Perceptron konvergencia tétel

Mesterséges Intelligencia I. házi feladat

Lám István
A2D587
lam.istvan@gmail.com

1 A Perceptron

1.1 A perceptron definíciója

A perceptron az egyik legegyszerűbb előrecsatolt neurális hálózat (1), feltalálója Frank Rosenblatt, amerikai informatikus-kutató. A működését tekintve úgy írható le, mint egy bináris függvény, aminek a hozzárendelési szabálya:

$$f(\overline{x}) = \begin{cases} 1 & ha\ \overline{w} \cdot \overline{x} + b > 0 \\ 0 & egyébként \end{cases}$$

$$ahol\ f: \mathbb{R}^n \to \{0,1\}, és\ \overline{w}, \overline{x} \in \mathbb{R}^n, b \in \mathbb{R}$$

Továbbiakban a ***w*** vektorra mint súlyvektorra, a b számra pedig mint előfeszültség fogunk hivatkozni.

Látható, hogy a perceptron mint egy osztályozó képes működni: képes eldönteni, hogy $\overline{x}$ benne van-e egy halmazban ($f(\overline{x}) = 1$), vagy sem ($f(\overline{x}) = 0$).

A perceptron egy fontos tulajdonsága, hogy „tanítható". A tanításhoz viszont szükség van egy példahalmazra, amely tartalmaz bizonyos bemeneti mintákat ($\{\overline{x}_1, \overline{x}_2, \dots, \overline{x}_m\}$) és az ezekre elvárt kimenetet ($y_1, y_2, \dots, y_m, \forall i \in \{1,2 \dots, m\}\ y_i \in \{0,1\}$). Szemléletesen: a perceptronnak „be kell mutatni" elemeket, és megmondani, hogy benne van-e egy halmazban vagy sem. A tanítás során iterálva közelítjük a helyes $\overline{w}$ vektort és a b előfeszültséget. Az iterálás (vagy másképpen, tanítási algoritmus) viszonylag egyszerű, és az esetek egy részében véges lépésben véget is ér. Sajnos, mint látni fogjuk, ez nem minden esetre áll fenn, bár kezdetben nagy lelkesedéssel fogadták a tudósok, mert remélték, hogy egy univerzális módszert kaptak bármilyen bináris kimenetű függvény leírására. Sajnos a reményük korán szertefoszlott…

1.2 Tanítási algoritmus

Ahogy az előző részben jeleztük, a perceptron viszonylag könnyen tanítható. Pillanatra tekintsünk el a b előfeszültségtől, így mutatjuk be a tanítási algoritmust. Legyen

$$In = \{\overline{x}_1, \overline{x}_2, \dots, \overline{x}_m\}, \forall i \in \{1,2, \dots, m\}\ x_i \in \mathbb{R}^n$$

a bemeneti minták halmaza,

$$Out = \{y_1, y_2, \dots, y_m, \forall i \in \{1,2 \dots, m\}\ y_i \in \{0,1\}$$

pedig az egyes bemenetekre elvárt válasz. Azaz, azt szeretnénk, hogy $\forall i \in \{1,2 \dots, m\}\ f(\overline{x}_i) = y_i$ ahol $f: \mathbb{R}^n \to \{0,1\}$ a perceptron hozzárendelési szabálya.

A tanítás során iteráljuk a $\overline{w} \in \mathbb{R}^n$ súlyvektort, a lent leírt algoritmussal. Az i-edik iteráció során meghatározott vektort $\overline{w}_i$-vel fogjuk jelölni, és az így iterált függvényt pedig $f_i(\overline{x})$-vel.

1. Kezdetben legyen $\overline{w}_0\ j = rand \quad j \in \{1..n\}$, azaz egy random kis szám. Legyen i=0
2. $\overline{w}_{i+1} = \overline{w}_i + \alpha\ (y_i - f_i(\overline{x}_i))\ \overline{x}_i$, ahol $\alpha \in]0,1]$ az úgynevezett tanulási ráta, amit kvázi szabadon megválaszthatunk, és állandó
3. Növeljük i-t eggyel, és ha i kisebb mint a megengedett maximum iteráció és a hiba nagyobb mint a beállított hibahatár, akkor ismételjük az eljárást a 2. ponttól kezdve.

