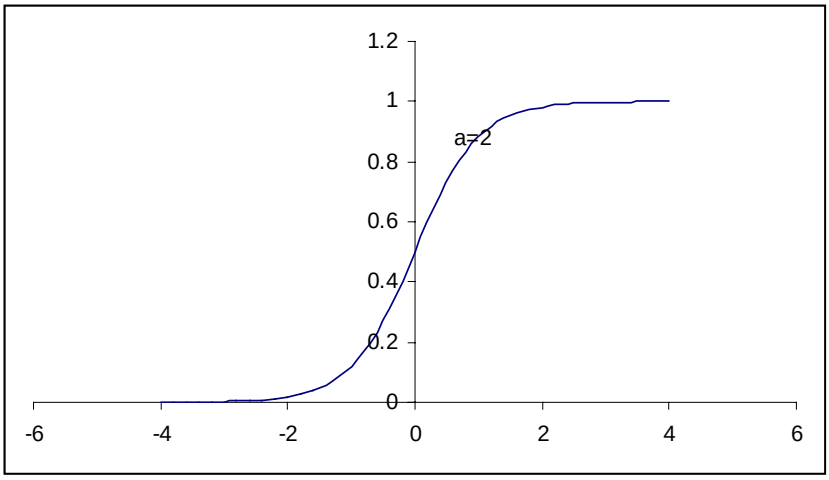
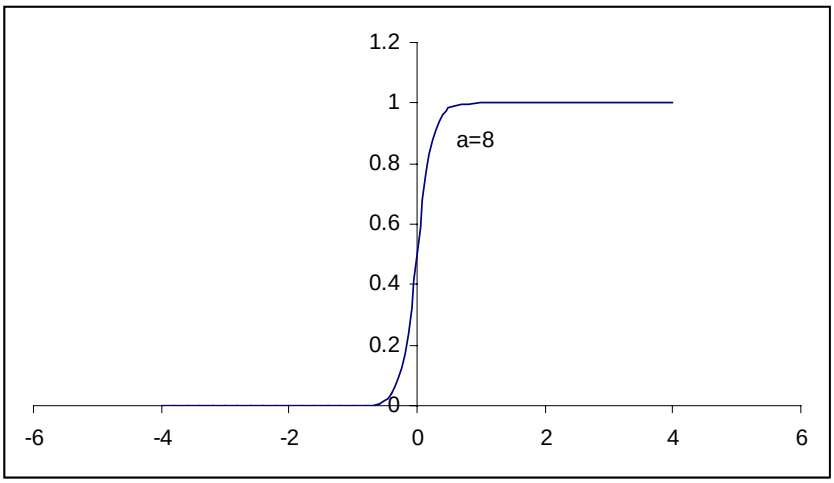
4.6. Μη γραμμικοί ταξινομητές – ΝΔ Back error propagation

Μία διαφορετική τεχνική σχεδιασμού ενός πολυεπίπεδου perceptron για την ταξινόμηση μη γραμμικά διαχωριζομένων κλάσεων βασίζεται στην αντικατάσταση της συνάρτησης d(x) από μία συνεχή και διαφορίσιμη συνάρτηση f(x) που την προσεγγίζει. Υπάρχουν διάφορες αναλυτικές μορφές για την f(x) με ποιο συνηθισμένη την λεγόμενη *λογιστική* που δίδεται από τον τύπο

(4.6.1) $$f(x) = \frac{1}{1+e^{-\alpha x}},\ \alpha > 0$$

Η παράμετρος *α* λέγεται *παράμετρος κλίσης* (*slope*) ή λοξότητος, Σχ 4.6.-1





Σχήμα 4.6-1 *Η λογιστική συνάρτηση για α=2 και α=8.*

Η χρήση της *f(x)* διευκολύνει την εφαρμογή μιας τυποποιημένης μεθόδου ελαχιστοποίησης της κατάλληλης συνάρτησης κόστους. Το όλο ΝΔ όπως θα παρουσιασθεί και θα αναλυθεί παρακάτω βασίζεται στα παραπάνω και ονομάζεται λόγω της μεθόδου εκπαίδευσής του, ΝΔ *Back error propagation* ή για συντομία *Back propagation.*

