3. Instruire prin minimizarea erorilor pătratice

Algoritmul perceptronului şi metoda Gallant reprezintă proceduri iterative de instruire care pot fi aplicate doar în cazul unor ieşiri binare ale neuronilor.Ele sunt metode de *corectare a erorilor* deoarece modifică vectorul pondere numai în cazul unei clasificări incorecte. Aceşti algoritmi converg numai pentru clase liniar separabile. Pentru clase neseparabile, algoritmii devin oscilanți la obținerea soluției, nefiind posibilă determinarea unui vector de separare aproximativ. În acest capitol vom expune o altă metodă de instruire a rețelelor neuronale, care se va introduce mai întâi pentru o rețea cu un singur neuron.

3.1. METODĂ DE INSTRUIRE GLOBALĂ (ÎNTR-UN SINGUR PAS) A REȚELELOR NEURONALE

Considerăm o mulțime de instruire formată din vectorii normalizați în raport cu semnul $\mathbf{z}^1,\ \mathbf{z}^2,\ ...,\ \mathbf{z}^k,\ ...,\ \mathbf{z}^p$. Aşa cum s-a arătat prin instruire se urmăreşte determinarea unui vector pondere $\mathbf{w}$, astfel încât

$$\mathbf{w}^T\mathbf{z}^k > 0 \qquad (\forall)\ k = 1,2,...,p\,. \tag{3. 1}$$

În cele ce urmează se va prezenta o metodă de instruire globală, care determină toate ponderile simultan. Deoarece rezolvarea unui sistem de inecuații este mai complicată decât a unui sistem de ecuații, se înlocuieşte problema rezolvării sistemului (3. 1) cu rezolvarea unui sistem de ecuații "echivalent". În acest scop se adoptă un vector $\mathbf{b}$ cu p componente, numere pozitive convenabil alese,

$$\mathbf{b} = [b_1 \quad b_2 \quad ... \quad b_p]^T, \text{cu} \quad b_k > 0 \qquad (\forall)\ k = \overline{1, p}\,, \tag{3. 2}$$

şi se înlocuieşte inegalitatea

$$\mathbf{w}^T\mathbf{z}^k > 0\,, \tag{3. 3}$$

cu egalitatea

$$\mathbf{w}^T\mathbf{z}^k = b_k\,. \tag{3. 4}$$

Astfel, determinarea ponderilor rețelei, respectiv a vectorului $\mathbf{w}$ se reduce la problema rezolvării sistemului liniar de ecuații

$$\mathbf{w}^T\mathbf{z}^k = b_k \qquad (\forall)\ k = 1,2,...,p\,. \tag{3. 5}$$

Rezolvarea exactă a acestui sistem nu este posibilă în toate cazurile, pentru că, de multe ori, numărul p al ecuațiilor este mai mare decât numărul necunoscutelor egal cu $N+1$ (se consideră că valoarea pragului este absorbită în mulțimea ponderilor pe mărimea de intrare x_0 cu valoarea 1).

Fie $\mathbf{w}$ o soluție aproximativă a sistemului (3. 5). Eroarea pătratică asociată vectorului de instruire $\mathbf{z}^k$ este

$$e_k^2 = (\mathbf{w}^T\mathbf{z}^k - b_k)^2\,. \tag{3. 6}$$

Se introduce acum funcția criteriu $J : \Re^{N+1} \to \Re$ definită ca suma erorilor pătratice a tuturor vectorilor de instruire

$$J = \sum_{k=1}^{p} (\mathbf{w}^{\mathrm{T}} \mathbf{z}^k - b_k)^2 \tag{3. 7}$$

Pentru descrierea matriceală a funcției criteriu introducem matricea $\mathbf{Z}$ de dimensiune $p \times (N+1)$ care conține pe linia k vectorul transpus $\mathbf{z}^k$

$$\mathbf{Z} = \left[\mathbf{z}^{1\mathrm{T}} \ \mathbf{z}^{2\mathrm{T}} \ldots \mathbf{z}^{k\mathrm{T}} \ldots \mathbf{z}^{p\,\mathrm{T}} \right]^{\mathrm{T}}. \tag{3. 8}$$