Könnyű látni, hogy a fenti algoritmusba a b előfeszültséget egy plusz tag hozzá vételével kaphatjuk: vegyünk fel minden bemeneti minta vektorhoz egy plusz dimenziót, azaz $\overline{x}_i[n+1] = 1$, és akkor az eredményben $b = \overline{w}[n+1]$ (azaz, a $\overline{w}$ vektor n+1-edik koordinátája) lesz az előfeszültség.

A fenti algoritmusra mint az egyik legegyszerűbb tanulási algoritmusra gondolhatunk. Mint látni fogjuk, bizonyos típusú (lineárisan szeparálható) példahalmazokra a fenti eljárás véges lépés után leáll, és megtalálja a kívánt függvényt. Ezen kívül sok más algoritmust dolgoztak ki ((2), (3), (4)) amelyek közül Gallant (2) „zsebes" algoritmusát emelném ki. Ez az algoritmus sokkal nagyobb adathalmazokra is hatékonyan működik, továbbá kezeli a zajos (azaz hibákat is tartalmazó), és ellentmondásokat is tartalmazó adathalmazokat. További előnye, hogy nem csak a lineárisan szeparálható adatokat képes kezelni. A perceptron konvergencia tétel mindazon által azt fogalmazza meg, hogy lineárisan szeparálható adathalmazokon még a fent ismertetett egyszerű algoritmus esetén is véges lépés után kapunk egy helyes $\overline{w}$-t.

2 Perceptron konvergencia tétel

2.1 A tétel kimondása

2.1.1 Definíció: lineáris szeparálhatóság (5)

Legyen $D \subseteq \mathbb{R}^n$. Legyen D két diszjunkt részhalmaza X_0 és X_1 (azaz $\exists X_0, X_1 \subseteq D, X_0 \cap X_1 = \emptyset$).

D lineárisan szeparálható X_0 és X_1 halmazokra, hogyha:

$$\exists \overline{w} \in \mathbb{R}^n, \exists h \in \mathbb{R}:$$

$$\forall \overline{x} \in X_0 \quad \overline{w} \cdot \overline{x} \geq h \quad \wedge (\forall \overline{x^*} \in X_1 \; \overline{w} \cdot \overline{x^*} < h)$$

ahol '·' a skaláris szorzás $\mathbb{R}^n$ felett. Másképpen fogalmazva:

$$\exists \overline{w} \in \mathbb{R}^n, \exists h \in \mathbb{R}:$$

$$\forall \overline{x} \in X_0 \quad \sum_{i=1}^{n} \overline{w}\, i \; * \overline{x}\, i \; \geq h \quad \wedge \quad \forall \overline{x^*} \in X_1 \quad \sum_{i=1}^{n} \overline{w}\, i \; * \overline{x^*}\, i \; < h$$

2.1.2 Tétel: perceptron konvergencia tétel:

Legyen

$$f_w\ \overline{x}\ = \begin{cases} 1 & ha\ \overline{w} \cdot \overline{x} > 0 \\ 0 & egyébként \end{cases}$$

$$ahol\ f_w \colon \mathbb{R}^n \rightarrow \ 0{,}1\ , és\ \overline{w}, \overline{x} \in \mathbb{R}^n, b \in \mathbb{R}$$

továbbá $D \subseteq \mathbb{R}^n$ lineárisan szeparálható ***példahalmaz***, melyre az ***elvárt kimenet*** $Out = \{y_1, y_2, \dots, y_m\}$. Tegyük fel, hogy

$$\exists \overline{w^*} \in \mathbb{R}^n \colon \forall i \in \ 1{,}2, \dots, m\ , \overline{x_i} \in D \quad f_{w*}\ \overline{x_i}\ = y_i$$

