Μηχανική μάθηση

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2022

Περίληψη

Ορισμός

1. Είναι τεχνητά νευρωνικά δίκτυα τα οποία διαθέτουν ένα ή περισσότερα ενδιάμεσα επίπεδα επεξεργασίας.
2. Απαραίτητο στοιχείο τους είναι οι συναρτήσεις ενεργοποίησης
3. Μπορούν να χρησιμοποιηθούν για μάθηση συναρτήσεων αλλά και για εύρεση κατηγοριών.
4. Επιλύουν προβλήματα που ένα απλό Perceptron δεν μπορεί να λύσει.
5. Αποτελούν την βάση για τα βαθιά νευρωνικά δίκτυα.

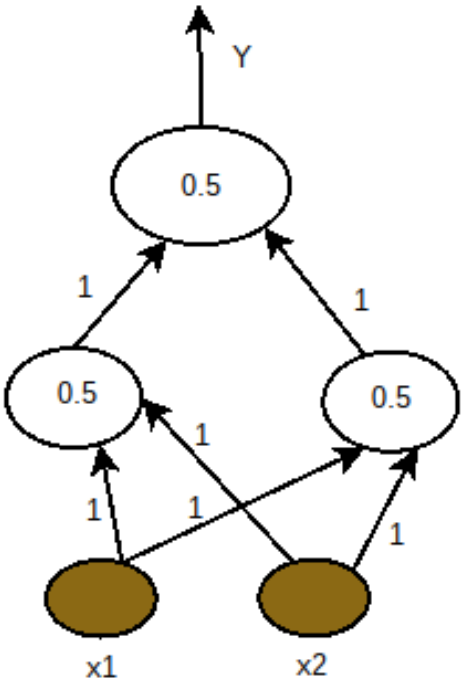
Το πρόβλημα XOR

1. Τα δίκτυα PERCEPTRON και ADALINE δεν είναι σε θέση να μάθουν πολύπλοκες συναρτήσεις όπως το κλασικό πρόβλημα XOR.
2. Το πρόβλημα αυτό δίνεται στον επόμενο πίνακα.

A	B	Υ
0	0	0
0	1	1
1	0	1
1	1	0

Επίλυση του XOR με MLP

Μια ενδεικτική υλοποίηση με χρήση MLP



Θεωρία παγκόσμιας προσέγγισης

Για τα δίκτυα MLP έχει αναπτυχθεί και αποδειχθεί η θεωρία της παγκόσμιας προσέγγισης:

1. Ένα δίκτυο δύο στρωμάτων (είσοδος - επεξεργασία - έξοδος) μπορεί να προσεγγίσει μια οποιαδήποτε συνάρτηση.
2. Δεν χρειάζεται πάνω από ένα στρώμα επεξεργασίας
3. Οι νευρώνες του κρυφού στρώματος έχουν σαν συνάρτηση ενεργοποίησης την σιγμοειδή.
4. Ο νευρώνας εξόδου έχει σαν έξοδο την γραμμική συνάρτηση ενεργοποίησης.

Το MLP σαν συνάρτηση

Ένα νευρωνικό δίκτυο MLP μπορεί να αναπαρασταθεί με πολλούς τρόπους αλλά ο πιο απλός είναι η ακόλουθη εξίσωση:

$$N\left(\vec{x},\vec{w}\right)=\sum_{i=1}^{H} w_{(d+2)i-(d+1)}\,\sigma\left(\sum_{j=1}^{d} x_j w_{(d+2)i-(d+1)+j}+w_{(d+2)i}\right) \tag{1}$$

