Μηχανική Μάθηση

Μιχάλης Τίτσιας

Διάλεξη 4ή

Μη γραμμικά μοντέλα και νευρωνικά δίκτυα

Περιεχόμενα

- Επανάληψη: Γραμμικά μοντέλα
- Μη γραμμικότητα
- Μη γραμμικότητα με πολυωνυμικές συναρτήσεις
- Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης
- Νευρωνικά δίκτυα

Επανάληψη: Γραμμικά μοντέλα

Γραμμική παλινδρόμηση

- Έστω ότι έχουμε τα ακόλουθα δεδομένα εκπαίδευσης

$$\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^{N}, \quad t_n \in \mathbb{R}$$

Μοντέλο: Υποθέτουμε μια γραμμική σχέση

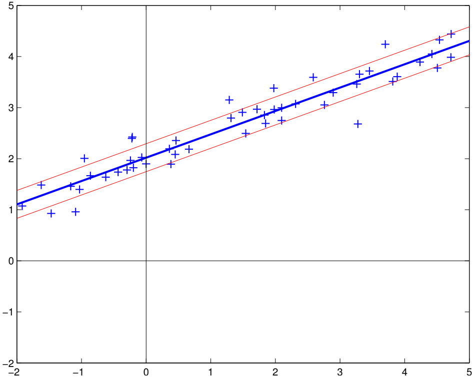
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D$$

Πιθανοτική κατανομή του t δοθέντος του δεδομένου εισόδου $\mathbf{x}$ είναι η Gaussian

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(t - y(\mathbf{x}, \mathbf{w})\right)^2\right\}$$

Επανάληψη: Γραμμικά μοντέλα

Γραμμική παλινδρόμηση



Η κεντρική μπλε γραμμή δείχνει την $w_0 + w_1 x$, ενώ οι δύο κόκκινες βρίσκονται σε απόσταση μιας τυπικής απόκλισης

Οι τιμές (w_0, w_1, β) είναι όλες βέλτιστες. Η λογαριθμική πιθανοφάνεια είναι η μέγιστη $\mathcal{L} = -6.1644$

Επανάληψη: Γραμμικά μοντέλα

Γραμμική λογιστική παλινδρόμηση

- Έστω ότι έχουμε τα ακόλουθα δεδομένα εκπαίδευσης

$$\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^{N}, \quad t_n \in \{0, 1\}$$

Μοντέλο: Υποθέτουμε μια γραμμική σχέση

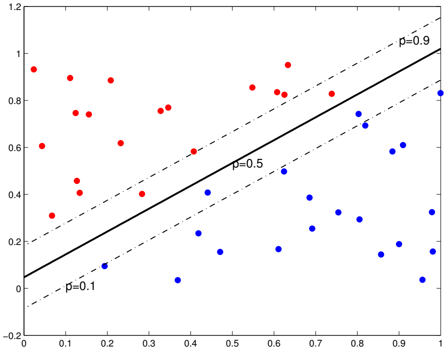
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = \mathbf{w}^T \mathbf{x}$$

Πιθανοτική κατανομή του t δοθέντος του δεδομένου εισόδου $\mathbf{x}$ είναι η

$$p(t|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})^t \left(1 - \sigma(\mathbf{w}^T \mathbf{x})\right)^{1-t}$$

Επανάληψη: Γραμμικά μοντέλα

Γραμμική λογιστική παλινδρόμηση



Επανάληψη $k = 300$ του αλγορίθμου της ανοδικής κλίσης

Επανάληψη: Γραμμικά μοντέλα

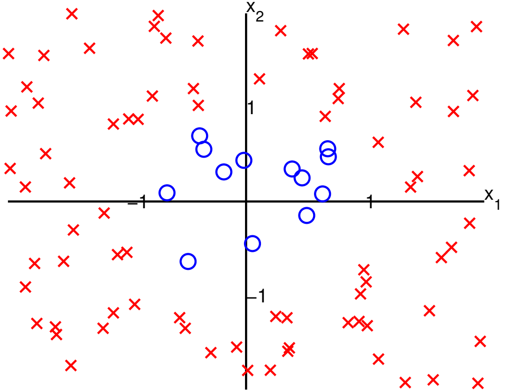
Ο κοινός παρανομαστής αυτών των μεθόδων είναι η γραμμική υπόθεση

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = \mathbf{w}^T \mathbf{x}$$

που οδηγεί σε γραμμικά σύνορα απόφασης κτλ

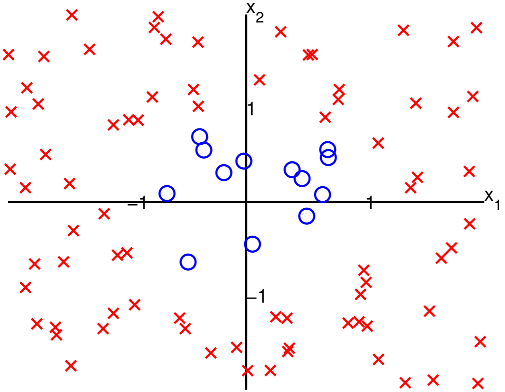
Πώς θα μπορούσαμε να κατασκευάσουμε μη γραμμικά μοντέλα;

Μη γραμμικά σύνορα απόφασης



Πώς θα μπορούσαμε να διαχωρίσουμε τις κατηγορίες με λογιστική παλινδρόμηση;

Μη γραμμικά σύνορα απόφασης

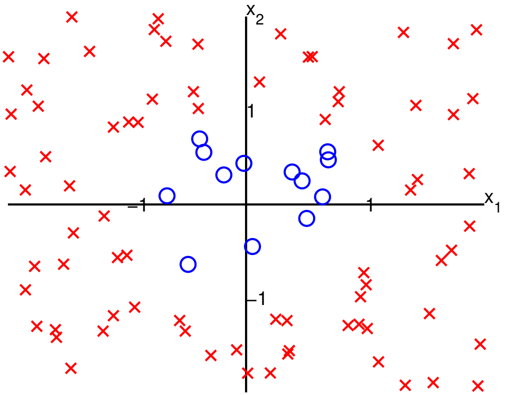


Η γραμμική υπόθεση $y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2$ δεν είναι κατάλληλη