Στο νευρωνικό δίκτυο υπάρχουν επίπεδα δηλαδή ομάδες νευρώνων, πλήρως διασυνδεδεμένα μεταξύ τους, Σχ.4.6-1 Η έξοδος κάθε νευρώνα ενός επιπέδου εισέρχεται σε κάθε νευρώνα του επόμενου επιπέδου, με αντίστοιχη σύναψη (βάρος), Σχ.4.6-2 Οι τιμές του πίνακα εισόδου αποτελούν το επίπεδο εισόδου. Εκτός από το επίπεδο εισόδου υπάρχουν *Z* το πλήθος επόμενα διαδοχικά επίπεδα, το τελευταίο των οποίων ονομάζεται επίπεδο εξόδου. Τα επίπεδα που ενδεχομένως υπάρχουν μεταξύ των επιπέδων εισόδου και εξόδου, λέγονται *κρυφά επίπεδα*. Στο Σχ. 4.6-2 φαίνεται ένα νευρωνικό όπου το πρότυπο εισόδου εχει δύο χαρακτηριστικά (x_1,x_2) που εισέρχονται σε ένα κρυφό επίπεδο τριών νευρώνων οι έξοδοι των οποίων είναι είσοδοι ενός επόμενου επιπέδου με δύο νευρώνες και δύο τελικές εξόδους. Για την αρίθμηση των επιπέδων θα χρησιμοποιούμε τον δείκτη ζ με τιμές $\zeta=0,\ldots,Z$, με $\zeta=0$ για το επίπεδο εισόδου και $\zeta=Z$ για το επίπεδο εξόδου Σχ. 4.6-1

Ονομάζουμε N_ζ το πλήθος των νευρώνων ενός επιπέδου ζ. Ως εκ τούτου το πλήθος των στοιχείων του πίνακα εισόδου $\boldsymbol{x}$ είναι N_0 και του πίνακα εξόδου $\boldsymbol{y}$ είναι N_Z. Αν ν είναι δείκτης για αρίθμηση των νευρώνων του επιπέδου ζ, τότε οι συνάψεις του συνθέτουν έναν πίνακα γραμμής $\mathbf{w}_\nu^\zeta$. Αν μ δείκτης για την αρίθμηση των νευρώνων του προηγουμένου επιπέδου ζ-1 που αποτελείται από $N_{\zeta-1}$ νευρώνες, τότε

(4.6.2) $$\mathbf{w}_\nu^\zeta = [w_{\nu 0}^\zeta,\ldots,w_{\nu\mu}^\zeta,\ldots,w_{\nu N_{\zeta-1}}^\zeta]$$

Η έξοδος του αθροιστή του νευρώνα θα είναι σ_ν^ζ

(4.6.3) $$\sigma_\nu^\zeta = \sum_{\mu=1}^{N_{\zeta-1}} w_{\nu\mu}^\zeta \cdot y_\mu^{\zeta-1} + w_{0\nu}^\zeta = \sum_{\mu=0}^{N_{\zeta-1}} w_{\nu\mu}^\zeta \cdot y_\mu^{\zeta-1} \text{ με } y_0^{\zeta-1} = 1$$

Η τελική έξοδος του νευρώνα, y_ν^ζ προκύπτει από την σχέση

(4.6.4) $$y_\nu^\zeta = f(\sigma_\nu^\zeta)$$

όπου $f(.)$ κατάλληλη συνάρτηση με την οποία θα ασχοληθούμε παρακάτω. Στο Σχ. 4.6-3 φαίνεται αναλυτικά η δομή ενός νευρώνα όπως περιγράφηκε παραπάνω.