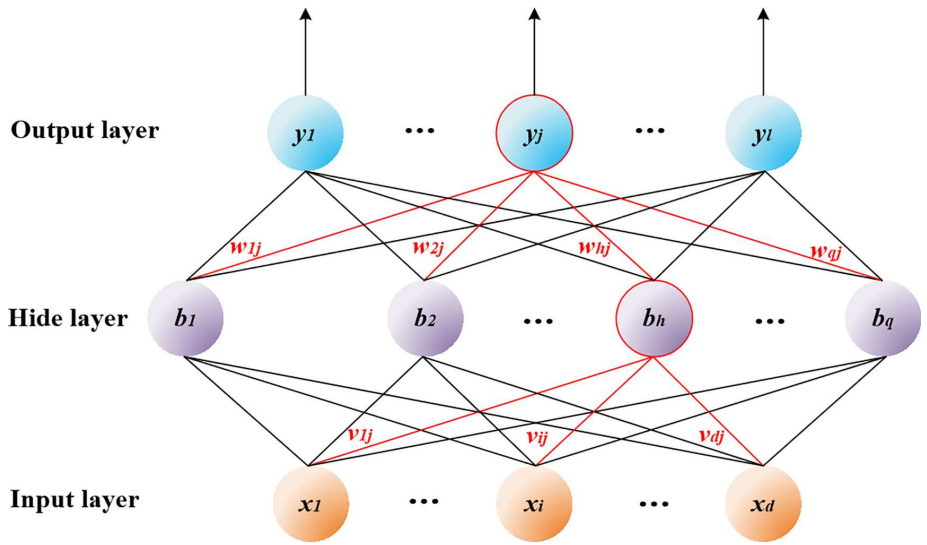
όπου

1. H είναι ο συνολικός αριθμός μονάδων επεξεργασίας του νευρωνικού δικτύου
2. d είναι η διάσταση του προβλήματος εισόδου.
3. $\vec{w}$ είναι τα βάρη του ΤΝΔ.
4. Η πόλωση κάθε μονάδας επεξεργασίας: $w_{(d+2)i}$.
5. Αν d=3 και έχουμε 2 κόμβους επεξεργασίας τότε συνολικά θα υπάρχουν $(d+2)H=10$ συνολικά παράμετροι στο ΤΝΔ

Μάθηση κατηγοριών με πολλές εξόδους

1. Χρησιμοποιούνται τόσες έξοδοι όσες και το πλήθος των κατηγοριών του προβλήματος (πχ στο wine 3)
2. Το αποτέλεσμα του ΤΝΔ θεωρείται εκείνη η έξοδος με την μεγαλύτερη τιμή εξόδου
3. Απαιτείται μεγαλύτερος αριθμός βαρών και πιο αργοί χρόνοι εκπαίδευσης

ΤΝΔ πολλών εξόδων



Κατώφλια στην μάθηση κατηγοριών

1. Για τις περιπτώσεις που θέλουμε το παραπάνω δίκτυο να κάνει πρόβλεψη κατηγορίας μπορούν να χρησιμοποιηθούν κατώφλια στις τιμές.
2. Για παράδειμα έστω ο ακόλουθος ενδεικτικός πίνακας:

y(x)	N(x)	CLASS
1	2.4	1
0	0.4	0
0	1.3	1
1	0.8	1

3. Στην στήλη y(x) είναι η πραγματική έξοδος και στην στήλη N(x) η έξοδος του ΤΝΔ.
4. Με την χρήση κατωφλιών (κοντινότερη κατηγορία) παράγεται η στήλη CLASS όπου βλέπουμε πως το ΤΝΔ "μαντεύει" σωστά τις 3 από τις 4 πραγματικές εξόδους.