Sistemul de ecuații (3. 5) a cărui soluție aproximativă o căutăm se poate scrie acum sub forma

$$\mathbf{Z}\mathbf{w} = \mathbf{b}. \tag{3. 9}$$

Pe de altă parte, având în vedere că $\mathbf{z}^{\mathrm{T}} \mathbf{w} = \mathbf{w}^{\mathrm{T}} \mathbf{z}$, vectorul $\mathbf{Z}\mathbf{w}$ se poate exprima prin

$$\mathbf{Z}\mathbf{w} = \left[\mathbf{w}^{\mathrm{T}} \mathbf{z}^1 \ \mathbf{w}^{\mathrm{T}} \mathbf{z}^2 \ldots \mathbf{w}^{\mathrm{T}} \mathbf{z}^p \right]^{\mathrm{T}}, \tag{3. 10}$$

deci

$$\sum_{k=1}^{p} (\mathbf{w}^{\mathrm{T}} \mathbf{z}^k - b_k)^2 = \left\| \mathbf{Z}\mathbf{w} - \mathbf{b} \right\|^2 = (\mathbf{Z}\mathbf{w} - \mathbf{b})^{\mathrm{T}} (\mathbf{Z}\mathbf{w} - \mathbf{b}). \tag{3. 11}$$

Rezultă că funcția criteriu J se poate scrie şi conform expresiei

$$J(\mathbf{w}) = \left\| \mathbf{Z}\mathbf{w} - \mathbf{b} \right\|^2 = (\mathbf{Z}\mathbf{w} - \mathbf{b})^{\mathrm{T}} (\mathbf{Z}\mathbf{w} - \mathbf{b}). \tag{3. 12}$$

Soluția aproximativă poate fi determinată prin minimizarea funcției criteriu, problemă care se rezolvă prin anularea derivatei în raport cu w:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{Z}^{\mathrm{T}} (\mathbf{Z}\mathbf{w} - \mathbf{b}) = 0, \tag{3. 13}$$

de unde rezultă:

$$\mathbf{Z}^{\mathrm{T}} \mathbf{Z}\mathbf{w} = \mathbf{Z}^{\mathrm{T}} \mathbf{b}. \tag{3. 14}$$

Dacă matricea pătratică $\mathbf{Z}^{\mathrm{T}} \mathbf{Z}$ este nesingulară, atunci sistemul precedent are soluția:

$$\mathbf{w}^* = (\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{b}. \tag{3. 15}$$

Matricea $\mathbf{Z}^+ = (\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}}$ se numeşte matrice inversă generalizată sau pseudo-inversă a matricei $\mathbf{Z}_{p \times (N+1)}$. Este evidentă proprietatea

$$\mathbf{Z}^+ \mathbf{Z} = \mathbf{I}, \tag{3. 16}$$

unde $\mathbf{I}$ este matricea unitate de dimensiune $(N+1) \times (N+1)$. Această proprietate arată că $\mathbf{Z}^+$ este inversa la stânga a matricei $\mathbf{Z}$. În particular dacă $\mathbf{Z}$ este o matrice pătratică nesingulară, atunci avem

$$\mathbf{Z}^+ = (\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} = \mathbf{Z}^{-1} (\mathbf{Z}^{\mathrm{T}})^{-1} \mathbf{Z}^{\mathrm{T}} = \mathbf{Z}^{-1}. \tag{3. 17}$$

Aşadar, dacă inversa unei matrice pătratice există, atunci inversa şi inversa generalizată coincid. În general, nu orice matrice dreptunghiulară admite o inversă generalizată. Condiția de existență a inversei generalizate este dată de următoarea teoremă:

Fie $\mathbf{A}$ o matrice de dimensiune $p \times q$ cu $p > q$. Matricea $\mathbf{A}$ admite o inversă generalizată dacă şi numai dacă

$$\text{rang } \mathbf{A} = q \tag{3. 18}$$

Din cele prezentate rezultă că vectorul pondere soluție a problemei de instruire se calculează cu relația matriceală

$$\mathbf{w}^* = \mathbf{Z}^+ \mathbf{b}, \tag{3. 19}$$

unde $\mathbf{Z}^+$ este inversa generalizată a matricei $\mathbf{Z}$.

Deoarece funcția criteriu J este convexă, minimul $\mathbf{w}^*$ al acestei funcții este un minim global. Vectorul $\mathbf{w}^*$ reprezintă o soluție de minimizare a erorii pătratice (MEP) pentru problema de instruire. Această soluție depinde în mod esențial de vectorul $\mathbf{b}$.

