Mesterséges intelligencia II.

Perceptron modell

Legyen adott egy $S = \{(\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l)\} \subseteq \mathbb{R}^d \times \{0, 1\}$ tanítóhalmaz. A perceptron modell célja a tanulás során, hogy az $S$ tanítóhalmaz alapján meghatározza azt a $\bar{w}_{opt}$ vektort, melyre a $\bar{w}_{opt}^T \bar{x} + b$ hipesík a lehető legjobban szeparál, azaz azt a hipersíkot, mely a lehető legjobban elválasztja egymástól a különböző címkével rendelkező pontokat. Ezt a tanító halmazon mért tévesztési hiba minimalizálásával teszi. A levezetés előtt, vezessük be, hogy $x_0 = 1$ és $w_0 = -b$. Ekkor a fenti hípersík egyenlete a $\bar{w}^T \bar{x}$ alakú lesz, ahol

$$\bar{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}, \text{ illetve } \bar{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{pmatrix}.$$

A tanítás után – vagy során, amennyiben egy kezdeti $\bar{w}$ vektor rendelkezésre áll – a modell használata rendkívül egyszerű, egy új osztályozandó példa esetén azt kell eldönteni, hogy a $\bar{w}$ által meghatározott hipersík melyik oldalára esik az adott bemenet. Ehhez, a bemenet alapján elkészítjük a korábban megfogalmazott $d+1$ dimenziós, $\bar{x}$-ként hivatkozott alakját a bemenetnek (hozzátesszük nulladik komponensként az $x_0$-at). Ezután kiszámoljuk a $\bar{w}^T \bar{x}$ értéket. Ha ez nulla vagy pozitív, akkor az hipersíkon vagy a hipersík felett van az adott pont, így ennek megfelelően címkézzük, különben a hipesík alatt van az adott pont, tehát ennek megfelelő címkével kell ellátni. Ennek megfelelően a $f$ diszkriminancia függvény több féle lehet:

$$f(\bar{x}) = sign(\bar{w}^T \bar{x}), \text{ vagy } f(\bar{x}) = sigm(\bar{w}^T \bar{x}),$$

ahol a $sign(x)$ az előjel függvény, $sigm(x) = \frac{1}{1+e^{-cx}}$, ahol $c$ konstans. (A $sigm(x)$ (szigmoid) függvény az előjel függvény folytonos közelítésének tekinthető, ez akkor lesz fontos, amikor az $f$ deriváltjára is szükség lesz)

Az optimális $\bar{w}$ meghatározása egy egyszerű tanítóalgoritmus sémával történik:

$i \leftarrow 0$
$\bar{w}^{(0)} \leftarrow$ véletlen választás
**while** nem konvergál **do**
  $i \leftarrow i + 1$
  $\bar{w}^{(i)} \leftarrow \bar{w}^{(i-1)} + \Delta^{(i-1)}$
**end while**

Az előbb leírt algoritmus sémából konkrét algoritmusokat kapunk, ha megadjuk az iterációnkénti elmozdulást ($\Delta^{(i-1)}$-t):

*A perceptron tanulási szabály (delta szabály lineáris kimenet esetén)*

Ezesetben a

$$\Delta^{(i-1)} = -\eta(y_j - f(\bar{x}_j))\bar{x}_j,$$

ahol a $0 < \eta < 1$ konstans, az $(\bar{x}_j, y_j)$ a tanítóhalmaz véletlen eleme, $f$ a modell döntési függvénye, amely az előző iteráció beli, $\bar{w}^{(i-1)}$ vektort használja. Ebben az esetben az $f$ előjel függvényt használó változatát használtuk.

*A sztochasztikus grádiens módszer*

A leggyakrabban használt változat, melyben a

$$\Delta^{(i-1)} = -\eta \nabla E(\bar{w}^{(i-1)}),$$

ahol a $0 < \eta < 1$ szintén konstans, a $\nabla E(\bar{w}^{(i-1)})$ pedig az $E$ hiba függvény parciális deriváltjait tartalmazó vektor, a hiba függvény a következő:

$$E(\bar{w}^{(i-1)}) = \frac{1}{2} \left( f(\bar{x}_j) - y_j \right)^2,$$

ahol az $(\bar{x}_j, y_j)$ a tanítóhalmaz véletlen eleme, $f$ pedig a döntési függvény szigmoidot használó változata, mely a $\bar{w}^{(i-1)}$ vektort használja.

Az $E$ függvény fent leírt változata esetén a

$$\nabla E(\bar{w}^{(i-1)}) = \begin{pmatrix} \frac{\partial E\left(\bar{w}^{(i-1)}\right)}{\bar{w}_0^{(i-1)}} \\ \frac{\partial E\left(\bar{w}^{(i-1)}\right)}{\bar{w}_1^{(i-1)}} \\ \vdots \\ \frac{\partial E\left(\bar{w}^{(i-1)}\right)}{\bar{w}_d^{(i-1)}} \end{pmatrix},$$

ahol tetszőleges $0 \geq k \geq d$ esetén

$$\frac{\partial E\left(\bar{w}^{(i-1)}\right)}{\partial \bar{w}_k^{(i-1)}} = \left( f(\bar{x}^{(j)}) - y_j \right) f(\bar{x}^{(j)}) \left( 1 - f(\bar{x}^{(j)}) \right) \bar{x}_k^{(i-1)}.$$