Azaz, létezik olyan súlyozás, amivel a példahalmaz bemeneteire a függvény az elvárt választ adja. Ekkor a perceptron tanulási algoritmusa véges lépésen belül talál egy olyan $\overline{w} \in \mathbb{R}^n$ vektort, amire $f_w(\overline{x})$ az elvárt kimenetet adja. Azaz, véges lépésen belül konvergens

Megjegyzés: elhagytuk a *b* előfeszültséget; de a 1.2 pontban leírtak miatt minden továbbiakban ismertetett állítás igaz lesz az előfeszültséggel együtt tárgyalt *f* függvényre

Bizonyítás (eredetileg Rosenblatt (6) bizonyítása, további források (7), (8)):

1. Ha nem történt hiba, akkor $\overline{w}_{i+1} = \overline{w}_i$ (1.2 alfejezet, 2. pont a felsorolásban)
2. Legyen $j_1, j_2, \dots, j_k$ egy olyan indexsorozat, amelyre hiba történt
3. Ekkor rekurzívan visszafejtve $\overline{w}_{j_i+1} = \ \sum_{i=1}^{k} \alpha\ \ y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i}$
4. Megmutatható, hogy hiba esetén $\overline{w_i}\ \alpha\ \ y_i - f_i\ \overline{x_i}\ \ \overline{x_i} = -\alpha |\overline{w_i} \cdot \overline{x_i}| < 0$
5. Mivel D lineárisan szeparálható, $\exists \overline{w_0} \in \mathbb{R}^n \colon \forall \overline{x} \in X_0 \subseteq D\ \overline{w_0} \cdot \overline{x} \geq h > 0$
6. Az előzőek miatt $\overline{w_0} \cdot \overline{w}_{j_i+1} = \ \sum_{i=1}^{k} \alpha\ \ y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{w_0} \cdot \overline{x}_{j_i} \geq kh$
7. A Cauchy-Schwarz-Bunyakovsky egyenlőtlenség miatt (9)

$$\overline{w_0}\ ^2\ \overline{w_{j_i+1}}\ ^2 \geq\ \overline{w_0} \cdot \overline{w}_{j_i+1}\ ^2$$

Az 6-sben és a fentebb leírtak miatt:

$$\overline{w_0}\ ^2\ \overline{w_{j_i+1}}\ ^2 \geq\ \overline{w_0} \cdot \overline{w}_{j_i+1}\ ^2 \geq\ kh\ ^2$$

Azaz

$$\left\|\overline{w_{j_i+1}}\right\|^2 \geq \frac{kh^2}{\left\|\overline{w_0}\right\|^2}$$

8. Továbbá,

$$\overline{w}_{i+1} = \overline{w}_i + \alpha\ (y_i - f_i\ \overline{x}_i\ \ \overline{x}_i$$

Így akkor:

$$\left\|\overline{w_{j_i+1}}\right\|^2 = \left\|\overline{w}_{j_i} + \alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i}\right\|^2 =$$

$$\left\|\overline{w_{j_i}}\right\|^2 + 2\overline{w_{j_i}}\,\alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i} + \alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i}\right\|^2$$

REF _Ref279440492 \r \h 4.pontban leírtak miatt

$$\left\|\overline{w_{j_i}}\right\|^2 + 2\overline{w_{j_i}}\,\alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i} + \alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i}\right\|^2 \leq \left\|\overline{w_{j_i}}\right\|^2 + \alpha\ (y_{j_i} - f_{j_i}\ \overline{x}_{j_i}\ \ \overline{x}_{j_i}\right\|^2 \leq k\beta^2$$

ahol

$$\beta^2 \triangleq \max\ \left\|\alpha\ (y_i - f_i\ \overline{x}_i\ \ \overline{x}_i\right\|^2, i \in j_1, j_2, \ldots, j_k$$

