u>:

$$\sigma(w) = \frac{1}{1 + e^{-w}} \in (0, 1).$$

Historically, tanh and sigmoid were preferred for their smoothness.

ReLU makes $v \mapsto F(x, v)$ a <u>continuous piecewise linear</u> function.