Μέσο τετραγωνικό σφάλμα σε μάθηση συναρτήσεων

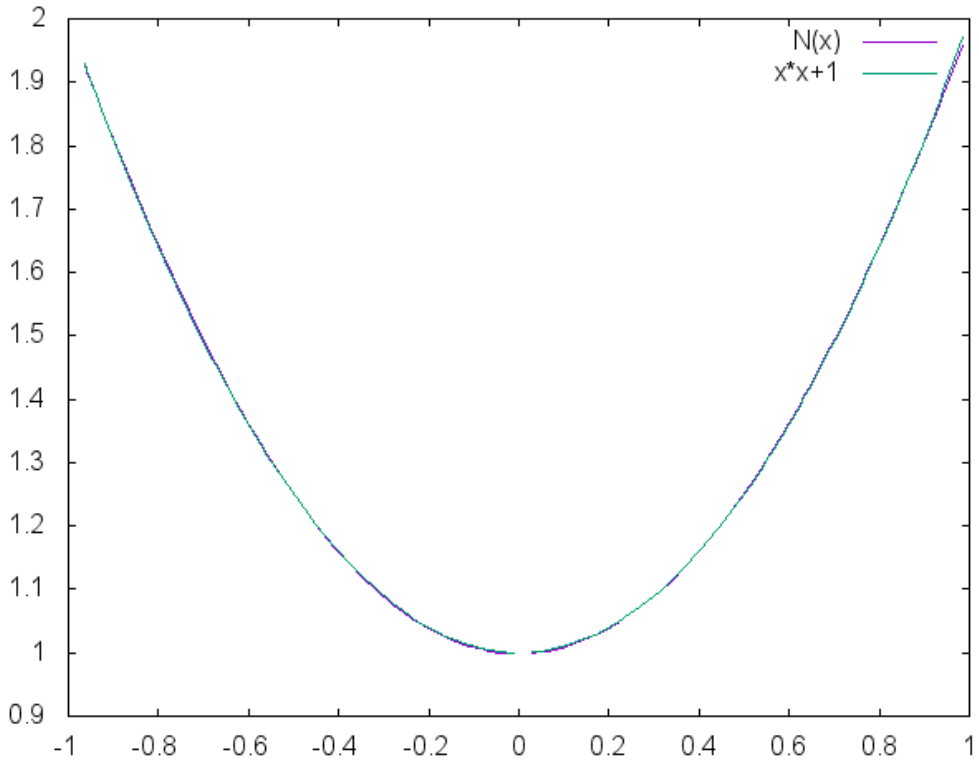
1. Έστω $x = (x_1, x_2, ..., x_M)$ τα πρότυπα εισόδου
2. Έστω $y = (y_1, y_2, ..., y_M)$ οι πραγματικές έξοδοι
3. Το σφάλμα ορίζεται

$$E = \frac{1}{M} \sum_{i=1}^{M} (N(x_i) - y_i)^2$$

4. Μπορεί να χρησιμοποιηθεί άμεσα σε μεθόδους βελτιστοποίησης

Παράδειγμα μάθησης συνάρτησης

Στο σχήμα παρατίθεται η μάθησης της συνάρτησης $x^2 + 1$ την οποία το δίκτυο Adaline δεν κατόρθωσε να μάθει.



Σφάλμα κατηγοριοποίησης

1. Έστω $x = (x_1, x_2, ..., x_M)$ τα πρότυπα εισόδου
2. Έστω $y = (y_1, y_2, ..., y_M)$ οι πραγματικές έξοδοι
3. Το σφάλμα ορίζεται

$$\mathrm{E}=\frac{1}{M}\sum_{i=1}^{M}\left(C\left(N(x_i)\right) \neq y_i\right)$$

4. Η συνάρτηση C(X) βρίσκει και επιστρέφει την κοντινότερη κατηγορία στην οποία είναι η τιμή X.

Σφάλμα κατηγοριοποίησης σε imbalanced data

1. Είναι δεδομένα στα οποία τα πρότυπα που ανήκουν σε κάποιες κατηγορίες είναι πολύ λιγότερα από τον μέσο όρο.
2. Σε αυτήν την περίπτωση το ΤΝΔ τείνει να μάθει μόνο τα δεδομένα που ανήκουν στην κατηγορία με τα περισσότερα πρότυπα.
3. Μια καλύτερη μέθοδος υπολογισμού
   1. Έστω οι κατηγορίες $C = (c_1, c_2, .., c_K)$
   2. Υπολογίζουμε το σφάλμα κατηγοριοποίησης μεμονωμένο για κάθε κατηγορία $E(c_i)$
   3. Το σφάλμα ορίζεται ως

$$E = \frac{1}{K} \sum_{i=1}^{K} E(c_i)$$

Ο αλγόριθμος Back Propagation (gradient descent)

Το μέσο τετραγωνικό σφάλμα ορίζεται ως:

$$E\left(N\left(\vec{x},\vec{w}\right)\right)=\sum\left(N\left(\vec{x}_i,\vec{w}\right)-y_i\right)^2 \tag{2}$$

