Definición de perceptrón

Consideremos un problema de aprendizaje supervisado con un conjunto de datos

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N},$$

donde $x^{(i)} \in \mathbb{R}^d$ representa el vector de características y $y^{(i)} \in \mathbb{R}$ la etiqueta asociada.

Un **perceptrón** se define mediante la composición de las siguientes funciones:

- **Función de combinación**

$$a\colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R},$$

Por ejemplo,

$$a(x, w, b) = w \cdot x + b.$$

- **Función de activación**

$$\sigma\colon \mathbb{R} \to \mathbb{R},$$

Por ejemplo,

$$\sigma(a) = a.$$

- **Función de salida**

$$h\colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R},$$

definida como la composición

$$\hat{y} = h(x, w, b) = \sigma(a(x, w, b)).$$

- **Función de pérdida**

$$L\colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}.$$

Por ejemplo,

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2.$$

Con esto,

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\big(y^{(i)} - h(x^{(i)}, w, b)\big)^2 = \frac{1}{N}\sum_{i=1}^{N}\big(y^{(i)} - \sigma(a(x^{(i)}, w, b))\big)^2.$$

El objetivo del aprendizaje es encontrar los parámetros (w^*, b^*) que minimizan la función de pérdida sobre el conjunto de datos:

$$(w^*, b^*) = \arg\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N} L\big(y^{(i)}, \sigma(a(x^{(i)}, w, b))\big).$$

En el ejemplo, esto se traduce en:

$$(w^*, b^*) = \arg\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - (w\cdot x^{(i)} + b)\right)^2,$$

lo cual es equivalente a una regresión lineal.

Entrenamiento

El entrenamiento del perceptrón implica ajustar los parámetros w y b utilizando un algoritmo de optimización, como el descenso por gradiente.

Para esto, calculamos las derivadas parciales de la función de pérdida con respecto a los parámetros:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial w} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \sigma}{\partial a}\cdot\frac{\partial a}{\partial w} = \frac{2}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})\cdot 1\cdot x^{(i)},$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial b} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \sigma}{\partial a}\cdot\frac{\partial a}{\partial b} = \frac{2}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})\cdot 0\cdot 1.$$

De esta manera, tomamos valores aleatorios iniciales para w y b, y actualizamos iterativamente:

$$\begin{pmatrix} w \\ b \end{pmatrix} \leftarrow \begin{pmatrix} w \\ b \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \end{pmatrix},$$

donde η es la tasa de aprendizaje. A cada actilización se le denomina **época**.

Perceptrón Sigmoide

Por simplicidad, podemos agregar a los datos una columna adicional con valor 1, de modo que el sesgo b se incorpore en el vector de pesos w. Así, en lugar de $a(x, w, b) = w\cdot x + b$, usamos

$$a(x, w) = w\cdot x.$$

Con esto, consideremos las siguientes funciones:

- Función de combinación: $a(x, w) = w\cdot x$.
- Función de activación sigmoide: $\sigma(a) = \dfrac{1}{1+e^{-a}}$.
- Función de salida: $\hat{y} = h(x, w) = \sigma(a(x, w))$.
- Función de pérdida (entropía cruzada):

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right].$$

Calculemos las derivadas parciales necesarias para el entrenamiento:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \sigma}{\partial a} \cdot \frac{\partial a}{\partial w}.$$

El valor de cada término es:

$$\begin{aligned}
\frac{\partial a}{\partial w} &= x, \\
\frac{\partial \sigma}{\partial a} &= \frac{e^{-a}}{(1+e^{-a})^2} = \sigma(a)(1-\sigma(a)) = \hat{y}(1-\hat{y}), \\
\frac{\partial L}{\partial \hat{y}} &= -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{y}^{(i)} - y^{(i)}}{\hat{y}^{(i)}(1-\hat{y}^{(i)})}\right).
\end{aligned}$$

Por lo tanto,

$$\frac{\partial L}{\partial w} = -\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})x^{(i)}.$$

Perceptrón Multiclase

Consideremos un problema de clasificación con K clases. Utilizamos la codificación one-hot para las etiquetas, es decir, $y^{(i)} \in \{0,1\}^K$ con un único 1 en la posición correspondiente a la clase correcta.

Definimos las siguientes funciones:

- Funciones de combinación: $a_k(x, w_k) = w_k \cdot x$ para $k = 1, \dots, K$.
- Función de activación softmax:

$$\sigma_k(a) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}}.$$

- Función de salida:

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_K) = (\sigma_1(a), \dots, \sigma_K(a)).$$

- Función de pérdida (entropía cruzada multiclase):

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)}).$$

Calculamos el gradiente de la función de pérdida respecto a los pesos w_k para cada clase k:

$$\begin{aligned}
\frac{\partial L}{\partial w} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial a} \cdot \frac{\partial a}{\partial w} \\
&= \frac{\partial L}{\partial \hat{y}} \cdot J_\sigma(a) \cdot \frac{\partial a}{\partial w},
\end{aligned}$$

o, de otra forma:

$$\begin{aligned}\frac{\partial L}{\partial w_k} &= \sum_{j=1}^{C} \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial w_k} \\ &= \sum_{j=1}^{C} \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial a_k} \frac{\partial a_k}{\partial w_k}\end{aligned}$$

El valor de cada derivada parcial es:

$$\begin{aligned}\frac{\partial a_k}{\partial w_k} &= x, \\ \frac{\partial \hat{y}_j}{\partial a_k} &= J_\sigma(a)_{jk} = \sigma_k(a)(\delta_{jk} - \sigma_j(a)), \\ \frac{\partial L}{\partial \hat{y}_j} &= -\frac{y_j}{\hat{y}_j}.\end{aligned}$$

Por lo tanto, el gradiente de la función de pérdida respecto a los pesos w_k es:

$$\frac{\partial L}{\partial w_k} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_k^{(i)} - y_k^{(i)} \right) x^{(i)}.$$