Observație: Chiar dacă clasele de instruire sunt liniar separabile, soluția MEP nu constituie neapărat un vector de separare în situația unei alegeri arbitrare a vectorului $\mathbf{b}$.

Conform teoremei de mai înainte, matricea $\mathbf{Z}$ de dimensiune $p \times (N+1)$ admite o inversă generalizată la stânga dacă şi numai dacă

$$\text{rang}\, \mathbf{Z} = (N+1). \tag{3. 20}$$

Acest lucru are loc dacă cel puțin $(N+1)$ dintre vectorii de instruire $\mathbf{z}^k$, cu $k = \overline{1,p}$, $p > (N+1)$ sunt liniar independenți. Geometric aceasta înseamnă că există $(N+1)$ vectori de instruire care nu aparțin aceluiaşi hiperplan.

3.2. Rețeaua ADALINE. Principiul algoritmului Widrow-Hoff determinist

Privind critic, metoda MEP are inconvenientul că necesită folosirea şi inversarea unor matrice de dimensiuni mari, cu pericolul ca matricea $\mathbf{Z}^T\mathbf{Z}$ să fie singulară (deci neinversabilă). Un alt dezavantaj de natură intrinsecă, mai puțin evident, provine din faptul că această metodă calculează ponderile într-un singur pas. Aşadar ea nu permite ameliorarea valorilor ponderilor în paşi succesivi, prin utilizarea de mai multe ori a fiecărui obiect din mulțimea de instruire.

Algoritmul Widrow-Hoff este un algoritm iterativ de instruire, care minimizează suma erorilor pătratice. Există mai multe variante de aplicare a acestui algoritm. Vom prezenta varianta de bază a algoritmului Widrow-Hoff aplicat unei rețele cu un singur neuron cu funcție de activare liniară numită de cei doi (Widrow şi Hoff) rețea de tip ADALINE (ADAptive LInear NEuron). Pentru fixarea notațiilor în fig. 3.1 se prezintă sub formă grafică structura acestei rețele. De aici se observă că funcția de activare este liniară, fiind descrisă de relația $f(s) = s$, unde s este activarea totală care se calculează cu expresia vectorială $s = \mathbf{w}^T\mathbf{x}$. Din acest motiv mărimea de ieşire are valorile cuprinse în mulțimea numerelor reale ($y \in \Re$), spre deosebire de perceptron la care ieşirea poate lua numai valorile ± 1 sau 0, 1. Diferă, de asemenea, metoda de instruire care se impune ca o consecință a faptului menționat mai înainte.

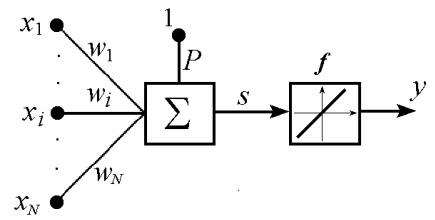


Fig. 3. 1 Structura unei rețele ADALINE.

Algoritmul Widrow-Hoff are la bază o metodă de optimizare de tip gradient care foloseşte ca funcție criteriu suma erorilor pătratice, considerând vectorii de intrare normalizați de semn. La metoda minimizării globale a erorii pătratice am obținut funcția criteriu:

$$J(\mathbf{w}) = (\mathbf{Zw} - \mathbf{b})^{\mathrm{T}} (\mathbf{Zw} - \mathbf{b}), \tag{3. 21}$$

în care:

$$\mathbf{Z} = \left[\mathbf{z}^{1\mathrm{T}} \quad \mathbf{z}^{2\mathrm{T}} \quad \ldots \quad \mathbf{z}^{k\mathrm{T}} \quad \ldots \quad \mathbf{z}^{p\mathrm{T}}\right]^{\mathrm{T}}, \tag{3. 22}$$

unde $\mathbf{z}^k$ este vectorul normalizat din mulțimea de instruire care cuprinde, neglijând semnul, componentele $x_1^k \quad x_2^k \quad \ldots \quad x_j^k \quad \ldots \quad x_{N+1}^k$ ale mărimilor de intrare.