Στην μέθοδο αυτή ισχύουν τα ακόλουθα:

1. Κινούμαστε αντίθετα από ότι λέει η παράγωγος της εξίσωσης 2. Αν η παράγωγος είναι θετική τότε μειώνεται το αντίστοιχο βάρος αλλιώς αυξάνεται.
2. Ο συντελεστής μάθησης είναι θετικός αριθμός και μικρότερος του 1.
3. Τα βάρη ενημερώνονται από τον κανόνα

$$w=w-n\frac{\partial E}{\partial w} \tag{3}$$

Gradient Descent

1. Ο αλγόριθμος Back Propagation ουσιαστικά είναι ο αλγόριθμος Gradient Descent
2. Έχει πολύ αργή σύγκλιση
3. Απαιτείται παραγώγιση (σε δίκτυα πολλών επιπέδων είναι δύσκολο).
4. Εξαρτάται άμεσα από την επιλογή του n
5. Υπάρχει online τρόπος μάθησης (ένα - ένα τα πρότυπα) και offline.
6. Τα βάρη ενημερώνονται διαρκώς είτε μέχρι να υπάρξει σύγκλιση είτε μέχρι να φτάσουμε σε μέγιστο αριθμό επαναλήψεων.

Momentum Back propagation

1. Κρατάμε και παλιές μεταβολές στα βάρη
2. Η ενημέρωση γίνεται από τον τύπο

$$w^{(k+1)} = w^{(k)} - n\frac{\partial E}{\partial w^{(k)}} + m\delta w^{(k-1)}$$

3. Σε περιοχές που έχουμε αργή σύγκλιση η ορμή επιταχύνει τον αλγόριθμο.
4. Σε περιοχές με μεγάλη ταλάντωση επιβραδύνει τον αλγόριθμο.
5. Και αυτή η μέθοδος έχει προβλήματα όπως η προηγούμενη
6. Συνήθως είναι ταχύτερες και πιο αποτελεσματικές μέθοδοι με παραγώγους δεύτερου βαθμού, όπως η BFGS

RPROP

1. Στην μέθοδο αυτή το βήμα είναι μεταβλητό στην ενημέρωση των βαρών.
2. Επίσης για την ενημέρωση των βαρών χρησιμοποίει τα πρόσημα των παραγώγων και όχι τις ίδιες τις παραγώγους.
3. Η ενημέρωση γίνεται

$$w^{(k+1)} = w^{(k)} - n^{(k)} \times \text{sign}\left(\frac{\partial E}{\partial w^{(k)}}\right)$$

Το βήμα στον RPROP

1. Το βήμα δίνεται από τον τύπο

$$n^{(k)} = \begin{cases} \min\left(n^{(k-1)} \times a, n_{\max}\right) & , \frac{\partial E^{(k)}}{\partial w^{(k)}} \times \frac{\partial E^{(k-1)}}{\partial w^{(k-1)}} > 0 \\ \max\left(n^{(k-1)} \times b, n_{\min}\right) & , \frac{\partial E^{(k)}}{\partial w^{(k)}} \times \frac{\partial E^{(k-1)}}{\partial w^{(k-1)}} < 0 \\ n^{(k-1)} & , \text{otherwise} \end{cases}$$

2. Όπου $a > 1 > b$
3. Το βήμα περιορίζεται μεταξύ των τιμών $n_{\min}$ και $n_{\max}$ για να μην γίνει πολύ μεγάλο ή πολύ μικρό.
4. Σε περίπτωση που η παράγωγος γίνει 0, τότε βρισκόμαστε σε τοπικό ελάχιστο της συνάρτησης σφάλματος και επομένως το βήμα δεν αλλάζει.