Οι τιμές των εξόδων κάθε επιπέδου είναι οι είσοδοι του επόμενου επιπέδου όπως περιγράφτηκε παραπάνω ξεκινώντας από το επίπεδο εισόδου και προχωρώντας προοδευτικά μέχρι το επίπεδο εξόδου (*forward propagation*).

Εκπαίδευση *back-error propagation*

Έστω I το πλήθος ζευγών από πίνακες εισόδου και των αντίστοιχων επιθυμητών πινάκων εξόδου με γνωστές τιμές, το σύνολο εκπαίδευσης S ορίζεται ως

$S=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)/(\boldsymbol{x}_i,\boldsymbol{y}_i)$ ζεύγος με x_i πίνακα στήλης εισόδου και
$\boldsymbol{y}_i$ τον αντίστοιχο επιθυμητό πίνακα στήλης εξόδου, $i=1,\ldots,I\}$.

Αν ο δείκτης ν αριθμεί τους νευρώνες του επιπέδου εξόδου Z, $\nu=1,\ldots,N_Z$ και ο πίνακας εξόδου του ΝΔ είναι $\mathbf{y}^Z=[y_1^z,\ldots,y_\nu^z,\ldots,y_{N_Z}^z]^T$ για συγκεκριμένο ζεύγος $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ ορίζουμε μία συνάρτηση $\Delta(i)$ σύμφωνα με την σχέση:

(4.6.5) $$\Delta(i)=\frac{1}{2}\sum_{\nu=1}^{N_Z}(y_\nu^Z-y_{\nu i})^2$$

Η σχέση $\Delta(i)$ είναι ένα άθροισμα τετραγωνικών σφαλμάτων μεταξύ της παραγόμενης εξόδου $\boldsymbol{y}^Z$ του ΝΔ όταν η είσοδος είναι $\boldsymbol{x}_i$ και της επιθυμητής εξόδου $\boldsymbol{y}_i$. Μπορούμε να ορίσουμε τώρα την συνάρτηση κόστους $K(.)$ που θα έχει ανεξάρτητες μεταβλητές όλα τα $\mathbf{w}_\nu^\zeta$ για το συγκεκριμένο σύνολο εκπαίδευσης S σύμφωνα με την σχέση

(4.6.6) $$K(w_\nu^\zeta,S)=\sum_{i=1}^{I}\Delta(i)$$

Αντιμετωπίζοντας την εύρεση ελάχιστης τιμής της K ως πρόβλημα βελτιστοποίησης (*optimization*) οι τιμή των $\mathbf{w}_\nu^\zeta$ μπορούν να εκτιμηθούν επαναληπτικά σύμφωνα με σχέση

(4.6.7) $$\mathbf{w}_v^\zeta(t+1) = \mathbf{w}_v^\zeta(t) - \rho \frac{\partial \mathrm{K}}{\mathbf{w}_v^\zeta}\Big|_{\mathbf{w}_v^\zeta(t)} = \mathbf{w}_v^\zeta(t) - \rho \sum_{i=1}^{I} \frac{\partial \Delta(\mathrm{i})}{\mathbf{w}_v^\zeta}\Big|_{\mathbf{w}_v^\zeta(t)}$$

Σύμφωνα με τον κανόνα παραγώγισης της αλυσίδας

(4.6.8) $$\frac{\partial \Delta(i)}{\partial \mathbf{w}_v^\zeta} = \frac{\partial \Delta(i)}{\partial \sigma_v^\zeta} \cdot \frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_v^\zeta}$$

Ο όρος $\frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_v^\zeta}$ για όλα τα $\mathbf{w}_v^\zeta$ (για κάθε v και ζ>0) είναι