Ας δοκιμάσουμε την

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$

Μη γραμμικά σύνορα απόφασης

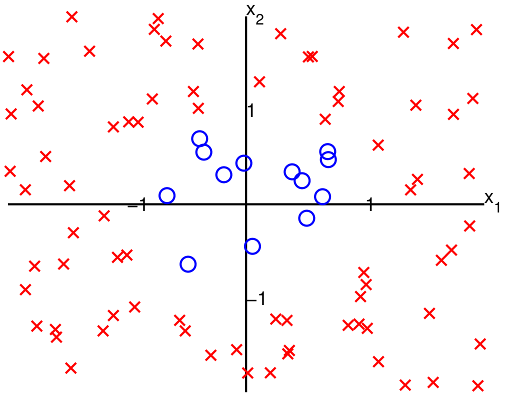


$$p(t=1|\mathbf{x}) = \frac{1}{1+e^{-w_0-w_1x_1-w_2x_2-w_3x_1^2-w_4x_2^2}}$$

Έστω

$w_0 = -1, w_1 = 0, w_2 = 0, w_3 = 1, w_4 = 1$

Μη γραμμικά σύνορα απόφασης

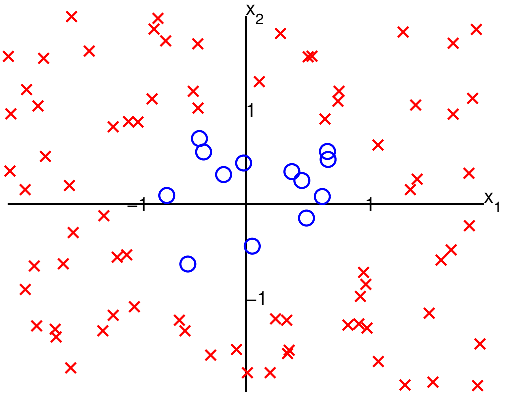


$\Rightarrow p(t=1|\mathbf{x}) = \frac{1}{1+e^{1-x_1^2-x_2^2}}$

Αποφάσισε $t = 1$ αν $-1 + x_1^2 + x_2^2 \geq 0$

Αποφάσισε $t = 0$ αν $-1 + x_1^2 + x_2^2 < 0$

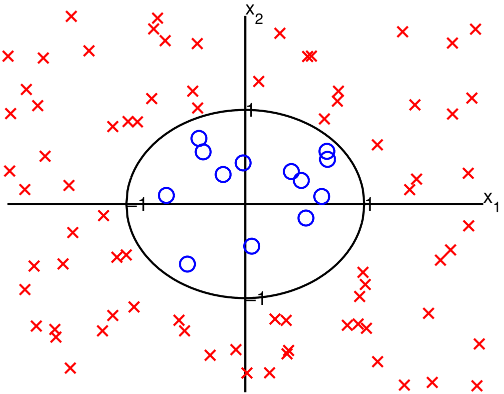
Μη γραμμικά σύνορα απόφασης



Αποφάσισε $t = 1$ αν $-1 + x_1^2 + x_2^2 \geq 0 \Rightarrow x_1^2 + x_2^2 \geq 1$

Αποφάσισε $t = 0$ αν $-1 + x_1^2 + x_2^2 < 0 \Rightarrow x_1^2 + x_2^2 < 1$

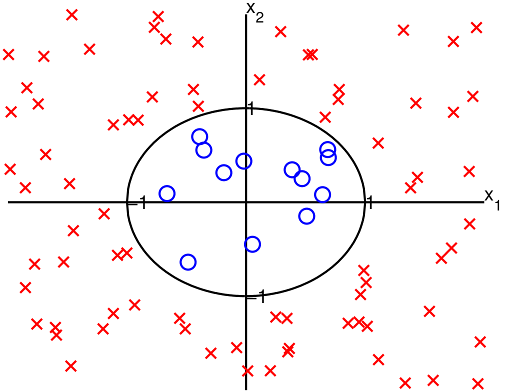
Μη γραμμικά σύνορα απόφασης



Σύνορο απόφασης $\Rightarrow x_1^2 + x_2^2 = 1$

Μη γραμμικότητα

Μη γραμμικά σύνορα απόφασης



$p(t=1|\mathbf{x}) = \frac{1}{1+e^{-w_0-w_1x_1-w_2x_2}} \Rightarrow p(t=1|\mathbf{x}) = \frac{1}{1+e^{-w_0-w_1x_1-w_2x_2-w_3x_1^2-w_4x_2^2}}$

αλλάζοντας τα αρχικό διάνυσμα εισόδου από $(1, x_1, x_2)$ σε $(1, x_1, x_2, x_1^2, x_2^2)$ προκύπτουν μη γραμμικά σύνορα απόφασης

Μη γραμμικότητα

$$\begin{array}{ccc} \text{Initial data vector} & \Rightarrow & \text{new feature vector} \\ (1, x_1, x_2) & & (1, x_1, x_2, x_1^2, x_2^2) \end{array}$$

Θα μπορούσαμε να προσθέσουμε και άλλα χαρακτηριστικά/features που αποτελούν όροι πολυωνύμου

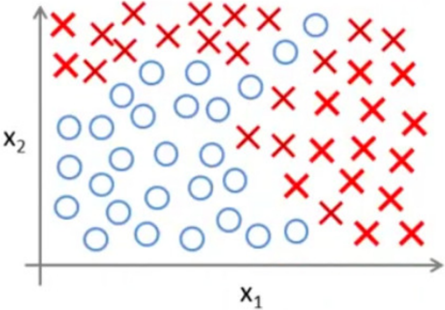
Π.χ.

$$(1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3, \ldots,)$$

Κάτι τέτοιο θα μας επέτρεπε να εκφράζουμε περισσότερο μη γραμμικές υποθέσεις

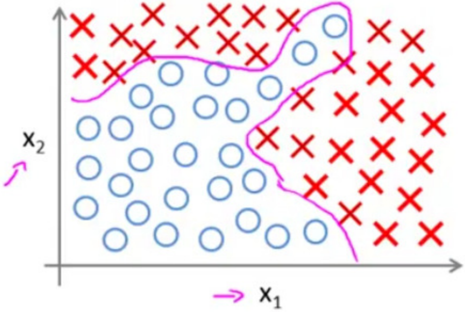
- Ωστόσο η χρήση πολυωνυμικών χαρακτηριστικών δεν είναι καλή ιδέα
- ας δούμε γιατί μέσω παραδειγμάτων

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις



Έστω τα δεδομένα σου σχήματος

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

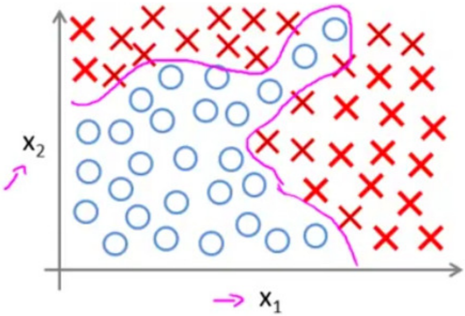


Θα θέλαμε να κατασκευάσουμε μοντέλο λογιστικής παλινδρόμησης της μορφής

$$p(t = 1|\mathbf{x}) = \sigma(y(\mathbf{x}, \mathbf{w}))$$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 + w_5 x_1 x_2^2 + \dots$$

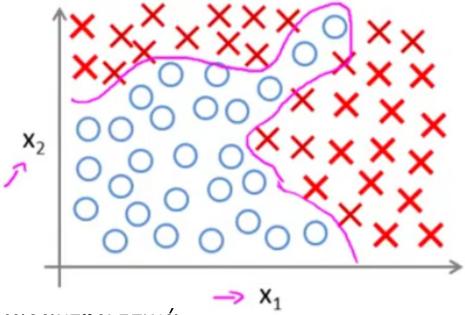
Μη γραμμικότητα με πολυωνυμικές συναρτήσεις



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 + w_5 x_1 x_2^2 + \ldots$$

Παρατηρούμε ότι αν και το δεδομένο εισόδου έχει διάσταση μόνο 2, ο αριθμός των πολυωνυμικών χαρακτηριστικών αυξάνει δραματικά με την πολυωνυμική τάξη

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

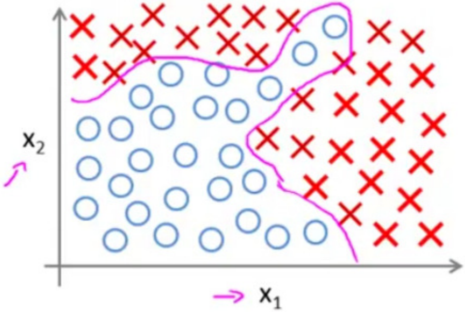


Αν είχαμε 100 χαρακτηριστικά

$$x_1$$
$$x_2$$
$$x_3$$
$$\dots$$
$$x_{100}$$

το διάνυσμα χαρακτηριστικών με όλους τους τετραγωνικούς όρους θα είχε διάσταση 5000 ($O(D^2)$), με όλους τους κυβικούς όρους θα είχε διάσταση 170000 ($O(D^3)$) κτλ