Αρχικοποίηση σε χαμηλές τιμές

1. Είναι ο πιο απλός τρόπος αρχικοποίησης.
2. Χρησιμοποιείται στην πράξη περισσότερο από όλους.
3. Τα βάρη αρχικοιποιούνται ομοιόμορφα σε χαμηλό διάστημα τιμών πχ. [-0.1,0.1]

Αρχικοποίηση Xavier

1. Τα βάρη αρχικοποιούνται ομοιόμορφα στο διάστημα $\left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]$
2. d είναι η διάσταση των προτύπων εισόδου.
3. Υπάρχει και η κανονικοποιημένη εκδοχή: $\left[-\frac{\sqrt{6}}{\sqrt{d+m}}, \frac{\sqrt{6}}{\sqrt{d+m}}\right]$, όπου m είναι ο αριθμός των κόμβων στο τρέχον επίπεδο.

Γενίκευση

1. Σκοπός της μάθησης είναι η εύρεση κρυμμένων συσχετίσεων
2. Γενίκευση είναι η μάθηση δεδομένων σε άγνωστα πρότυπα, τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση
3. Χρήση πολλών νευρώνων: το δίκτυο μαθαίνει πολύπλοκες συναρτήσεις αλλά δεν μπορεί να μάθει σε άγνωστα δεδομένα το ίδιο καλά
4. Χρήση λίγων νευρώνων: το δίκτυο μαθαίνει απλές μόνο συναρτήσεις.

Validation set

1. Διαίρεση του συνόλου εκπαίδευσης σε εκπαίδευσης και επικύρωσης.
2. Εκπαιδεύεται το δίκτυο στο μικρότερο πλέον σύνολο εκπαίδευσης
3. Η εκπαίδευση διακόπτεται όταν στο σύνολο επικύρωσης ξεκινά να ανεβαίνει στο σφάλμα.
4. Χρειάζεται προσεκτική επιλογή του ποσοστού για το σύνολο επικύρωσης
5. Απαραίτητα προϋπόθεση να έχουμε αρκετά δεδομένα στα χέρια μας.

N-Fold

1. Θέτουμε την παράμετρο N
2. Διαιρούμε το σύνολο των δεδομένων σε N σύνολο $S_1, S_s, ..., S_N$
3. Για i=1..N
   1. Δημιουργία συνόλου εκπαίδευσης T με όλα τα S_j για j διαφορετικό του i
   2. Εκπαίδευση στο σύνολο T και αποτίμηση του σφάλματος στο S_i με σφάλμα E_i
4. Ο μέσος όρος των σφαλμάτων E_i είναι και το τελικό σφάλμα.

Στις πιο πολλές περιπτώσεις N=10 αλλά και η τιμή N=2 χρησιμοποιείται αρκετά συχνά.

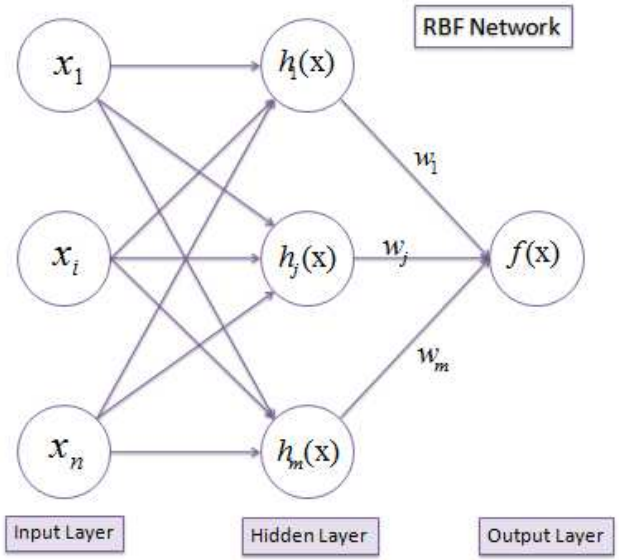
Leave One Out

- Η μέθοδος Folding χρησιμοποιείται πάρα πολύ αλλά σε κάποιες περιπτώσεις τα δεδομένα δεν είναι αρκετά για να γίνει η παραπάνω διαδικασία.
- Σε αυτήν την περίπτωση χρησιμοποιείται η τεχνική leave one out, στην οποία το σύνολο ελέγχου αποτελείται μόνον από ένα πρότυπο.