Vectorul pondere se calculează iterativ cu relația cunoscută

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha_k \mathbf{p}^k, \tag{3. 23}$$

unde

$$\mathbf{p}^k = \operatorname{grad} J(\mathbf{w}) = \nabla J(\mathbf{w}) = \left.\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^k}. \tag{3. 24}$$

La metoda MEP globală am stabilit că:

$$\mathbf{p}^k = \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{Z}^{\mathrm{T}} (\mathbf{Zw}^k - \mathbf{b}), \tag{3. 25}$$

deci corecția la pasul k este de forma

$$\mathbf{w}^{k+1} = \mathbf{w}^k + c_k \mathbf{Z}^{\mathrm{T}} (\mathbf{b} - \mathbf{Zw}^k), \tag{3. 26}$$

unde $c_k = 2\alpha_k$.

În continuare vom modifica această regulă de instruire, considerând o ipoteză, general valabilă la instruirea rețelelor neuronale, potrivit căreia la pasul k se modifică numai un vector şi anume $\mathbf{z}^k$. În acest sens matricea $\mathbf{Z}$ se modifică adoptându-se conform expresiei

$$\mathbf{Z} = \left[0 \;\; 0 \ldots \mathbf{z}^{k\mathrm{T}} \ldots 0\right]^{\mathrm{T}}, \tag{3. 27}$$

în care toate liniile au componentele nule mai puțin linia k ce conține vectorul linie $\mathbf{z}^{k\mathrm{T}}$. Altfel spus matricea $\mathbf{Z}$ se înlocuieşte în regula de corecție cu $\mathbf{z}^{k\mathrm{T}}$ (vector linie). De asemenea, în relația de calcul recurent se înlocuieşte vectorul $\mathbf{b}$ cu componenta b_k, obținându-se astfel expresia:

$$\mathbf{w}^{k+1} = \mathbf{w}^k + c_k \mathbf{z}^k [b_k - \mathbf{z}^{k\mathrm{T}} \mathbf{w}^k] \,, \tag{3. 28}$$

care se poate rescrie sub forma

$$\mathbf{w}^{k+1} = \mathbf{w}^k + c_k [b_k - (\mathbf{w}^k)^{\mathrm{T}} \mathbf{z}^k] \mathbf{z}^k \,. \tag{3. 29}$$

Metoda de căutare recurentă astfel obținută este cunoscută drept *regula de corecție Widrow-Hoff* sau *regula delta*. Numărul b_k poate fi interpretat ca fiind valoarea dorită a ieşirii. În acest caz valoarea corecției rezultă proporțională cu diferența (delta) dintre ieşirea dorită şi ieşirea reală a rețelei:

$$\mathbf{w}^{k+1} - \mathbf{w}^k = c_k \delta_k \mathbf{z}^k \,, \tag{3. 30}$$

unde

$$\delta_k = d^k - (\mathbf{w}^k)^{\mathrm{T}} \mathbf{z}^k \,. \tag{3. 31}$$

Pentru convergența metodei este necesar ca c_k să descrească cu creşterea lui k. O alegere convenabilă pentru c_k este

$$c_k = c_1 / k \,, \tag{3. 32}$$

unde $c_1 > 0$.

Observație: În prezența unor condiții normale impuse coeficienților c_k procedeul iterativ Widrow-Hoff converge spre soluția de minim global a metodei MEP.

3. 3 ABORDAREA PROBABILISTĂ A PROBLEMEI DE MINIMIZARE A ERORII PĂTRATICE (MEP)

Formularea problemei. Funcția criteriu

Pentru a dezvolta o abordare statistică a instruirii vom considera vectorul mărimilor de intrare, la pasul k, o variabilă aleatoare vectorială $(N+1)$ - dimensională, notată $\mathbf{x}^k$. Familia de variabile aleatoare

$$\{\mathbf{x}^k \Big| \quad k = 1, \quad 2, \quad ..., \quad p \} \tag{3. 33}$$

constituie un proces aleator. În mod analog ieşirile, reală y^k şi dorită d^k ale rețelei vor fi considerate drept variabile aleatoare. Aşadar secvențele $\{d^k\}$ şi $\{y^k\}$ definesc procese aleatoare sau secvențe de variabile aleatoare. Eroarea pătratică instantanee e_k^2 reprezintă de asemenea o variabilă aleatoare a cărei valoare medie statistică este