9. Így akkor

$$\frac{kh^2}{\left\|\overline{w_0}\right\|^2} \leq \left\|\overline{w_{j_i+1}}\right\|^2 \leq k\beta^2$$

$$0 \leq k \leq \frac{\beta^2\ \left\|\overline{w_0}\right\|^2}{h^2}$$

Azaz, korlátos j_k indexsorozatra befejeződik az algoritmus ■

Megjegyzés: a bizonyításban $\overline{w_{j_i+1}}$ helyett írhattunk volna $\overline{w_{j_{i+1}}}$-et is, azaz, a következő olyan elem, amire a hiba nem nulla, nem változtatott volna a bizonyítás menetén.

2.1.3 További bizonyítások

További bizonyítások találhatók Novikoff (10),Minksy és Papert (11) és később (12), stb. cikkeiben.

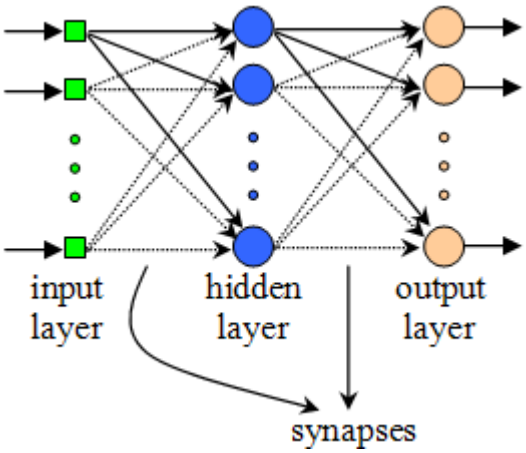
3 Nem konvergens esetek

Bár a perceptron konvergencia tétel tévesen azt sugallhatja, hogy innentől bármilyen függvényt képesek leszünk megtanítani ennek a mesterséges neuronnak, van egy óriási bökkenő: a perceptron tétel bizonyításánál felhasználtuk, hogy a.) létezik jó súlyozás b.) lineárisan szeparálható adathalmazt eredményez a megtanítandó függvény. Ez bizony igen erős feltételezés, de sokáig nem derült ki a kutatók előtt, hogy tényleg komoly akadályokat gördít a perceptronnal kapcsolatos kutatások elé. Minksy és Papert rámutatott, hogy bizony léteznek olyan függvények, amit a „klasszikus" perceptronnak nem lehet megtanítani, pl.: bizonyítottan nem lehet a XOR-t megvalósítani (11) perceptronnal. Maga a könyv (11) egyébként az egyik legteljesebb és legismertebb könyv a témában, mégis sokan támadták a szerzőket azért, mert nem a művükben nem utaltak arra, hogy ilyen probléma bizony csak az egyrétegű neurális hálózatokban fordul elő (13). Ez többek szerint helytelen támadás, ugyanis a könyv írásakor már ismeretes volt a többrétegű hálózatok előnyei. Ennek ellenére később a könyvet bővített (javított) formában ismét kiadták (12).

Bár eddig csak a teljes konvergenciát tárgyaltuk, bizonyos esetekben viszont elegendő, hogy egy függvényt csak jól tudunk becsülni. Bár ahogy láttuk, perceptronnal nem lehet minden $f: \mathbb{R}^n \to \{0,1\}$ függvényt leírni, az bizonyított, hogy approximálni bármilyen $f: \mathbb{R} \to -1,1$ folytonos függvényt lehet (14) még egyrétegű hálózattal is (természetesen itt már nem a „klasszikus" perceptront használják a szerzők).

3.1 Többrétegű hálózatok és az Univerzális Approximációs tétel

A perceptronnal bár sok probléma van, mégis talán az egyik legjelentősebb felfedezés a neurális hálózatok fejlődésében. Jelentősége hasonlítható a Newtoni modellhez: bár tudjuk, hogy az esetek döntő többségében nem használható, mégis az egyszerűségében rejlő zseniális konstrukciója miatt a mai napig használják. Ennek ellenére a következőekben rövid kitekintést teszünk a perceptronoktól a többrétegű hálózatok irányába.