Ορισμός

1. Τα δίκτυα ακτινικής βάσης μπορούν να χρησιμοποιηθούν για μάθηση συναρτήσεων και για ταξινόμηση δεδομένων.
2. Διαθέτουν ένα στρώμα εισόδου, ένα στρώμα επεξεργασίας και μία ή περισσότερες εξόδους.

Σχήμα δικτύου RBF



$$f(\mathrm{x}) = \sum_{j=1}^{m} w_j h_j(\mathrm{x})$$

$$h(x) = \exp\left(-\frac{(x-c)^2}{r^2}\right)$$

Ακτινικές συναρτήσεις

1. Μια συνάρτηση f(x) ονομάζεται ακτινικού τύπου (radial function) αν υπάρχει κάποιο διάνυσμα c (το οποίο θα αποκαλούμε κέντρο) και η τιμή της συνάρτησης εξαρτάται μόνον από την απόσταση του x από αυτό το κέντρο.
2. Πρέπει να είναι πάντα στην μορφή

$$h(x) = f\left(\|x - c\|\right)$$

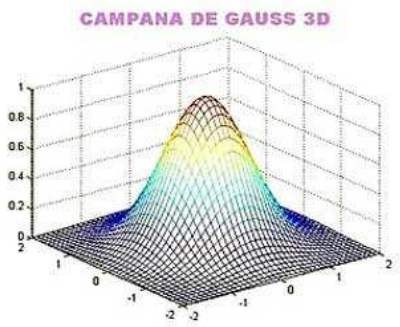
όπου x είναι το διάνυσμα εισόδου και c κάποιο κέντρο.

3. Η συνάρτηση gauss δίνεται από τον τύπο

$$f(x) = e^{-\frac{\|x-c\|^2}{\sigma^2}} \tag{4}$$

4. Αυτή η συνάρτηση χρησιμοποιείται περισσότερο στις διάφορες υλοποιήσεις RBF νευρωνικών δικτύων.

Γραφική αναπαράσταση συνάρτησης Gauss



Σύγκριση με MLP

1. Το RBF έχει μόνο δύο στρώματα επεξεργασίας
2. Δεν υπάρχει αλγόριθμος εκπαίδευσης για παραπάνω από ένα επίπεδα
3. RBF,MLP: και τα δύο πραγματοποιούν μάθηση με επίβλεψη.
4. Και στα δύο ισχύει το θεώρημα της παγκόσμιας προσέγγισης

Φάσεις εκπαίδευσης

1. Εκπαίδευση δύο σταδίων
    1. Στην πρώτη φάση εκπαιδεύονται οι παράμετροι c,σ για τις ακτινικές συναρτήσεις
    2. Στην δεύτερη φάση εκπαιδεύονται τα βάρη των νευρώνων εξόδου.
2. Εκπαίδευση ενός σταδίου
    1. Όλες οι παράμετροι του δικτύου εκπαιδεύονται ταυτόχρονα πχ με την χρήση ενός γενετικού αλγορίθμου.

Βασικός αλγόριθμος εκπαίδευσης

Για κάθε νευρώνα επεξεργασίας i=1..N υπολογισμός της τιμής

$$a_i = f\left(\|x - c_i\|, \sigma_i\right) \tag{5}$$

Για κάθε νευρώνα εξόδου i=1..K υπολογισμός της τιμής

$$o_i = \sum_{j=1}^{K} w_{ij} a_j \tag{6}$$

Σε πολλές περιπτώσεις K=1, αφού αρκεί και για μάθηση συναρτήσεων αλλά και για κατηγοριοποίηση δεδομένων.

Εκπαίδευση επιπέδου επεξεργασίας

1. Για να γίνει η εκπαίδευση ενός RBF θα πρέπει να υπάρχει διαφορετικός αλγόριθμος για το επίπεδο επεξεργασίας και διαφορετικός για το επίπεδο εξόδου.
2. Στο επίπεδο επεξεργασίας πρέπει να προσδιοριστούν τα κέντρο c_i καθώς και το εύρος των συναρτήσεων σ_i.
3. Δύο αλγόριθμοι που αντιμετωπίζουν αυτό το πρόβλημα είναι ο αλγόριθμος "κάθε πρότυπο και κέντρο" και ο κλασικός αλγόριθμος ομαδοποίησης K-means.