$$\mathrm{E}\{e_k^2\} = \mathrm{E}\{(d^k - y^k)^2\} \,, \tag{3. 34}$$

unde $\mathrm{E}\{\cdot\}$ reprezintă operatorul de mediere statistică. În cele ce urmează se va presupune că vectorul ponderilor $\mathbf{w}^k$ este determinist. Din acest motiv vom avea proprietatea specifică operatorului de mediere statistică

$$\mathrm{E}\left\{\mathbf{w}^{k\mathrm{T}} \mathbf{x}^k\right\} = \mathbf{w}^{k\mathrm{T}} \mathrm{E}\left\{\mathbf{x}^k\right\}. \tag{3. 35}$$

Înlocuind $y^k = \mathbf{w}^{k\mathrm{T}}\mathbf{x}^k$ în expresia valorii medii a erorii pătratice şi având în vedere proprietățile cunoscute ale operatorului de mediere statistică putem scrie

$$\mathrm{E}\{e_k^2\} = \mathrm{E}\{(d^k - \mathbf{w}^{k\mathrm{T}}\mathbf{x}^k)^2\} = \mathrm{E}\{(d^k)^2\} + \mathrm{E}\{[\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k]^2\} - 2\mathrm{E}\{d^k\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k\} \qquad (3.\ 36)$$

În continuare, deoarece $\mathbf{w}^k$ este un vector determinist vom avea

$$\mathrm{E}\{[\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k]^2\} = \mathrm{E}\{(\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k)(\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k)\} = \mathbf{w}^{k\mathrm{T}}\mathrm{E}\{\mathbf{x}^k\mathbf{x}^{k\mathrm{T}}\}\mathbf{w}^k \qquad (3.\ 37)$$

şi

$$\mathrm{E}\{d^k\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k\} = \mathrm{E}\{d^k\mathbf{x}^{k\mathrm{T}}\mathbf{w}^k\} = \mathrm{E}\{d^k\mathbf{x}^{k\mathrm{T}}\}\mathbf{w}^k\,. \qquad (3.\ 38)$$

În ultimele două relații s-a utilizat proprietatea produsului scalar $\mathbf{w}^{k\mathrm{T}}\mathbf{x}^k = \mathbf{x}^{k\mathrm{T}}\mathbf{w}^k$.

În continuare să analizăm mărimile $\mathrm{E}\{\mathbf{x}^k\mathbf{x}^{k\mathrm{T}}\}$ şi $\mathrm{E}\{d^k\mathbf{x}^{k\mathrm{T}}\}$. Pentru simplificarea prezentării, renunțăm temporar la scrierea indicelui superior k. În aceste condiții $\mathrm{E}\{\mathbf{x}^k\mathbf{x}^{k\mathrm{T}}\}$ va fi o matrice notată:

$$\mathbf{R} = \mathrm{E}\{\mathbf{x}\cdot\mathbf{x}^{\mathrm{T}}\} = \begin{bmatrix} \mathrm{E}\{x_1^2\} & \mathrm{E}\{x_1x_2\} & \ldots & \mathrm{E}\{x_1x_{N+1}\} \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{E}\{x_{N+1}x_1\} & \ldots & \ldots & \mathrm{E}\{x_{N+1}^2\} \end{bmatrix} \qquad (3.\ 39)$$

de dimensiune $(N+1)\times(N+1)$, similară cu matricea de autocorelație a vectorului mărimilor de intrare, iar $\mathrm{E}\{d^k\mathbf{x}^{k\mathrm{T}}\}$ un vector

$$\mathbf{q} = \mathrm{E}\{d\,\mathbf{x}\} = [\mathrm{E}\{d_1x_1\}\,\mathrm{E}\{d_2x_2\}\ldots\mathrm{E}\{d_{N+1}x_{N+1}\}]^{\mathrm{T}}, \qquad (3.\ 40)$$

identic cu vectorul de intercorelație ieşire-intrare.