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις



Το διάνυσμα πολυωνυμικών χαρακτηριστικών αυξάνει εκθετικά και όποτε ο αριθμός των παραμέτρων που πρέπει να υπολογίσουμε αυξάνεται εκθετικά

- $\Rightarrow$ curse of dimensionality

Σε πολλές πραγματικές εφαρμογές η διάσταση του $\mathbf{x}$, δεν θα είναι της τάξης του 100, αλλά πολύ μεγαλύτερη $\Rightarrow$ ας δούμε ένα παράδειγμα από τον τομέα της υπολογιστικής όρασης computer vision

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

Έστω ότι θα θέλαμε κατηγοριοποίησουμε εικόνες σε δύο κατηγορίες

- δείχνει αυτοκίνητο
- δεν δείχνει αυτοκίνητο

Ένα τέτοιο πρόβλημα αποτελεί αντικείμενο της υπολογιστικής όρασης

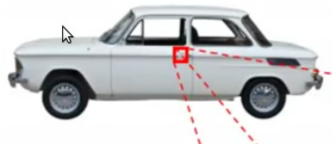
- σε αυτή την περιοχή συναντούμε ενδεχομένως τα δυσκολότερα προβλήματα μηχανικής μάθησης
- Γιατί η υπολογιστική όραση είναι τόσο δύσκολη;

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

Όταν έμεις βλέπουμε ένα αυτοκίνητο, ο υπολογιστής βλέπει κάτι άλλο!

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

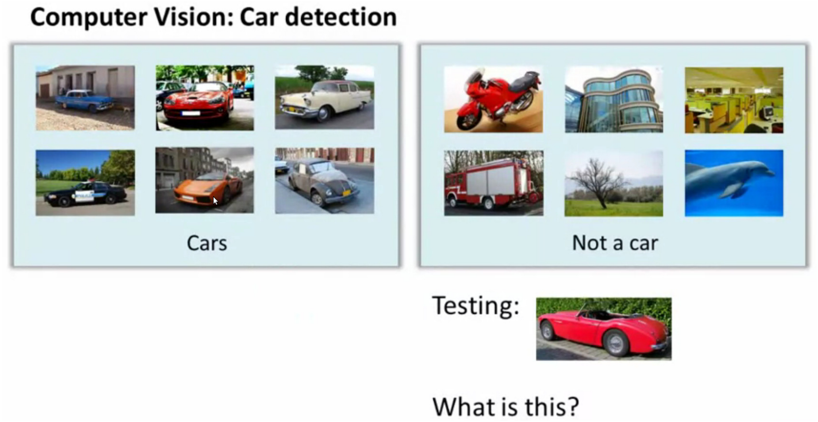
Όταν έμεις βλέπουμε ένα αυτοκίνητο, ο υπολογιστής βλέπει κάτι άλλο!



But the camera sees this:

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	75	65
20	43	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

Μη γραμμικότητα με πολυωνυμικές συναρτήσεις



Μη γραμμικότητα με πολυωνυμικές συναρτήσεις

Ακόμα και για μικρού μεγέθους εικόνες, έστω 50×50, η διάσταση του δεδομένου εισόδου είναι 2500 (για εικόνες με χρώμα είναι 7500)

Αν χρησιμοποιήσουμε πολυωνυμικά χαρακτηριστικά ως δευτέρου βαθμού, θα πρέπει να μάθουμε $O(7500^2)$ παραμέτρους!

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Γενικά θα θέλαμε να αντικαταστήσουμε κάθε αρχικό δεδομένο εισόδου $\mathbf{x}$ με νέο feature vector

$$\boldsymbol{\phi} = (1, \phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$$

$\boldsymbol{\phi}$ μπορεί να εκληφθεί ως το νέο (μετασχηματισμένο) δεδομένο

Τα πολυωνυμικά feature vectors αποτελούν ειδική περίπτωση

Το μοντέλο έπειτα είναι γραμμικό ως προς το $\boldsymbol{\phi}$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

Αλλά δεν είναι γραμμικό ως προς το $\mathbf{x}$!

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Feature vector

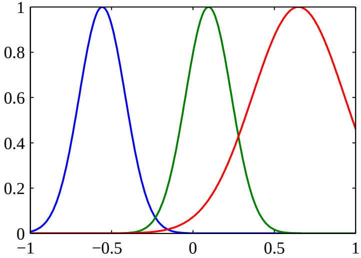
$$\boldsymbol{\phi} = (1, \phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$$

Μοντέλο έπειτα είναι γραμμικό ως προς το $\boldsymbol{\phi}$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

Θα θέλαμε να ορίζουμε το $\boldsymbol{\phi}$ ώστε το M να μην αυξάνει δραματικά με τη διάσταση του $\mathbf{x}$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

Για features συνήθως χρησιμοποιούμε ακτινικές (δηλ. που έχουν μια τοπική εμβέλεια) συναρτήσεις βάσης

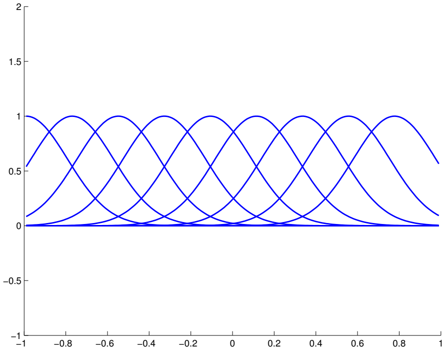
$$\phi_j(\mathbf{x}) = h\left(||\mathbf{x} - \boldsymbol{\mu}_j||\right)$$

Οι συνηθέστερες είναι οι Gaussian συναρτήσεις βάσεις

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Προκειμένου να κατανοήσουμε τα μοντέλα αυτά μπορούμε να παράγουμε τυχαίες συναρτήσεις

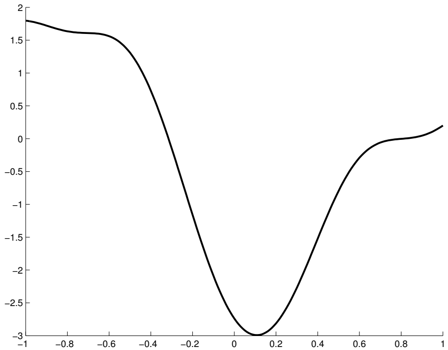


$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Για να δούμε τι συναρτήσεις παίρνουμε ας δώσουμε τυχαίες τιμές στις παραμέτρους $\mathbf{w}$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

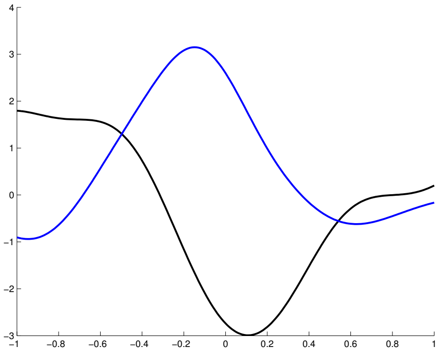


$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Στο σχήμα φαίνεται μία συνάρτηση $y(\mathbf{x}, \mathbf{w})$ που προέκυψε επιλέγοντας τυχαίες τιμές για το $\mathbf{w}$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

