Μηχανική Μάθηση

Μιχάλης Τίτσιας

Διάλεξη 3ή

Γραμμικά μοντέλα παλινδρόμησης και λογιστικής παλινδρόμησης

Τι θα πούμε στο σημερινό μάθημα

Δεδομένα:

- Σας αυτό το μάθημα θα μας αποσχολήσουν προβλήματα μάθησης με επίβλεψη όπου στα δεδομένα οι έξοδοι παίρνουν πραγματικές τιμές (παλινδρόμηση) ή δυαδικές τιμές (κατηγοριοποίηση με δύο κατηγορίες)

Μοντέλα/υπόθεσεις:

- Θα ασχοληθούμε αποκλειστικά με πιθανοτικά γραμμικά μοντέλα παλινδρόμησης και κατηγοριοποίησης

Αλγόριθμοι εκπαίδευσης:

- Συναρτήσεις κόστους θα προκύψουν από μέγιστη πιθανοφάνεια και Bayesian κανονικοποίηση
- Αλγόριθμοι βελτιστοποίησης: ελάχιστα τετραγώνα, αλγόριθμος ανοδικής κλίσης

Τι θα πούμε στο σημερινό μάθημα

$$E(\mathbf{w}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2}_{\text{data term}} + \underbrace{\lambda\frac{||\mathbf{w}||^2}{2}}_{\text{regularization term}}$$

1. Πώς επιλέγουμε την τιμή του λ; $\Rightarrow$ μιλήσαμε για αυτό στο προηγούμενο μάθημα
2. Ποια είναι η πιθανοτική ερμηνεία πίσω από την χρήση της $\frac{1}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2$ και $\lambda\frac{||\mathbf{w}||^2}{2}$;
3. Ποια θα είναι η μορφή της $E(\mathbf{w})$ στο πρόβλημα κατηγοριοποίησης;

Στο σημερινό μάθημα θα κουβεντιάσουμε για το 1 και 2

Περιεχόμενα

- Πιθανοτική γραμμική παλινδρόμηση
- Εκπαίδευση με μέγιστη πιθανόφανεια
- Εκπαίδευση με κανονικοποίηση
- Κατηγοριοποίηση
- Γραμμική λογιστική παλινδρόμηση
- Αλγόριθμο ανοδικής κλίσης
- Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση
- Λογιστική παλινδρόμηση για πολλές κατηγορίες

Πιθανοτική γραμμική παλινδρόμηση

- Έστω ότι έχουμε τα ακόλουθα δεδομένα εκπαίδευσης

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

  όπου κάθε $\mathbf{x}_n$ είναι ένα δεδομένο εισόδου και t_n το αντίστοιχο δεδομένο εξόδου
- Πρόβλημα μάθησης: Κατασκευή ενός συστήματος που να μαθαίνει να προβλέπει την έξοδο t_* για κάθε άγνωστο δεδομένο εισόδου $\mathbf{x}_*$

Πιθανοτική γραμμική παλινδρόμηση

Μοντέλο: Υποθέτουμε μια γραμμική σχέση

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

Θόρυβος: Υποθέτουμε ότι η κάθε εξόδος t είναι δομή συν ένα σφάλμα (ονομάζεται θόρυβος) που ακολουθεί Gaussian κατανομή

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon | 0, \beta^{-1})$$

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

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

το οποίο αποτελεί το πιθανοτικό μοντέλο για γραμμική παλινδρόμηση με Gaussian θόρυβο

- μας λέει πως η έξοδος παράγεται στοχαστικά μέσω της εισόδου

Εκπαίδευση με μέγιστη πιθανόφανεια

Θέλουμε να εκτιμήσουμε τις παραμέτρους $(\mathbf{w}, \beta)$ βελτιστοποιώντας μια συνάρτηση κόστους $\Rightarrow$ θα χρησιμοποιήσουμε την τεχνική της μέγιστης πιθανοφάνειας

- Από κοινού κατανομή: Υποθέτουμε ότι το κάθε t_n παράγεται ανεξάρτητα (από όλα τα άλλα ts) δοθέντος του $\mathbf{x}_n$