(4.6.9) $$\frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_v^\zeta} = \begin{bmatrix} \frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_{v0}^\zeta} \\ \vdots \\ \frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_{v\mu}^\zeta} \\ \vdots \\ \frac{\partial \sigma_v^\zeta}{\partial \mathbf{w}_{v\mathrm{N}_{\zeta-1}}^\zeta} \end{bmatrix} = \begin{bmatrix} 1 \\ y_0^{\zeta\text{-}1} \\ \vdots \\ y_\mu^{\zeta\text{-}1} \\ \vdots \\ y_{\mathrm{N}_{\zeta-1}}^{\zeta\text{-}1} \end{bmatrix} = \mathbf{y}^{\zeta\text{-}1}$$

Απομένει τώρα ο υπολογισμός του πρώτου όρου του γινομένου, τον οποίο ονομάζουμε

(4.6.10) $$\delta_v^\zeta = \frac{\partial \Delta(i)}{\partial \sigma_\mathrm{v}^\zeta}$$

Θα υπολογίσουμε πρώτα το δ_v^ζ για έναν νευρώνα v του επιπέδου εξόδου (ζ=Z, N_ζ=N_Z) και δείκτη αρίθμησης των νευρώνων μ=1,…,N_Z.

(4.6.11) $$\delta_v^\zeta = \frac{\partial \Delta(i)}{\partial \sigma_\mathrm{v}^\zeta} = \frac{\partial}{\partial \sigma_\mathrm{v}^\zeta}\left(\frac{1}{2}\sum_{\mu=1}^{\mathrm{N_Z}}(y_\mu^\zeta - y_{\mu\mathrm{i}})^2\right) \text{ με } y_\mu^\zeta = f(\sigma_\mu^\zeta) \Rightarrow$$
$$\delta_v^\zeta = \frac{1}{2}\sum_{\mu=1}^{\mathrm{N_Z}} \frac{\partial}{\partial \sigma_v^\zeta}(f(\sigma_\mu^\zeta) - y_{\mu\mathrm{i}})^2 = \frac{2}{2}(f(\sigma_\mathrm{v}^\zeta) - y_{\mathrm{vi}}) f'(\sigma_\mathrm{v}^\zeta) \Rightarrow$$

(4.6.12) $$\delta_v^\zeta = (y_\mathrm{v}^\zeta - y_{\mathrm{vi}}) f'(\sigma_\mathrm{v}^\zeta)$$

Για τα κρυφά επίπεδα (0<ζ<Z) ο υπολογισμός του δ_ν^ζ είναι περιπλοκότερος, επειδή δεν είναι δεδομένη η τιμή της εξόδου y_{vi}, κάθε νευρώνα. Ο υπολογισμός σε κάθε κρυφό επίπεδο θα βασισθεί στις τιμές διόρθωσης του επομένου του, αρχίζοντας από το επίπεδο που προηγείται του επιπέδου εξόδου και οδεύοντας διαδοχικά προς τα πίσω (*back-error propagation*). Συγκεκριμένα αν ζ ένα κρυφό επίπεδο, το επόμενό του θα είναι το $\zeta+1$. Έστω ακόμη μ ένας μετρητής αρίθμησης των νευρώνων του ζ επιπέδου και κ ένας μετρητής αρίθμησης των νευρώνων του $\zeta+1$ επιπέδου. Η συνάρτηση $\Delta(i)$ εξαρτάται από τα $\sigma_1^{\zeta+1},...,\sigma_\kappa^{\zeta+1},...,\sigma_{N_{\zeta+1}}^{\zeta+1}$ και κάθε $\sigma_\kappa^{\zeta+1}$ εξαρτάται από το σ_ν^ζ του ν-οστού νευρώνα του ζ επιπέδου. Σύμφωνα με τον κανόνα της αλυσιδωτής παραγώγισης