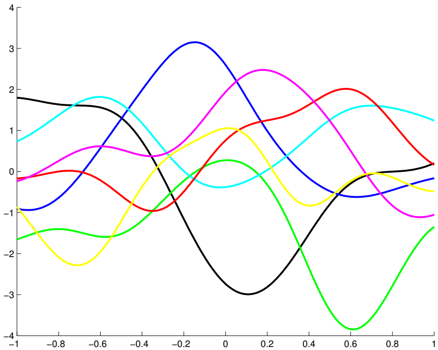


$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Στο σχήμα φαίνονται δύο διαφορετικές συναρτήσεις $y(\mathbf{x}, \mathbf{w})$ με τυχαία $\mathbf{w}$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

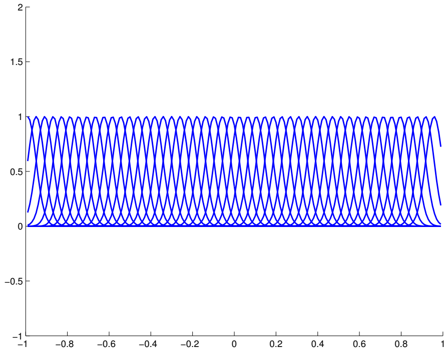


$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Στο σχήμα φαίνονται διάφορες συναρτήσεις $y(\mathbf{x}, \mathbf{w})$ που προκύπτουν επιλέγοντας τυχαίους παραμέτρους $\mathbf{w}$ κάθε φορά

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

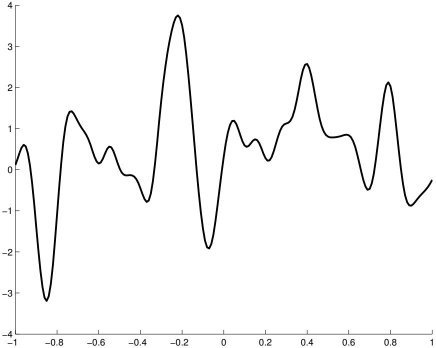


$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Αν αλλάζουμε το πλήθος των συναρτήσεων βάσης καθώς και την παράμετρο ℓ που καθορίζει την ακτίνα εμβέλειας της κάθε ακτινικής συνάρτησης $\Rightarrow$ θα προκύψουν διαφορετικού τύπου συναρτήσεις

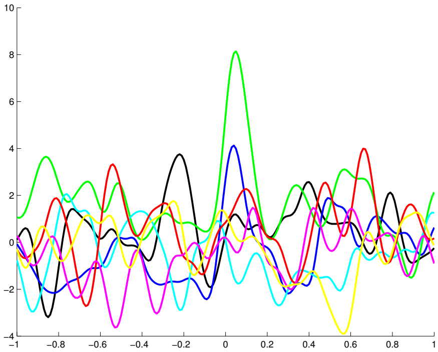
Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης



Οι συναρτήσεις αυτές έχουν μικρότερο lengthscale

- διαισθητικά το lengthscale είναι η απόσταση που πρέπει να διανύσουμε κατά μήκος του άξονα x προκειμένου να παρατηρήσουμε αλλαγή στην κατεύθυνση της συνάρτησης

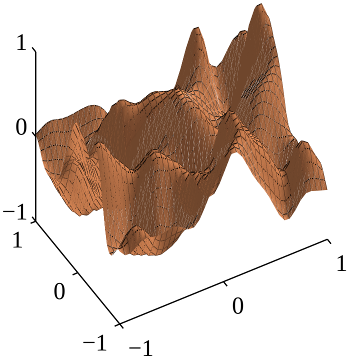
Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης



Οι συναρτήσεις αυτές έχουν μικρότερο lengthscale

- διαισθητικά το lengthscale είναι η απόσταση που πρέπει να διανύσουμε κατά μήκος του άξονα x προκειμένου να παρατηρήσουμε αλλαγή στην κατεύθυνση της συνάρτησης

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

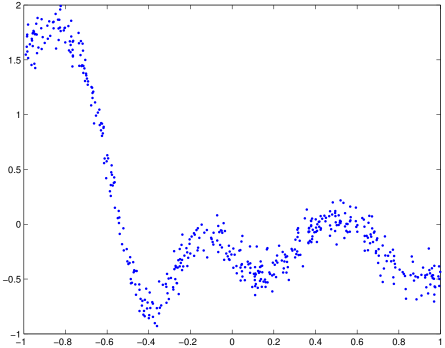
Στο σχήμα φαίνεται μια συνάρτηση στον δισδιάστατο χώρο ($\mathbf{x} = [x_1 \; x_2]$) που προκύπτει επιλέγοντας τυχαίες παραμέτρους $\mathbf{w}$

Οι αλγόριθμοι εκπαίδευσης με την τεχνική της μέγιστης πιθανοφάνειας καθώς και της Bayesian κανονικοποίησης εφαρμόζονται ακριβώς όπως μάθαμε στα προηγούμενα μαθήματα

απλώς αντικαθιστούμε το $\mathbf{x}_n$ με ϕ_n!

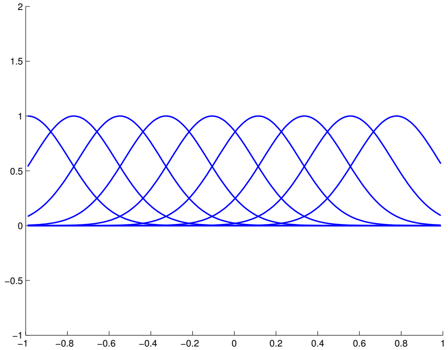
Ας δούμε κάποια παραδείγματα παλινδρόμησης και κατηγοριοποίησης

Παράδειγμα παλινδρόμησης



Έστω τα δεδομένα του σχήματος

Παράδειγμα παλινδρόμησης



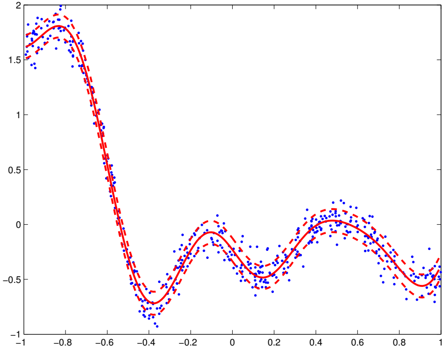
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Υπόθετουμε ένα μοντέλο με τις Gaussian ακτινικές συναρτήσεις βάσης που απεικονίζονται στην εικόνα

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

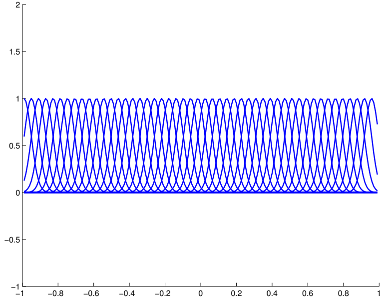
Παράδειγμα παλινδρόμησης



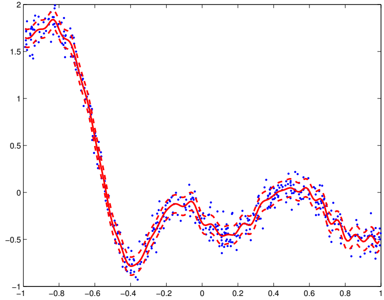
Η λύση που προκύπτει με την τεχνική της μέγιστης πιθανοφάνειας

