Perceptron ciągły i dyskretny

Jacek Kluska

Politechnika Rzeszowska

Klasyfikator ciągły

Perceptron z ciągłą funkcją aktywacji - wejścia $(x_1, \ldots x_N)$, wyjście y

$$y = f\left(n\right); \qquad n = \langle \mathbf{w}, \mathbf{x} \rangle + b = w_1 x_1 + \ldots + w_N x_N + b$$

f - różniczkowalna, np.
Znany jest zbiór par uczących $\left\{ \left(\mathbf{x}_1, d_1\right), \left(\mathbf{x}_2, d_2\right), \ldots, \left(\mathbf{x}_Q, d_Q\right) \right\}$.

- $y = f\left(n\right) = \dfrac{2}{1 + e^{-\lambda n}} - 1, \qquad \lambda > 0,\ y \in (-1, 1)$ - bipolarna
- $y = f\left(n\right) = \dfrac{1}{1 + e^{-\lambda n}}, \qquad \lambda > 0,\ y \in (0, 1)$ - unipolarna

Dane: Zbiór par uczących $\left\{ \left(\mathbf{x}_1, d_1\right), \left(\mathbf{x}_2, d_2\right), \ldots, \left(\mathbf{x}_Q, d_Q\right) \right\}$, gdzie $d_i \in \{-1,\ 1\}$ lub $d_i \in \{0,\ 1\}$.
Błąd

$$E\left(k\right) = \frac{1}{2}\left[d\left(k\right) - y\left(k\right)\right]^2, \qquad k = 0, 1, 2, \ldots$$

Problem

Znaleźć taki algorytm uczenia $\underline{w}$ i b, aby

$$E\left(k\right) \longrightarrow \min \quad , \qquad \forall\, k$$

- Gradienty

$$\nabla_{\mathbf{w}} E = \frac{\partial E}{\partial y}\frac{\partial y}{\partial n}\nabla_{\mathbf{w}} n = \frac{\partial E}{\partial y}\frac{\partial y}{\partial n}\begin{bmatrix} \dfrac{\partial n}{\partial w_1} \\ \vdots \\ \dfrac{\partial n}{\partial w_N} \end{bmatrix}$$

$$\nabla_b E = \frac{\partial E}{\partial y}\frac{\partial y}{\partial n}\frac{\partial n}{\partial b}$$

Algorytm uczenia perceptronu - metoda "delta"

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta \nabla_{\mathbf{w}} E$$
$$b(k+1) = b(k) - \eta \nabla_{b} E \,, \qquad \eta > 0$$

- f - bipolarna

$$\frac{\partial f}{\partial n} = \frac{\partial}{\partial n}\left(\frac{2}{1+e^{-\lambda n}} - 1\right) = \frac{\lambda}{2}\left(1 - y^2\right)$$

- f - unipolarna

$$\frac{\partial f}{\partial n} = \frac{\partial}{\partial n}\left(\frac{1}{1+e^{-\lambda n}}\right) = \lambda y\left(1 - y\right)$$

Algorytm uczenia perceptronu dla funkcji bipolarnej i unipolarnej

Algorytm uczenia perceptronu dla funkcji bipolarnej

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta\frac{\lambda}{2}\left[d(k) - y(k)\right]\left[1 - y^2(k)\right]\mathbf{x}(k)$$
$$b(k+1) = b(k) + \eta\frac{\lambda}{2}\left[d(k) - y(k)\right]\left[1 - y^2(k)\right]$$

Algorytm uczenia perceptronu dla funkcji unipolarnej

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta\lambda\left[d(k) - y(k)\right]y(k)\left[1 - y(k)\right]\mathbf{x}(k)$$
$$b(k+1) = b(k) + \eta\lambda\left[d(k) - y(k)\right]y(k)\left[1 - y(k)\right]$$

Uwaga: Uczenie klasyfikatora **ciągłego** nawet dla problemów liniowo separowalnych nie zawsze prowadzi do poprawnej klasyfikacji (Wittner, Denker, 1988).

Klasyfikator dyskretny

Perceptron ze skokową funkcją aktywacji - wejścia $(x_1, \ldots x_N)$, wyjście y

$$y = f(n); \qquad n = \langle \mathbf{w}, \mathbf{x} \rangle + b = w_1 x_1 + \ldots + w_N x_N + b$$

Funkcja aktywacji

$$y = f(n) = sign(n) = \begin{cases} 1 & \Leftrightarrow n \geqslant 0 \\ -1 & \Leftrightarrow n < 0 \end{cases}$$

Dane: Zbiór par uczących $\{(\mathbf{x}_1, d_1), (\mathbf{x}_2, d_2), \ldots, (\mathbf{x}_Q, d_Q)\}$, gdzie $d_i \in \{-1, 1\}$ lub $d_i \in \{0, 1\}$.
Błąd

$$E(k) = \frac{1}{2}\left[d(k) - y(k)\right]^2, \qquad k = 0, 1, 2, \ldots$$

Problem

Znaleźć taki algorytm uczenia $\underline{w}$ i b, aby

$$E(k) = 0 \ , \qquad \forall k \geqslant k_0$$

Algorytm uczenia perceptronu dyskretnego

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \frac{1}{2} c \left[d(k) - y(k) \right] \mathbf{x}(k)$$
$$b(k+1) = b(k) + \frac{1}{2} c \left[d(k) - y(k) \right]$$

<u>**Wskazówka**</u>: współczynnik korekcji

$$c(k) = c_0 \frac{\left| \langle \mathbf{w}(k), \mathbf{x}(k) \rangle \right|}{\left\| \mathbf{x}(k) \right\|^2} \ , \qquad c_0 > 0$$

Uwaga: Uczenie klasyfikatora <u>**dyskretnego**</u> według wzoru jak wyżej, dla problemów liniowo separowalnych - zawsze prowadzi do poprawnej klasyfikacji w skończonej liczbie kroków.