3.1.1 Többrétegű hálózatok

Talán az egyik legjelentősebb konstrukció a 3 rétegű neurális hálózatok: az első réteg a bemeneti réteg, az utolsó réteg a kimeneti réteg, és e kettőt köti össze a belső, „rejtett" réteg. Az egyes neuronokat szinapszisok kötik össze. Az így kialakított konstrukció a mai napig kutatott téma, rengeteg nyitott kérdés létezik még.

3.1.2 Univerzális Approximációs Tétel

A perceptron konvergencia tételhez hasonló jelentőségű a többrétegű hálózatok témakörében az univerzális approximációs tétel (15), amely kimondja, hogy $\mathbb{R}^n$ tetszőleges kompakt részhalmazában univerzális approximálóként képes működni egy többrétegű perceptron (16) hálózat. További kapcsolódó eredmény, hogy ez esetben már nem az aktiváló függvény miatt lesz univerzális approximátor, hanem a hálózat felépítéséből adódóan (17). Ez igencsak meglepő lehet a Perceptron konvergencia tétel után, ugyanis ott elég erőteljesen a perceptron függvényének tulajdonságaira alapoztuk a bizonyítást.

Formálisan a tétel a következőt mondja:

Cybenko(-Hornik) tétel: legyen egy tetszőleges $\varphi: \mathbb{R} \to A \subset \mathbb{R}$ egy nem konstans, korlátos, monoton növő, folytonos függvény. Vegyük $K \subset \mathbb{R}^n$ kompakt halmazt. Adott egy $f: K \to \mathbb{R}$ folytonos függvény és $\varepsilon > 0$, akkor

$$\exists m \in \mathbb{N}, \exists a_i, b_i \in \mathbb{R}, \exists \overline{w}_i \in \mathbb{R}^n : F(\overline{x}) = \sum_{i=1}^{m} a_i \varphi(\overline{w}_i \cdot \overline{x} + b_i), F: K \to \mathbb{R}$$

úgy, hogy F approximálja f –t, azaz

$$\forall \overline{x} \in K: |F(x) - f(x)| < \varepsilon$$

Ennek bizonyítását itt mellőzzük.

4 Történet

Santiago Ramón y Cajal 1906-os Nobel díjas felfedezése óta (a neuron felfedezése) az ember igyekezett egyrészt megérteni agyunk működését, másrészt valahogyan modellezni annak működését. Az egyik nagyon egyszerű ilyen modell volt az első mesterséges neuron, a perceptron (18). A percepeptront egyébként nagyon sokszor a retina működéséhez szükséges neuronokhoz hasonlítják (19). Néhány egyszerűbb élfelismerés feladat például egészen jól megoldható ezzel.

Rosenblatt felfedezése nem egyedi: nagyon sokan tanulmányozták és írták le a neuronok működését, és adtak valamilyen közelítő modellt. Például, a perceptronhoz nagyon hasonló leírást adtak Pitts és McCulloch (20). Később, egyre bonyolultabb és egyre összetettebb leírásokat adtak, így egyre összetettebb feladatok ellátására lett alkalmas (pl.: a dolgozatban tárgyalt többrétegű neurális hálózat vagy univerzális approximációs tétel). Mégis a perceptronok egyszerűségük ellenére nem elhanyagolható jelentőségűek, sőt óriási hatással voltak a későbbi fejlődésre, és sok kifinomultabb mesterséges neuron tervezésénél a perceptronnal kapcsolatos tapasztalatokra alapoztak. A legfontosabb művek a témában egyrészt a feltaláló Rosenblatt (6), illetve Minksy és Papert (11) könyve. Ez utóbbival kapcsolatban sokan támadták a szerzőket (lásd 3. bekezdés), leginkább hogy nem ismertették azt, hogy bár perceptronnal nem, de többszintű hálózatokkal megvalósítható a XOR függvény. Később egyébként kiadtak egy bővített változatot, ami már tartalmazza az ezekkel kapcsolatos tapasztalatokat (12).