Cu notațiile astfel introduse, valoarea medie a erorii pătratice la momentul k se va putea scrie sub forma:

$$\mathrm{E}\{e_k^2\} = \mathrm{E}\{(d^k)^2\} + \mathbf{w}^{k\mathrm{T}}\mathbf{R}^k\mathbf{w}^k - 2\mathbf{q}^{k\mathrm{T}}\mathbf{w}^k, \qquad (3.\ 41)$$

sau dacă suprimăm în continuare, pentru simplitate, indicele k

$$J(\mathbf{w}) = \mathrm{E}\{e^2\} = \mathrm{E}\{d^2\} + \mathbf{w}^{\mathrm{T}}\mathbf{R}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{q}\,. \qquad (3.\ 42)$$

Soluția MEP probabilistă

Vectorul ponderilor sinaptice este o soluție a problemei de minim

$$\begin{cases} J(\mathbf{w}) \to \min \\ \mathbf{w} \in \Re^{N+1} \end{cases}. \qquad (3.\ 43)$$

Punctul de minim al funcției criteriu rezultă rezolvând ecuația

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0\,. \qquad (3.\ 44)$$

Având în vedere relația de derivare a unei forme pătratice şi faptul că matricea de autocorelație $\mathbf{R}$ este simetrică rezultă ecuația echivalentă:

$$2\mathbf{R}\mathbf{w} - 2\mathbf{q} = 0\,. \tag{3. 45}$$

Dacă matricea $\mathbf{R}$ este inversabilă se obține soluția analitică a problemei de instruire.

$$\mathbf{w}^* = \mathbf{R}^{-1}\mathbf{q}\,. \tag{3. 46}$$

În plus, pentru ca extremul obținut să fie un punct de minim trebuie să fie îndeplinită condiția

$$\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}\,\partial \mathbf{w}^{\mathrm{T}}} > 0\,, \tag{3. 47}$$

respectiv $\mathbf{R} > 0$.

Uneori în literatura de specialitate soluția MEP probabilistă se exprimă în funcție de pseudo-inversa matricei $\mathbf{R}$. Acest rezultat se obține simplu din condiția de extrem

$$\mathbf{R}\mathbf{w} = \mathbf{q}\,, \tag{3. 48}$$

înmulțind la stânga cu $\mathbf{R}^{\mathrm{T}}$, de unde rezultă

$$\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{w} = \mathbf{R}^{\mathrm{T}}\mathbf{q}\,. \tag{3. 49}$$

Dacă matricea $\mathbf{R}^{\mathrm{T}}\mathbf{R}$ este nesingulară atunci există matricea

$$\mathbf{R}^{+} = (\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}} \tag{3. 50}$$

care este inversa generalizată a matricei $\mathbf{R}$. Soluția MEP probabilistă se scrie acum sub forma

$$\mathbf{w}^* = \mathbf{R}^{+}\mathbf{q}\,. \tag{3. 51}$$

Se constată că această soluție a problemei MEP probabilistă este similară formal cu soluția MEP deterministă. Remarcăm, totuşi, că matricea $\mathbf{R}$ care apare în această expresie este pătratică, pe când matricea $\mathbf{Z}$ care apare în soluția MEP deterministă este dreptunghiulară de dimensiune $p \times (N+1)$.

Procedeul iterativ de determinare pe bază statistică a vectorului pondere optim

Metoda MEP probabilistă prezentată mai înainte are dezavantajul major că necesită cunoaşterea matricei de autocorelație statistică $\mathbf{R}$ a vectorului de intrare şi a vectorului de intercorelație statistică ieşire-intrare $\mathbf{q}$. Calculul explicit al lui $\mathbf{R}$ şi $\mathbf{q}$ impune cunoaşterea statistică a semnalelor de intrare şi ieşire, adică a funcțiilor de densitate de probabilitate ale acestora la fiecare moment k. O altă dificultate este legată de calculul matricei $\mathbf{R}^{-1}$ şi de resursele de memorie pe care le implică acest calcul. O soluție alternativă o constituie folosirea unei metode iterative de optimizare de tip gradient. În acest caz vectorul pondere se deplasează pe suprafața funcției criteriu J, pe direcția celei mai mari pante.

Procedura iterativă se va scrie şi în acest caz conform relației:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha_k \mathbf{p}^k , \tag{3. 52}$$

unde direcția de coborâre este gradientul funcției criteriu în punctul $\mathbf{w}^k$:

$$\mathbf{p}^k = \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2(\mathbf{R}^k \mathbf{w}^k - \mathbf{q}^k) . \tag{3. 53}$$

Rezultă că:

$$\mathbf{w}^{k+1} = \mathbf{w}^k + c_k (\mathbf{R}^k \mathbf{w}^k - \mathbf{q}^k) , \tag{3. 54}$$

unde $c_k = 2\alpha_k$.