Κάθε πρότυπο και κέντρο

1. Σε αυτήν την περίπτωση θεωρούμε πως κάθε πρότυπο είναι και κέντρο
2. $c_i = x_i, \ i = 1..m$
3. Αυτή η μέθοδος δεν απαιτεί κάποια εκπαίδευση
4. Είναι αρκετά αργή όταν τα πρότυπα εισόδου είναι πολλά σε πλήθος.
5. Για τον υπολογισμό της παραμέτρου σ_i έχει προταθεί μόνον για την ακτινική συνάρτηση Gauss η τιμή

$$\sigma = \frac{D}{\sqrt{m}} \tag{7}$$

όπου D είναι η απόσταση μεταξύ των πιο απομακρυσμένων κέντρων.

Kmeans

1. **Αρχικοποίηση** των K κέντρων $c_i,\ i = 1..K$
2. **Επανέλαβε**
   1. $S_i = \{\},\ i = 1..K$
   2. Εύρεση της ομάδας που ανήκει το κάθε στοιχείο $x_i,\ \iota = 1..N$ ως ακολούθως: α) εύρεση $j^* = \min_{i=1}^{K} \{D(x_i, c_j)\}$ β) $S_{j^*} = S_{j^*} \cup x_i$
   3. Έστω M_j το πλήθος των μελών της ομάδας. Ανανέωση του κέντρου της ομάδας

$$c_j = \frac{1}{M_j} \sum_{i=1}^{M_j} x_i \tag{8}$$

   όπου x_i τα μέλη της ομάδας.
3. **Αν** τα κέντρα δεν έχουν αλλάξει **τότε τερματισμός**, **αλλιώς** μετάβαση στο βήμα 2.
4. Υπολογισμός των παραμέτρων σ_i ως την διακύμανση των προτύπων κάθε ομάδας S_i

Εκπαίδευση εξωτερικού στρώματος

1. Αν θεωρήσουμε πως το εξωτερικό στρώμα έχει μόνον μια έξοδο, τότε θα υπάρχει η εξίσωση

$$o = \sum_{i=1}^{N} w_i a_i \tag{9}$$

2. Για την μάθηση των βαρών w μπορεί να χρησιμοποιηθεί και ο κανόνας ADALINE και να γίνει μάθηση των βαρών με χρήση και ενός ρυθμού μάθησης n, $n<1$. Για παράδειγμα gradient descent ή BFGS.
3. Ωστόσο σε πολλές περιπτώσεις η εκπαίδευση του w γίνεται με την επίλυση ενός γραμμικού συστήματος.

Εκπαίδευση εξωτερικού επιπέδου με γραμμικό σύστημα

1. Θεωρούμε τον πίνακα

$$A = \begin{bmatrix} f(x_1, c_1) & f(x_1, c_2) & \dots & f(x_1, c_N) \\ f(x_2, c_1) & f(x_2, c_2) & \dots & f(x_2, c_N) \\ \dots & \dots & \dots & \dots \\ f(x_m, c_1) & f(x_m, c_2) & \dots & f(x_m, c_N) \end{bmatrix}$$

2. όπου $f(x, c)$ είναι η ακτινική συνάρτηση βάσης για είσοδο x και κέντρο c.
3. Τα βάρη ενημερώνονται με την πράξη

$$w = \left(A^T A\right)^{-1} A^T y \tag{10}$$

4. Με y την επιθυμητή έξοδο του δικτύου RBF
5. Ο πίνακας $\left(A^T A\right)^{-1} A^T$ ονομάζεται και ψευδοαντίστροφος του A.

Σύνοψη

- Παρουσίαση δικτύων MLP
- Μέθοδοι εκπαίδευσης
- Ειδικά θέματα MLP
- Παρουσίαση δικτύων RBF

Βιβλιογραφία I


Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386–408.


Freund, Y. and Schapire, R. E. 1998. Large margin classification using the perceptron algorithm. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT' 98). ACM Press.