(4.6.13)

$$\delta_\nu^\zeta = \frac{\partial\Delta(i)}{\partial\sigma_\nu^\zeta} = \sum_{\kappa=1}^{N_{\zeta+1}} \frac{\partial\Delta(i)}{\partial\sigma_\kappa^{\zeta+1}} \frac{\partial\sigma_\kappa^{\zeta+1}}{\partial\sigma_\nu^\zeta} \overset{(1)}{=} \sum_{\kappa=1}^{N_{\zeta+1}} \delta_\kappa^{\zeta+1} \frac{\partial\sigma_\kappa^{\zeta+1}}{\partial\sigma_\nu^\zeta} = \sum_{\kappa=1}^{N_{\zeta+1}} \delta_\kappa^{\zeta+1} \frac{\partial}{\partial\sigma_\nu^\zeta}\left(\sum_{\mu=1}^{N_\zeta} w_{\kappa\mu}^{\zeta+1}\cdot y_\mu^\zeta\right) =$$

$$= \sum_{\kappa=1}^{N_{\zeta+1}} \delta_\kappa^{\zeta+1}\cdot w_{\kappa\nu}^{\zeta+1}\cdot f'\left(\sigma_\nu^\zeta\right) \Rightarrow$$

(4.6.14) $$\delta_\nu^\zeta = \left[\sum_{\kappa=1}^{N_{\zeta+1}} \delta_\kappa^{\zeta+1}\cdot w_{\kappa\nu}^{\zeta+1}\right]\cdot f'\left(\sigma_\nu^\zeta\right)$$

Τελικά από τις σχέσεις (4.6.6 - 9) συνάγεται ότι

(4.6.15) $$\mathbf{w}_\nu^\zeta(t+1) = \mathbf{w}_\nu^\zeta(t) - \rho \sum_{\forall(x_i,y_i)} \delta_\nu^\zeta \cdot y^{\zeta\text{-}1}$$

όπου δ_ν^ζ δίνεται από την σχέση (4.6.12)

(4.6.16) $$\delta_\nu^\zeta = (y_\nu^\zeta - y_{\nu i}) f'(\sigma_\nu^\zeta)$$

όταν ο νευρώνας ν βρίσκεται στο επίπεδο εξόδου και από την (4.6.14)

(4.6.17) $$\delta_\nu^\zeta = \left[\sum_{\kappa=1}^{N_{\zeta+1}} \delta_\kappa^{\zeta+1}\cdot w_{\kappa\nu}^{\zeta+1}\right]\cdot f'\left(\sigma_\nu^\zeta\right)$$

όταν ο νευρώνας *ν* βρίσκεται σε κρυφό επίπεδο ξεκινώντας από το τελευταίο και υποχωρώντας προοδευτικά μέχρι το πρώτο επίπεδο (*back-error propagation*).

Το Back error propagation ΝΔ είναι ίσως το πλέον χρησιμοποιούμενο ΝΔ και έχει εφαρμοσθεί σε πληθώρα εφαρμογών από διαφορετικές επιστημονικές περιοχές. Στα κρυφά επίπεδά του προσδιορίζονται ουσιαστικά αυτόματα τα αποτελεσματικά χαρακτηριστικά για την ταξινόμηση. Το πλήθος των κλάσεων μπορεί να είναι το ίδιο με αυτό των ανυσμάτων εισόδου του συνόλου εκπαίδευσης και έτσι το ΝΔ να προσεγγίζει ένα μετασχηματισμό των ανυσμάτων εισόδου στα ανύσματα εξόδου. Το σημαντικότερο μειονέκτημα του back propagation είναι χρόνος ολοκλήρωσης της εκπαίδευσης του ή χρόνος σύγκλισης, όπως αλλιώς λέγεται. Είναι δυνατόν να χρειασθούν εκατοντάδες χιλιάδες επαναλήψεις έως ότου συγκλίνει ακόμη και για σχετικά απλές εφαρμογές. Σε κάποιες εφαρμογές χρειάσθηκαν μερικές μέρες για την σύγκλιση του συστήματος. Ο εγκλωβισμός της διαδικασίας σύγκλισης σε τοπικά ελάχιστα της συνάρτησης κόστους είναι ένα επιπρόσθετο πρόβλημα που θα αναλύσουμε σε ακόλουθη παράγραφο.