- Η μεσαία κόκκινη γραμμή δείχνει την μέση πρόβλεψη, δηλ. την $y(x, \mathbf{w}) = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_M\phi_M(\mathbf{x})$
- Η διακεκομμένες γραμμές βρίσκονται σε απόσταση μιας τυπικής απόκλισης, δηλ. απόστασης $1/\sqrt{\beta}$

Παράδειγμα παλινδρόμησης



$M = 50$ ακτινικές συναρτήσεις

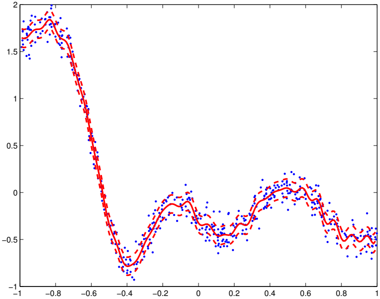


Το εκπαιδευμένο μοντέλο

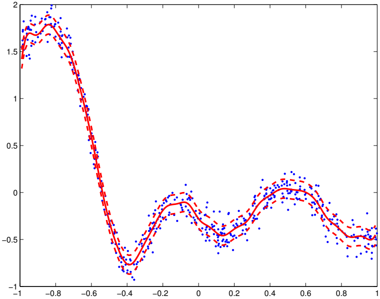
Αν αλλάξουμε το σύνολο των ακτινικών συναρτήσεων βάσης θα προκύψει διαφορετική λύση

- που στην προκειμένη περιπτώση εκδηλώνει σημάδια υπερεκπαίδευσης!

Παράδειγμα παλινδρόμησης



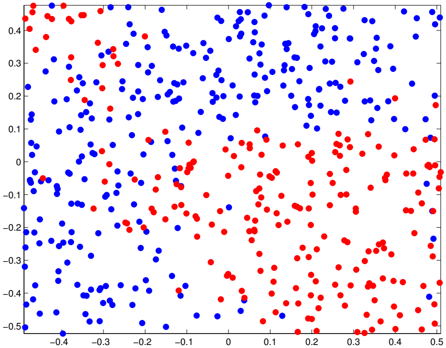
$\lambda = 0$



$\lambda = 1$

Για να αποφύγουμε την υπερεκπαίδευση θα μπορούσαμε να εφαρμόσουμε κανονικοποίηση

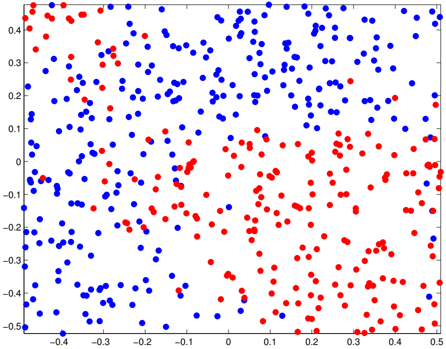
Παράδειγμα κατηγοριοποίησης



Στο σχήμα φαίνονται τα δεδομένα εκπαίδευσης δύο κατηγοριών

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Παράδειγμα κατηγοριοποίησης



Χρησιμοποιούμε το μοντέλο της λογιστικής παλινδρόμησης

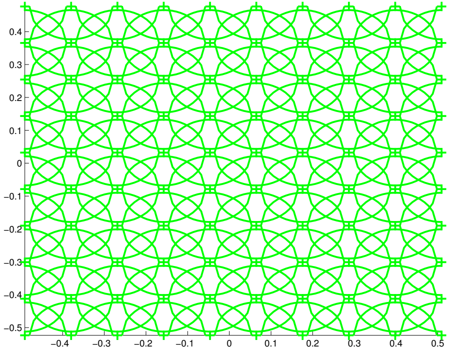
$$p(t=1|\phi_n, \mathbf{w}) = \sigma\left(\mathbf{w}^T \phi_n\right)$$

όπου το feature vector ορίζεται από τις Gaussian ακτινικές συναρτήσεις βάσης

$$\phi_n = (1, \phi_1(\mathbf{x}_n), \ldots, \phi_M(\mathbf{x}_n)), \quad \phi_j(\mathbf{x}) = e^{-\frac{1}{2\ell^2}||\mathbf{x}-\boldsymbol{\mu}_j||^2}$$

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

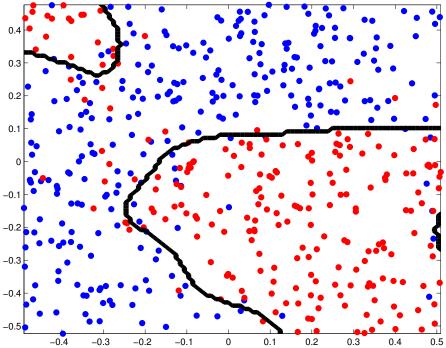
Παράδειγμα κατηγοριοποίησης



Αφού τα δεδομένα εισόδου βρίσκονται στον δισδιάστατο χώρο οι συναρτήσεις βάσης είναι συναρτήσεις δύο μεταβλητών

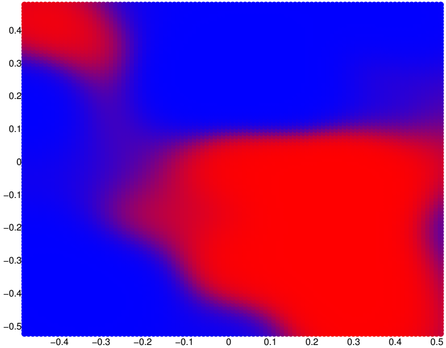
Έστω ότι χρησιμοποιούμε τις συναρτήσεις βάσης που φαίνονται στο σχήμα

Παράδειγμα κατηγοριοποίησης



Μεγιστοποιώντας την λογαριθμική πιθανοφάνεια μέσω του αλγορίθμου ανοδικής κλίσης προκύπτει το σύνορο απόφασης του παραπάνω σχήματος

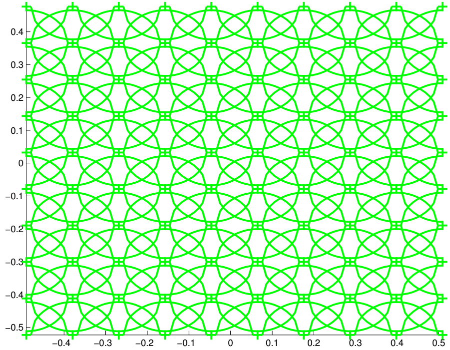
Παράδειγμα κατηγοριοποίησης



Πρόβλεψη σε κάθε σημείο του χώρου

- Μπλε χρώμα σημαίνει ότι η πιθανότητα $p(t = 1|\phi(\mathbf{x}), \mathbf{w})$ στο αντίστοιχο σημείο $\mathbf{x}$ είναι κοντά στο 1, ενώ κόκκινο χρώμα σημαίνει ότι η πιθανότητα αυτή είναι κοντά στο 0

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης



Στο προηγούμενο παράδειγμα χρησιμοποιήσαμε πάρα πολλές συναρτήσεις βάσης προκειμένου να καλύψουμε όλο το χώρο

- $\Rightarrow$ αυτό πάσχει από το πρόβλημα της κατάρας της διάστασης όπως η περίπτωση των πολυωνυμικών χαρακτηριστικών
- Πώς θα μπορούσαμε να αποφύγουμε το πρόβλημα της κατάρας της διάστασης;