Determinarea direcției de coborâre ridică aproape aceleaşi probleme tehnice ca şi determinarea, într-un singur pas, a soluției MEP probabiliste. Ar trebui să cunoaştem dinainte $\mathbf{R}$ şi $\mathbf{q}$, ceea ce înseamnă să ştim a priori suprafața corespunzătoare funcției criteriu. Pentru a depăşi aceste dificultăți vom folosi o versiune simplificată a funcției criteriu. În acest scop se înlocuieşte valoarea medie statistică a erorii pătratice cu valoarea instantanee a acesteia, deci:

$$J(\mathbf{w}) = e_k^2 . \tag{3. 55}$$

Direcția de coborâre este în acest caz

$$\mathbf{p}^k = \left. \frac{\partial (e_k^2)}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^k} , \tag{3. 56}$$

unde

$$e_k^2 = [d^k - \mathbf{w}^{k\mathrm{T}} \mathbf{x}^k]^2 . \tag{3. 57}$$

Efectuând derivata funcției pătratice, rezultă

$$\mathbf{p}^k = -2e_k \mathbf{x}^k , \tag{3. 58}$$

relație care se obține avându-se în vedere proprietatea de derivare a produsului scalar

$$\frac{\partial}{\partial \mathbf{w}} (\mathbf{w}^{\mathrm{T}} \mathbf{x}) = \frac{\partial}{\partial \mathbf{w}} (\mathbf{x}^{\mathrm{T}} \mathbf{w}) = \mathbf{x} . \tag{3. 59}$$

Am obținut astfel următoarea aproximare:

$$\left. \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^k} = \frac{\partial}{\partial \mathbf{w}} \mathrm{E}(e_k^2) \approx -2e_k \mathbf{x}^k . \tag{3. 60}$$

Regula de corecție se poate scrie acum:

$$\mathbf{w}^{k+!} = \mathbf{w}^k + c_k e_k \mathbf{x}^k , \tag{3. 61}$$

sau exprimând în mod explicit eroarea:

$$\mathbf{w}^{k+!} = \mathbf{w}^k + c_k [d^k - \mathbf{w}^{k\mathrm{T}} \mathbf{x}^k] \mathbf{x}^k . \tag{3. 62}$$

S-a regăsit, aşadar, procedeul iterativ de instruire Widrow-Hoff determinist sau regula delta. În legătură cu această formă a algoritmului sunt necesare câteva comentarii:

i) vectorii $\mathbf{x}^k$ nu au suferit vreo normalizare de semn;

ii) valorile ieşirii d^k pot fi pozitive sau negative spre deosebire de b_k de la algoritmul determinist care era în mod obligatoriu pozitiv;

iii) dacă ne situăm în cazul paradigmei clasificatorului, din regula de mai sus se poate deduce algoritmul Widrow-Hoff determinist:

$\mathbf{w}^{k+1} = \mathbf{w}^k + c_k[b_k - \mathbf{w}^{k\mathrm{T}}\mathbf{z}^k]\mathbf{z}^k$, unde $\mathbf{z}^k$ sunt vectori normalizați şi $b_k > 0$;

iv) sarcinile pe care le poate îndeplini o rețea ADALINE (neuron instruit cu algoritmul Widrow-Hoff) sunt mai generale decât stabilirea apartenenței unui obiect la o clasă (din două posibile).

3. 4 ALGORITMUL WIDROW-HOFF

În metoda Widrow-Hoff, la fiecare pas al instruirii se prezintă rețelei doar un singur vector de intrare. Mai întâi se consideră cazul unei rate de învățare constante, deci $c_k = c, \quad k = 1,2,...$

Descrierea algoritmului

P_0 Se inițializează ponderile rețelei alegând un vector $\mathbf{w}^1$ arbitrar. Se fixează rata de învățare $c > 0$. Se stabileşte numărul vectorilor de instruire, p. Se pune $k := 1$ şi $e^2 := 0$ (eroarea totală).

P_1 Se prezintă rețelei un vector de intrare $\mathbf{x}^k$ şi ieşirea corectă (dorită) d^k.