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

- Πιθανοφάνεια: Θεωρούμε το $p(\mathbf{t}|X, \mathbf{w}, \beta)$ ως συνάρτηση των παραμέτρων $(\mathbf{w}, \beta)$. Υπό αυτό το πρίσμα την ονομάζουμε πιθανοφάνεια
- Λογαριθμική πιθανοφάνεια: Θέλουμε να επιλέξουμε τα $(\mathbf{w}, \beta)$ που μεγιστοποιούν την πιθανοφάνεια ή ισοδύναμα τον λογάριθμό της

$$\mathcal{L}(\mathbf{w}, \beta) = \log \prod_{n=1}^{N} \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(t_n - y(\mathbf{x}_n, \mathbf{w})\right)^2\right\}$$

Εκπαίδευση με μέγιστη πιθανόφανεια

Θέλουμε να επιλέξουμε τα $(\mathbf{w}, \beta)$ που μεγιστοποιούν την πιθανοφάνεια ή ισοδύναμα τον λογάριθμό της

$$\mathcal{L}(\mathbf{w}, \beta) = \log \prod_{n=1}^{N} \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(t_n - y(\mathbf{x}_n, \mathbf{w})\right)^2\right\}$$

ή

$$\mathcal{L}(\mathbf{w}, \beta) = \sum_{n=1}^{N} \log \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(t_n - y(\mathbf{x}_n, \mathbf{w})\right)^2\right\}$$

ή

$$\mathcal{L}(\mathbf{w}, \beta) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\beta) - \frac{\beta}{2}\sum_{i=1}^{N}\left(t_n - y(\mathbf{x}_n, \mathbf{w})\right)^2$$

Εκπαίδευση με μέγιστη πιθανόφανεια

$$\mathcal{L}(\mathbf{w}, \beta) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\beta) - \frac{\beta}{2}\sum_{i=1}^{N}(t_n - y(\mathbf{x}_n, \mathbf{w}))^2$$

Μεγιστοποίηση ως προς $\mathbf{w}$ ισοδυναμεί με ελαχιστοποίηση της συνάρτησης ελαχίστων τετραγώνων

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

Ως προς β θα δώσει

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

Εκπαίδευση με μέγιστη πιθανόφανεια

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

Ας βγάλουμε την λύση για το $\mathbf{w}$ για την απλή περίπτωση που το $\mathbf{x}$ είναι μονοδιάστατο, δηλ.

$$y(x, \mathbf{w}) = w_0 + w_1 x$$

Έχουμε

$$E(w_0, w_1) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)^2$$

Παίρνωντας μερικές παραγώγους

$$\frac{\partial E(w_0, w_1)}{\partial w_0} = -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)$$

$$\frac{\partial E(w_0, w_1)}{\partial w_1} = -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)\, x_n$$

Εκπαίδευση με μέγιστη πιθανόφανεια

- Εξισώνοντας με το μηδέν έχουμε

$$\sum_{n=1}^{N} (t_n - w_0 - w_1 x_n) = 0$$

$$\sum_{n=1}^{N} (t_n - w_0 - w_1 x_n)\, x_n = 0$$

ή

$$N w_0 + w_1 \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} t_n$$

$$w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{N} x_n^2 = \sum_{n=1}^{N} t_n x_n$$

το οποίο είναι ένα γραμμικό σύστημα με δύο εξισώσεις και δύο αγνώστους (w_0, w_1)

Εκπαίδευση με μέγιστη πιθανόφανεια

$$N w_0 + w_1 \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} t_n$$

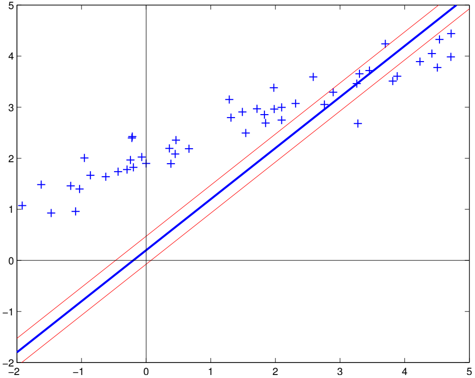
$$w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{N} x_n^2 = \sum_{n=1}^{N} t_n x_n$$