Dowód zbieżności reguły uczenia klasyfikatora dyskretnego

- Funkcja aktywacji każdego neuronu jest typu "signum": $y \in \{-1, 1\}$.
- Mamy 2 klasy punktów: K_1 i K_2.

$$\mathbf{x} \in K_1 \Leftrightarrow \langle \mathbf{w}, \mathbf{x} \rangle \leqslant 0 \Leftrightarrow \operatorname{sign} \langle \mathbf{w}, \mathbf{x} \rangle = -1$$

$$\mathbf{x} \in K_2 \Leftrightarrow \langle \mathbf{w}, \mathbf{x} \rangle > 0 \Leftrightarrow \operatorname{sign} \langle \mathbf{w}, \mathbf{x} \rangle = +1$$

- Fakt:

$$\mathbf{x} \in K_2 \Rightarrow (-\mathbf{x}) \in K_1$$

ponieważ $\langle \mathbf{w}, \mathbf{x} \rangle > 0 \Rightarrow \langle \mathbf{w}, (-\mathbf{x}) \rangle < 0$

- **Modyfikujemy ciąg uczący:** eliminujemy te $\mathbf{x}$, które nie powodują zmiany wag. Jeżeli $\mathbf{x} \in K_1$ i $[d(k) - y(k)] = [1 - (-1)] = 2$. Wtedy $\mathbf{w}(k+1) = \mathbf{w}(k) + c\mathbf{x}(k)$, $c > 0$. Nie będziemy podawać w trakcie uczenia punktów $\mathbf{x} \in K_2$, lecz $(-\mathbf{x}) \in K_1$. Zatem $\mathbf{w}(k+1) = \mathbf{w}(k) + c\mathbf{x}(k)$. **Zmodyfikowany ciąg uczący**:

$$K_1' = K_1 \cup \{(-\mathbf{x}) : \ \mathbf{x} \in K_2\}$$

Dowód zbieżności reguły uczenia klasyfikatora dyskretnego - założenia i fakty

Założenia:

1. Problem jest liniowo separowalny:

$$\exists \mathbf{w}^* = \mathbf{w}^{opt} : \qquad \langle \mathbf{w}^*, \mathbf{x} \rangle > 0, \qquad \forall \mathbf{x}$$

2. Bez utraty ogólności: $\mathbf{w}(1) = \mathbf{0}$.

Fakt 1

- wynika z modyfikacji ciągu uczącego i stąd, że $\mathbf{w}(1) = \mathbf{0}$:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + c\mathbf{x}(k), \qquad \forall \mathbf{x} \in K_1'$$

$$\Downarrow$$

$$\mathbf{w}(k+1) = c\mathbf{x}(1) + c\mathbf{x}(2) + ... + c\mathbf{x}(k)$$

Dowód zbieżności reguły uczenia klasyfikatora dyskretnego - fakty

Fakt 2

- wynika z liniowej separowalności:

$$\exists \delta : \qquad \inf_{\mathbf{x}} \langle \mathbf{w}^*, \mathbf{x} \rangle = \delta > 0$$

Rozpatrzmy $\langle \mathbf{w}^*, \mathbf{w}(k+1) \rangle$:

$$\begin{aligned} \|\mathbf{w}^*\| \, \|\mathbf{w}(k+1)\| &\geqslant \langle \mathbf{w}^*, \mathbf{w}(k+1) \rangle \\ &= \langle \mathbf{w}^*, c\mathbf{x}(1) \rangle + \langle \mathbf{w}^*, c\mathbf{x}(2) \rangle + \ldots + \langle \mathbf{w}^*, c\mathbf{x}(k) \rangle \\ &\geqslant ck\delta \end{aligned}$$

$$\|\mathbf{w}(k+1)\|^2 \geqslant \frac{c^2 k^2 \delta^2}{\|\mathbf{w}^*\|^2}$$

Dowód zbieżności reguły uczenia klasyfikatora dyskretnego - fakty, c.d.

Z drugiej strony:

$$\begin{aligned}\left\|\mathbf{w}(k+1)\right\|^2 &= \left\langle \mathbf{w}(k) + c\mathbf{x}(k), \mathbf{w}(k) + c\mathbf{x}(k)\right\rangle \\ &= \left\|\mathbf{w}(k)\right\|^2 + 2c\left\langle \mathbf{w}(k), \mathbf{x}(k)\right\rangle + c^2\left\|\mathbf{x}(k)\right\|^2\end{aligned}$$

Korekcja wag następuje dopóty, dopóki $\left\langle \mathbf{w}(k), \mathbf{x}(k)\right\rangle \leqslant 0$, więc

$$\begin{aligned}\left\|\mathbf{w}(k+1)\right\|^2 &\leqslant \left\|\mathbf{w}(k)\right\|^2 + c^2\left\|\mathbf{x}(k)\right\|^2 \\ &\leqslant \left\|\mathbf{w}(k)\right\|^2 + c^2M^2 \leqslant kc^2M^2\end{aligned}$$

gdzie $M = \sup\limits_{\mathbf{x}} \left\|\mathbf{x}\right\|$. Czyli

$$kc^2M^2 \geqslant \left\|\mathbf{w}(k+1)\right\|^2 \geqslant \frac{c^2k^2\delta^2}{\left\|\mathbf{w}^*\right\|^2} \Rightarrow k \leqslant \left(\frac{M\left\|\mathbf{w}^*\right\|}{\delta}\right)^2.$$