P_2 Se calculează valoarea instantanee a erorii pătratice e_k^2

$e_k^2 = [d^k - \mathbf{w}^{k\mathrm{T}}\mathbf{x}^k]^2$ şi eroarea totală, $e^2 := e^2 + e_k^2$.

P_3 Se adaptează vectorul pondere folosind regula delta probabilistă $\mathbf{w}^{k+!} = \mathbf{w}^k + c_k[d^k - \mathbf{w}^{k\mathrm{T}}\mathbf{x}^k]\mathbf{x}^k$.

P_4 Se pune $k := k + 1$. Pentru $k \le p$ se merge la pasul **P_1**, altfel la pasul **P_5**.

P_5 Se verifică dacă $e^2 \le \varepsilon$ (de exemplu $\varepsilon = 10^{-5}$). Dacă da STOP, alfel se merge la pasul **P_1**.

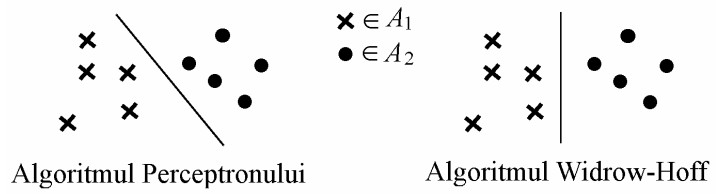
Alegerea ratei de instruire c are o importanță deosebită în cadrul algoritmului iterativ Widrow-Hoff. Această mărime controlează convergența spre vectorul pondere obținut prin metoda MEP. Dacă c are o valoare prea mică atunci algoritmul converge foarte lent. Dacă c este prea mare, atunci algoritmul poate să nu sesizeze punctele de minim. În acest caz este posibil ca procesul de instruire să nu conveargă niciodată. Rata de învățare se poate alege o constantă, sau variabilă, caz în care se calculează cu relația $c_k = 1/k$ sau $c_k = c/k$, $c < 0$, $k = 1,2,...$ unde k este indexul (pasul curent) al algoritmului de instruire. Alegerea constantei c se face pe baza experienței sau riguros folosindu-se unele noțiuni din teoria clasificării care nu vor fi expuse aici.

Algoritmul Widrow-Hoff versus algoritmul perceptronului

În această secțiune se va face un mic studiu comparativ al celor mai importanți algoritmi de instruire prezentați până acum: algoritmul perceptronului şi procedura Widrow-Hoff.

i) Algoritmul Widrow-Hoff a fost dezvoltat pentru a găsi o metodă de instruire pentru date neseparabile, ca o alternativă la algoritmul perceptronului;

ii) O altă motivație a fost de a găsi, pentru clase separabile, o soluție mai bună decât cele furnizate de algoritmul perceptronului. Instruirea cu algoritmul perceptronului se opreşte de îndată ce se obține o soluție care clasifică în mod corect fiecare formă de instruire. În cele mai multe cazuri, soluția găsită este abia acceptabilă (vezi figura 3. 2a). În figura 3. 2b se prezintă o soluție mai bună ce se obține cu metoda Widrow –Hoff;



a) b)

Fig. 3. 2 Comparația soluțiilor de clasificare realizate cu metoda perceptronului respectiv cu procedura W-H.

iii) Algoritmul Widrow-Hoff asigură față de algoritmul perceptronului o viteză de instruire mai mare. Acest lucru se realizează prin ajustarea ratei de instruire la fiecare iterație a procedurii.

iv) La metoda de instruire cu algoritmul Widrow-Hoff, ieşirile dorite d_k pot fi numere reale, nefiind limitate la valorile ±1 ca la perceptron.

v) Metoda MEP globală şi algoritmul Widrow-Hoff furnizează o soluție atât pentru clase separabile cât şi pentru clase neseparabile. Prezența însă a unui singur punct izolat poate face ca hiperplanul de separare să fie complet perturbat, astfel încât majoritatea datelor de instruire să fie eronat clasificate. Acest lucru reprezintă un dezavantaj serios deoarece punctul izolat poate fi un exemplu de instruire fals datorat zgomotului.

Nu există la această oră o comparație realizată în mod satisfăcător între algoritmul Widrow-Hoff şi algoritmul buzunarului (Gallant).