5 Irodalomjegyzék

1. *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.* **Rosenblatt, Frank J.** hely nélk. : Cornell Aeronautical Laboratory, 1958., Psychological Review, old.: 386–408.

2. *Perceptron-based learning algorithms.* **Gallant, Stephen I.** 2, hely nélk. : IEEE, 1990., IEEE Transactions on Neural Networks, 1. kötet, old.: 179-191.

3. *Perceptron Trees: A case study in hybrid conce.* **Utgoff, Paul E.** 1988., Proceedings Seventh Nationat Conference Artificial Intelligence, old.: 601-606.

4. *Learning in feedforward layered networks: the tiling algorithm.* **Mézard, Marc és Nadal, Jean-Pierre.** 12, 1989., Journal of Physics A: Mathematical and General , 22. kötet.

5. Linear separability. *Wikipedia.* [Online] 2010. http://en.wikipedia.org/wiki/Linearly_separable.

6. **Rosenblatt, J. Frank.** *Principles of Neurodynamics.* hely nélk. : Spartan Books, 1962.

7. **Csató, Lehel.** Perceptron Convergence Theorem Proof. *Lehel Csató homepage.* [Online] 2010. http://cs.ubbcluj.ro/~csatol/kozgaz_mestint/4_neuronhalo/PerceptConvProof.pdf.

8. Perceptron Convergence Theorem. [Online] http://annet.eeng.nuim.ie/intro/course/chpt2/convergence.shtml.

9. Cauchy–Schwarz_inequality. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Cauchy%E2%80%93Schwarz_inequality.

10. *On convergence proofs for perceptrons.* **Novikoff, Albert B.J.** 1963., In Proceedings of the Symposium on the Mathematical Theory of Automata, 12. kötet, old.: 615-622.

11. **Papert, Seymour A. és Minsky, Marvin L.** *Perceptrons: An Introduction to Computational Geometry.* 1969.

12. **Papert, Seymour A. és Minsky, Marvin L.** *Perceptrons: An Introduction to Computational Geometry - Expanded edition.* hely nélk. : The MIT Press; Exp Sub edition, 1987. ISBN-13: 978-0262631112.

13. Book: Perceptron. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Perceptrons_%28book%29.

14. *A learning rule for very simple universal approximators consisting of a single layer of perceptrons.* **Peter Auer, Harald Burgsteiner, Wolfgang Maass.** 5, 2007., Neural Networks, 21. kötet, old.: 786-795.

15. *Approximation by superpositions of a sigmoidal function.* **Cybenko, George.** 4, 1989., Mathematics of Control, Signals, and Systems (MCSS), 2. kötet, old.: 303-314.

16. Feedforward neural network. *Wikipedia.* [Online] 2010. http://en.wikipedia.org/wiki/Feedforward_neural_network.

17. *Approximation Capabilities of Multilayer Feedforward Networks.* **Hornik, Kurt.** 2, 1991., Neural Networks, 4. kötet, old.: 251-257 .

18. Perceptron. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Perceptron.

19. **Rojas, Raul.** Neural Networks - A Systematic Introduction. *Chapter 3 - Weighted Networks - The perceptron.* [Online] 1996. [Hivatkozva: 2010. december 2.] http://page.mi.fu-berlin.de/rojas/neural/chapter/K3.pdf.

20. *A logical Calculus of Ideas Immanent in nervous activity.* **McCulloch, Warren Sturgis és Pitts, Walter.** 1943., Bulletin of Mathematical Biophysics, old.: 115-133.

21. **Rojas, Raul.** *Neural Networks - A Systematic Introduction.* hely nélk. : Springer-Verlag, 1996. ISBN-13: 978-3540605058.