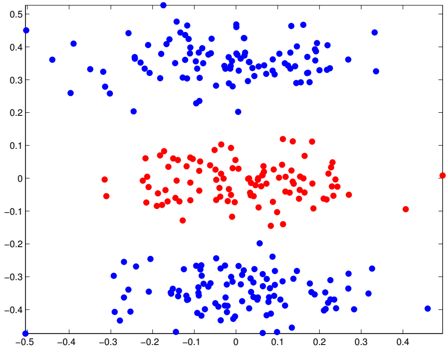
Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Πώς θα μπορούσαμε να αποφύγουμε το πρόβλημα της κατάρας της διάστασης; Η λύση είναι πολύ απλή

- Δεν χρειάζεται να γεμίσουμε άσκοπα το χώρο με συναρτήσεις βάσης
- $\Rightarrow$ αρκεί να βάλουμε τις συναρτήσεις βάσης εκεί που υπάρχουν δεδομένα εισόδου

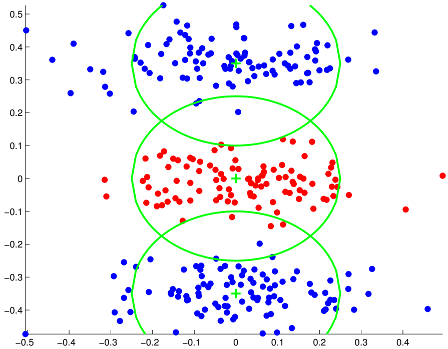
Ας δούμε ένα παράδειγμα

Χρήση λίγων συναρτήσεων βάσης



Έστω τα δεδομένα εκπαίδευσης δύο κατηγοριών

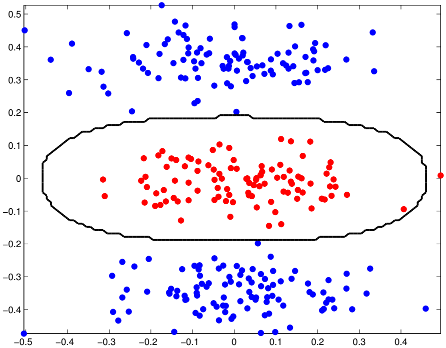
Χρήση λίγων συναρτήσεων βάσης



Χρησιμοποιούμε μόνο 3 συναρτήσεις που τοποθετούνται εκεί που τα δεδομένα έχουν μεγάλη πυκνότητα

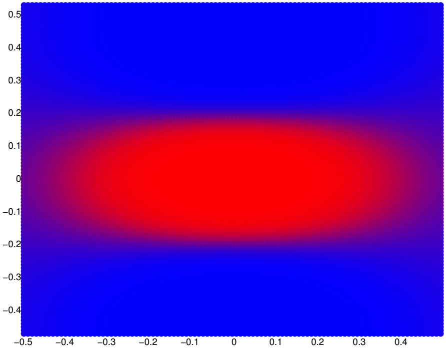
Τα κέντρα αυτών των συναρτήσεων βάσης θα μπορούσαν να βρεθούν εφαρμόζοντας αλγόριθμους ομαδοποίησης

Χρήση λίγων συναρτήσεων βάσης



Σύνορο απόφασης

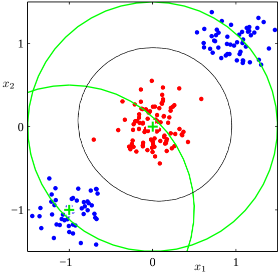
Χρήση λίγων συναρτήσεων βάσης

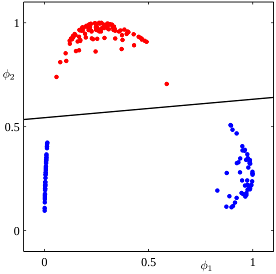


Πρόβλεψη σε όλο το χώρο

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

Χρήση λίγων συναρτήσεων βάσης





Ακόμα και με δύο συναρτήσεις βάσης (σε ένα όμοιο πρόβλημα) μπορούμε να πετύχουμε εύκολο διαχωρισμό των δύο κατηγοριών

Πρόσεξε ότι τα αρχικά δεδομένα είναι μη γραμμικά διαχωρίσιμα!

Ωστόσο τα μετασχηματισμένα δεδομένα (τα οποία είναι και αυτά δισδιάστατα λόγω ότι χρησιμοποιήσαμε δύο ακτινικές συναρτήσεις) είναι γραμμικά διαχωρίσιμα!

Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + w_2\phi_2 + \ldots + w_M\phi_M$$

$$\phi_j(\mathbf{x}) = h\left(||\mathbf{x} - \boldsymbol{\mu}_j||\right)$$

Στην πράξη επιλέγουμε το M και τα κέντρα $\{\boldsymbol{\mu}_j\}_{j=1}^{M}$ των συναρτήσεων βάσης με δύο εναλλακτικούς (αλλά όμοιους) τρόπους

1. Είτε μέσω ομαδοποίησης των δεδομένων εισόδου $\Rightarrow$ σε αυτή την περίπτωση θα έχουμε $M < N$ και τα $\{\boldsymbol{\mu}_j\}_{j=1}^{M}$ θα είναι τα κέντρα των ομάδων
2. Είτε χρησιμοποιώντας όσες συναρτήσεις βάσης όσο και το πλήθος των δεδομένων εκπαίδευσης, δηλ.
   - $\phi_n(\mathbf{x}) = h\left(||\mathbf{x} - \mathbf{x}_n||\right)$, $n = 1, \ldots, N$
     όπου $\mathbf{x}_n$, $n = 1, \ldots, N$ τα δεδομένα του συνόλου εκπαίδευσης

Νευρωνικά δίκτυα

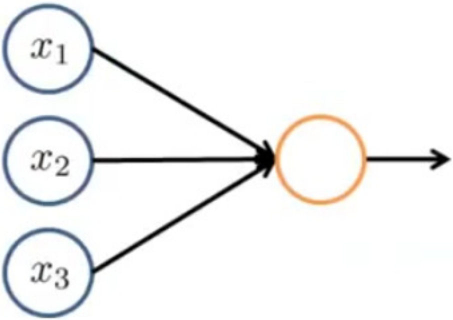
Μη γραμμικά μοντέλα με ακτινικές συναρτήσεις βάσης είναι πολύ ευέλικτα μοντέλα

Ωστόσο έχουν τον περιορισμό ότι το feature vector ϕ ειναι συγκεκριμένο $\Rightarrow$ δηλ. δεν μαθαίνουμε την μορφή του feature vector

Όταν προσπαθούμε να μάθουμε την μορφή του feature vector ϕ με ευέλικτο τρόπο τότε ουσιαστικά οδηγούμαστε στα νευρωνικά δίκτυα!

