Dimostrazione della non influenza del coefficiente η, quando l'algoritmo Perceptron viene inizializzato con il vettore dei pesi nullo

Federico Zanardo

28 Ottobre 2020

1 Perceptron

1.1 Introduzione

Un **perceptron** è un classificatore *binario* che prende come input un vettore di valori e restituisce 1 se il risultato è maggiore rispetto ad una certa *soglia*, -1 altrimenti. Più precisamente, dati i valori di input $x_1, \ldots, x_n$ (che indicheremo con $\vec{x}$), l'output σ sarà:

$$\sigma(\vec{w}, \vec{x}) = \begin{cases} 1 & \text{se } w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n > 0 \\ -1 & \text{altrimenti} \end{cases}$$

dove $\vec{w}$ sono dei pesi. In particolare, la funzione σ può essere riscritta come:

$$\sigma(\vec{w}, \vec{x}) = sign(\vec{w} \cdot \vec{x})$$

dove

$$sign(z) = \begin{cases} 1 & \text{se } z > 0 \\ -1 & \text{altrimenti} \end{cases}$$

L'apprendimento di un perceptron implica la scelta dei valori per i pesi $\vec{w}$.

1.2 Algoritmo

Sia l'*insieme di apprendimento* $\mathcal{S} = \{x, t\}$, con $x \in \mathbb{R}^{(n+1)}$ e $t \in \{-1, +1\}$. Sia $\eta > 0$ il *learning rate*, ovvero, un coefficiente per rendere più stabile l'apprendimento, in modo da evitare che il vettore dei pesi $\vec{w}$ possa subire delle brusche variazioni durante l'aggiornamento dei pesi stessi.

L'algoritmo esegue i seguenti passi:

1. Inizializza il vettore dei pesi $\vec{w}$ con dei valori randomici;
2. Ripeti
   (a) Seleziona un esempio $(x, t) \in \mathcal{S}$;
   (b) Se $o = sign(\vec{w} \cdot \vec{x}) \neq t$, allora

$$\vec{w} = \vec{w} + \eta \cdot (t - o) \cdot \vec{x}$$

2 Teorema

In questa relazione, si vuole dimostrare la non influenza del learning rate η nell'apprendimento dell'algoritmo Perceptron, quando il vettore iniziale dei pesi $w^{(0)}$ è nullo.

Teorema: Sia l'insieme di apprendimento $\mathcal{S} = \{x, t\}$, $x \in \mathbb{R}^{(n+1)}$, $t \in \{-1, +1\}$, il learning rate $\eta > 0$ e $\vec{w}^{(0)} \in \mathbb{R}^{(n+1)}$ il vettore iniziale dei pesi. Se il vettore iniziale dei pesi $\vec{w}^{(0)} = 0$ (nullo), allora il learning rate η non influenza l'apprendimento.

2.1 Dimostrazione

Per induzione si dimostri che per un qualsiasi vettore dei pesi $\vec{w}$ dopo k passi di apprendimento dell'algoritmo Perceptron, per ogni $k >= 1$, $\vec{w}^{(k)} \in \mathbb{R}^{(n+1)}$ e $\vec{r}^{(k)} \in \mathbb{R}^{(n+1)}$, se $\vec{w}^{(0)}$ è nullo, allora $\vec{w}^{(k)} = \eta \cdot \vec{r}^{(k)}$ e $\vec{r}^{(k)}$ non dipende da η.

1. **Caso base**: $k = 1$

$$\begin{aligned}\vec{w}^{(1)} &= \vec{w}^{(0)} + \eta(t^{(1)} - sign(\vec{w}^{(0)} \cdot \vec{x}^{(1)}) \cdot \vec{x}^{(1)}) \\ &= 0 + \eta(t^{(1)} - 0) \cdot \vec{x}^{(1)} \\ &= \eta \cdot t^{(1)} \cdot \vec{x}^{(1)}\end{aligned}$$

   Quindi,

$$\vec{w}^{(1)} = \eta \cdot \vec{r}^{(1)}$$

   con $\vec{r}^{(1)} = t^{(1)} \cdot \vec{x}^{(1)}$. É possibile notare come $\vec{r}^{(1)}$ non dipende dal learning rate η. Quindi, la proprietà è valida nel caso base;

2. **Passo induttivo**: assumo che la proprietà valga per k. Dimostro che la proprietà continui a valere anche per $k + 1$.

   (a) **Ipotesi induttiva**: per un qualsiasi vettore di pesi $\vec{w} \in \mathbb{R}^{(n+1)}$, per ogni $k >= 1$ e $\vec{r}^{(k)} \in \mathbb{R}^{(n+1)}$, $\vec{w}^{(k)} = \eta \cdot \vec{r}^{(k)}$ e $\vec{r}^{(k)}$ non dipende da η;

   (b) **Tesi**:

$$\vec{w}^{(k+1)} = \vec{w}^{(k)} + \eta(t^{(k+1)} - sign(\vec{w}^{(k)} \cdot \vec{x}^{(k+1)})) \cdot \vec{x}^{(k+1)}$$

   Applicazione dell'ipotesi induttiva:

$$\begin{aligned}&= \eta \cdot \vec{r}^{(k)} + \eta(t^{(k+1)} - sign(\eta \cdot \vec{r}^{(k)} \cdot \vec{x}^{(k+1)})) \cdot \vec{x}^{(k+1)} \\ &= \eta(\vec{r}^{(k)} + (t^{(k+1)} - sign(\eta \cdot \vec{r}^{(k)} \cdot \vec{x}^{(k+1)})) \cdot \vec{x}^{(k+1)}) \\ &= \eta \cdot \vec{r}^{(k+1)}\end{aligned}$$

   con

$$\vec{r}^{(k+1)} = \vec{r}^{(k)} + (t^{(k+1)} - sign(\eta \cdot \vec{r}^{(k)} \cdot \vec{x}^{(k+1)})) \cdot \vec{x}^{(k+1)}$$

   in quanto il termine $sign(\eta \cdot \vec{r}^{(k)} \cdot \vec{x}^{(k+1)})$ è sempre maggiore di zero e pertanto il learning rate η non influisce sull'apprendimento in alcun modo. Quindi,

$$\vec{w}^{(k+1)} = \eta \cdot \vec{r}^{(k+1)}$$

   Giunti a questo punto, $\vec{r}^{(k+1)}$ non dipende in alcun modo da η. Pertanto, la proprietà è stata dimostrata anche per il passo induttivo.

$\square$