Θα κουβεντιάσουμε την μεθοδολογία χωρίς πολλά μαθηματικά $\Rightarrow$ δίνοντας μόνο την κεντρική ιδεά

Νευρωνικά δίκτυα

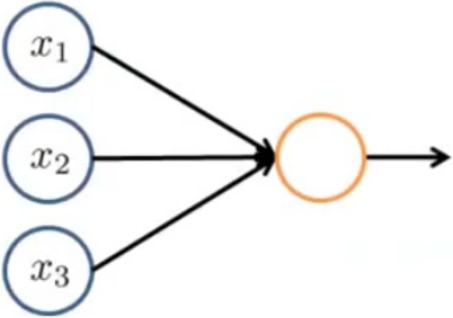


Ενα νευρωνικό δίκτυο προκύπτει ως η ιεραρχική σύνθεση της λογιστικής υπολογιστικής μονάδας (ή κάποιας όμοιας μη γραμμικής συνάρτησης)

$$\sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-w_0 - w_1 x_1 - w_2 x_2 - \dots}}$$

Το $\mathbf{x}$ είναι αυθαίρετο στο παράδειγμα αυτό, δηλ. ένα οποιοδήποτε input όχι απαραίτητα το δεδομένο εισόδου

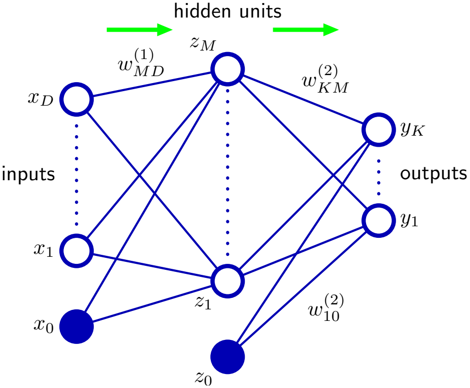
Νευρωνικά δίκτυα



$$\sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-w_0 - w_1 x_1 - w_2 x_2 - \dots}}$$

Στην ορολογία των νευρωνικών δικτυών η μη γραμμική συνάρτηση $\sigma(\cdot)$ ονομάζεται activation function

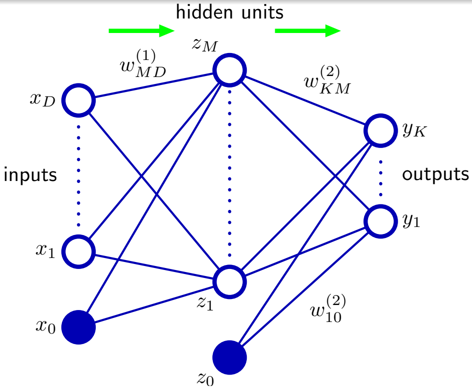
Νευρωνικά δίκτυα



Ένα νευρωνικό δίκτυο στην απλούστερη μορφή αποτελείται από τρία layers

- input layer: τα δεδομένα εισόδου
- hidden layer: ένα ενδιάμεσο επίπεδο
- output layer: οι συναρτήσεις που παίρνουμε ως έξοδο

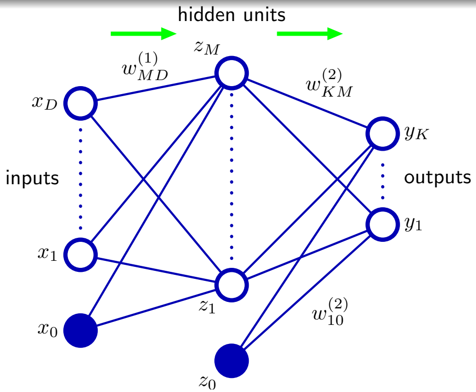
Νευρωνικά δίκτυα



Η γενική δομή ενός νευρωνικού δικτύου με ένα input layer, ένα hidden layer και ένα output layer και με activation function την $\sigma(\cdot)$

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

όπου $\mathbf{w}$ είναι το σύνολο όλων των παραμέτρων



Το ποια θα είναι η μορφή της activation function σε κάθε έξοδο εξαρτάται από το τι θέλουμε να προβλέψουμε. Αν στην k έξοδο θέλουμε να προβλεψουμε συνεχή τιμές (παλινδρόμηση), τότε η συνάρτηση αυτή είναι η γραμμική (δηλ. ουσιαστικά δεν χρειάζεται να χρησιμοποιήσουμε activation function)

$$y_k(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_{kj}^{(2)} \sigma \left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)}$$

Νευρωνικά δίκτυα

Η εκπαίδευση του νευρωνικού δικτύου γίνεται ομοίως με τα προηγούμενα μοντέλα. Π.χ.

- Αν θέλουμε να λύσουμε το πρόβλημα παλινδρόμησης με ένα νευρωνικό δίκτυο η τεχνική της μέγιστης πιθανοφάνειας μας οδηγεί στην ελαχιστοποίηση της

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, \mathbf{w}) - t_n\right)^2$$

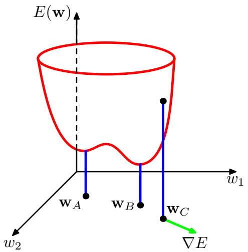
όπου $y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j^{(2)} \sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_0^{(2)}$

- Αν θέλουμε να λύσουμε ένα πρόβλημα κατηγοριοποίησης δύο κατηγοριών οδηγούμαστε

$$E(\mathbf{w}) = \sum_{n=1}^{N} t_n \log y(\mathbf{x}_n, \mathbf{w}) + (1 - t_n) \log\left(1 - y(\mathbf{x}_n, \mathbf{w})\right)$$

όπου $y(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_j^{(2)} \sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_0^{(2)}\right)$

Νευρωνικά δίκτυα



Η εκπαίδευση απαιτεί επίπονη βελτιστοποίηση όπου η συναρτήσεις κόστους μπορούν (αντιθέτως με τα γραμμικά ως προς τις παραμέτρους μοντέλα) να έχουν πολλά τοπικά ακρότατα

Ωστόσο για κάθε συνάρτηση κόστους $E(\mathbf{w})$ εργαζόμαστε όπως και στην γραμμική λογιστική παλινδρόμηση, δηλ. βρίσκουμε το διάνυσμα των μερικών παραγώγων $\nabla E(\mathbf{w})$ και εφαρμόζουμε κάποιο αλγόριθμο αριθμητικής βελτιστοποίησης

Ο υπολογισμός του $\nabla E(\mathbf{w})$ γίνεται εφαρμόζοντας τον κανόνα παραγώγισης της αλυσιδας με ένα έξυπνο τρόπο που ονομάζεται error backpropagation

Νευρωνικά δίκτυα

Σχέση με τα προηγούμενα μοντέλα

- Το αρχικό γραμμικό (ως προς παραμέτρους και εισόδους) μοντέλο είχε τη μορφή

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{D} w_j x_j + w_0$$

- Το γραμμικό μόνο ως προς τις παραμέτρους μοντέλο είχε τη μορφή

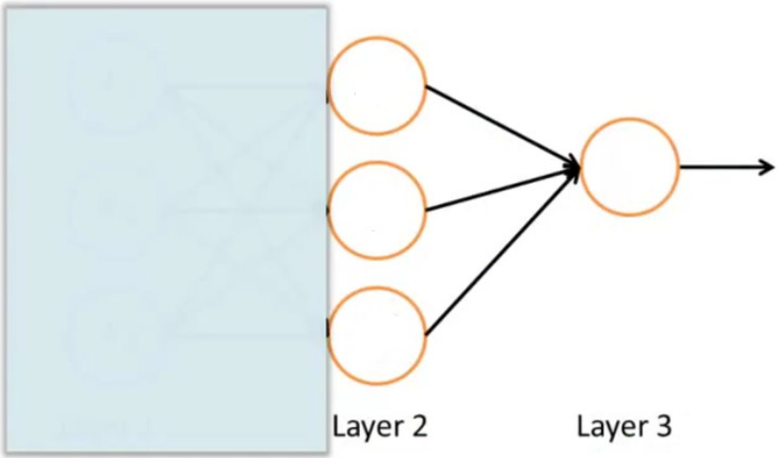
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j + w_0$$

όπου $\phi = (1, \phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x}))$ ήταν το feature vector

- Τα νευρωνικά δίκτυα πηγαίνουν ένα βήμα πιο πέρα υπό την έννοια ότι επιδιώκουν να μάθουν το feature vector

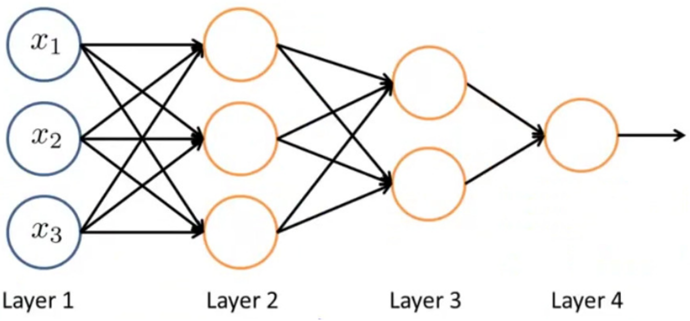
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j^{(2)} \sigma \left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_0^{(2)}$$

Νευρωνικά δίκτυα



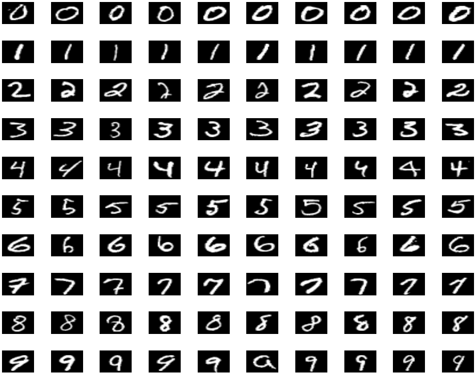
Ο,τιδήποτε υπάρχει πριν το τελευταίο hidden layer ουσιαστικά μαθαίνει απευθείας από τα δεδοδομένα ένα feature vector

Νευρωνικά δίκτυα



Θα μπορούσαμε να έχουμε πολλά hidden layers όποτε το feature vector που μαθαίνουμε να έχει μια ιεραρχική δομή

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

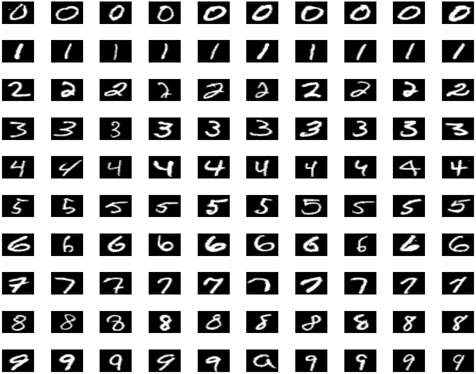


Έχουμε ένα σύνολο δεδομένων εκπαίδευσης που αποτελείται από 60000 χειρόγραφα ψηφία (10 κατηγορίες)

Κάθε δεδομένο εισόδου αποτελεί μια εικόνα διάστασης 28×28, οπότε κάθε δεδομένο εισόδου έχει διάσταση 784

Νευρωνικά δίκτυα

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες



Πρότου χρησιμοποιήσουμε ένα νευρωνικό δίκτυο ας δοκιμάσουμε ένα πιο απλό μοντέλο

- Αυτό της γραμμικής λογιστικής παλινδρόμησης για πολλές κατηγορίες

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

Θα εφαρμόσουμε λογιστική παλινδρόμηση για πολλές κατηγορίες

$$p(\mathbf{t}_n|\mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

όπου

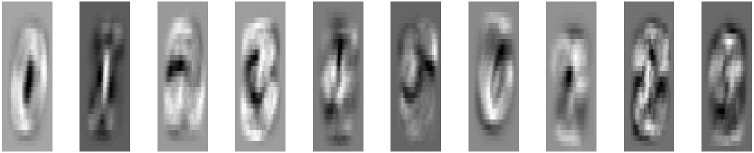
$$y_{nk} = \frac{e^{\mathbf{w}_k^T \mathbf{x}_n}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T \mathbf{x}_n}}$$

Για να εκπαιδεύσουμε τις παραμέτρους $\{\mathbf{w}_k\}_{k=1}^{K}$, μπορούμε να μεγιστοποιήσουμε την λογαριθμική πιθανοφάνεια

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}$$

Το μοντέλο αυτό μπορεί να θεωρηθεί ως ένα νευρωνικό δίκτυο με K εξόδους, D εισόδους και χωρίς hidden layer

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

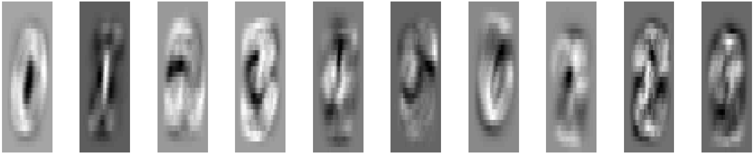


Στο σχήμα απεικονίζονται οι τιμές των παραμέτρων $\mathbf{w}_k$ (χωρίς το bias $w_{k,0}$) για τις 10 κατηγορίες

$$w_{k,d}, d = 1, \ldots, 784$$

που προκύπτουν από την εκπαίδευση με μέγιστη πιθανοφάνεια

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες



Αυτό που κάνει η λογιστική παλινδρόμηση στην προκειμένη περίπτωση είναι να περιγράψει κατά κάποιο τρόπο την κάθε κατηγορία με ένα template

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

Κατά το έλεγχο του συστήματος κατηγοριοποίησης χρησιμοποιούμε 10000 δεδομένα (τα οποία προφανώς είναι διαφορετικά από τα 60000 δεδομένα που χρησιμοποιήθηκαν για εκπαίδευση)

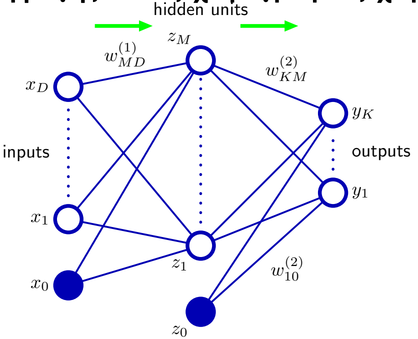
Η τελική τιμή της συνάρτησης κόστους ήταν

$$\mathcal{L} = -18205.100991$$

Το ολικό σφάλμα της μεθόδου ήταν

$$error = 8.18\%$$

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες



Χρησιμοποιούμε τώρα ένα νευρωνικό δίκτυο με 500 υπολογιστικές μονάδες στο hidden layer (δηλ. $M = 500$). Η εκπαίδευση γίνεται με την ίδια συνάρτηση κόστους

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}$$

με μόνη διαφορά ότι τώρα το y_{nk} ορίζεται μέσω του νευρωνικού δικτύου

Παράδειγμα εφαρμογής στους χειρόγραφους χαρακτήρες

Η τελική τιμή της συνάρτησης κόστους ήταν

$$\mathcal{L} = -430.170827$$

Το ολικό σφάλμα στα 10000 δεδομένα ελέγχου ήταν

$$error = 3.31\%$$

Οπότε παρατηρούμε ότι το παραπάνω νευρωνικό δίκτυο έχει πολύ καλύτερη επίδοση από την λογιστική παλινδρόμηση

Επίλογος

- Διάβασμα για το σπίτι: . Bishop: 3.1 μέχρι σελίδα 140, 6.3 (μόνο τη σελίδα 299) σελίδες 225-237
- Επόμενο μάθημα: K κοντινότεροι γείτονες, περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης, naive Bayes