Το σύστημα έχει ως λύση την

$$w_0 = \frac{1}{N} \sum_{n=1}^{N} t_n - w_1 \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$w_1 = \frac{N \sum_{n=1}^{N} t_i t_n - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} t_n}{N \sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n \right)^2}$$

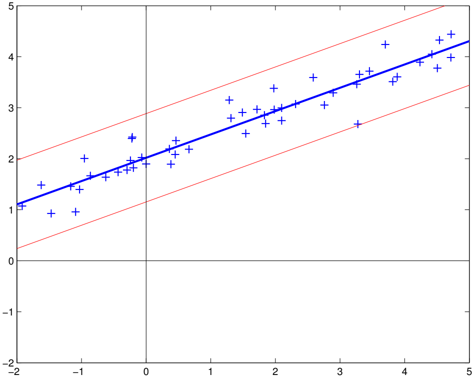
Εκπαίδευση με μέγιστη πιθανόφανεια



Η κεντρική μπλε γραμμή δείχνει την $w_0 + w_1 x$, ενώ οι δύο κόκκινες γραμμές βρίσκονται σε απόσταση μιας τυπικής απόκλισης (δηλ. $1/\sqrt{\beta}$) από την μπλε γραμμή

Οι τιμές (w_0, w_1) δεν έχουν επιλεγεί σωστά (η γραμμή δεν έχει «ταιριάξει» στα δεδομένα) για αυτό και η λογαριθμική πιθανοφάνεια είναι μικρή $\mathcal{L} = -736.2031$

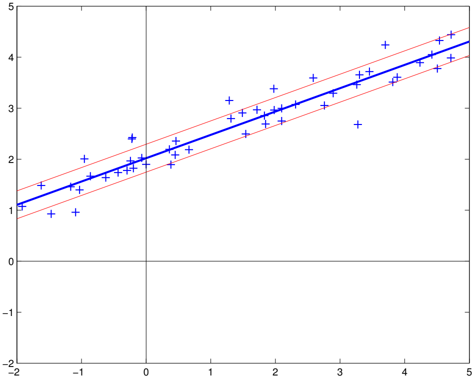
Εκπαίδευση με μέγιστη πιθανόφανεια



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) έχουν επιλεγεί βέλτιστα, το β δεν είναι βέλτιστο (οι κόκκινες γραμμές της τυπικής απόκλισης δεν ταιριάζουν στα δεδομένα). Η πιθανοφάνεια είναι $\mathcal{L} = -41.2290$

Εκπαίδευση με μέγιστη πιθανόφανεια



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 όλα βέλτιστα. Η λογαριθμική πιθανοφάνεια είναι η μέγιστη $\mathcal{L} = -6.1644$

Εκπαίδευση με μέγιστη πιθανόφανεια

Όταν η μεταβλητή εισόδου είναι διάνυσμα

- Η μεγιστοποίηση οδηγεί σε $D+1 \times D+1$ γραμμικό σύστημα

$$\left[\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T\right] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \dots \end{bmatrix} = \sum_{n=1}^{N} t_n \mathbf{x}_n \Rightarrow \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \dots \end{bmatrix} = \left[\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T\right]^{-1} \sum_{n=1}^{N} t_n \mathbf{x}_n$$

  όπου $\mathbf{x}_n = [1\ x_{n,1}\ x_{n,2} \dots x_{n,D}]^T$

- Με συμβολισμό διανυσμάτων και πινάκων αυτό γράφεται ως

$$\mathbf{w} = (X^T X)^{-1} X^T t$$

  όπου X είναι $N \times D+1$ πίνακας που περιέχει σε κάθε γραμμή ένα $\mathbf{x}_n = [1\ x_{n,1}\ x_{n,2} \dots x_{n,D}]^T$

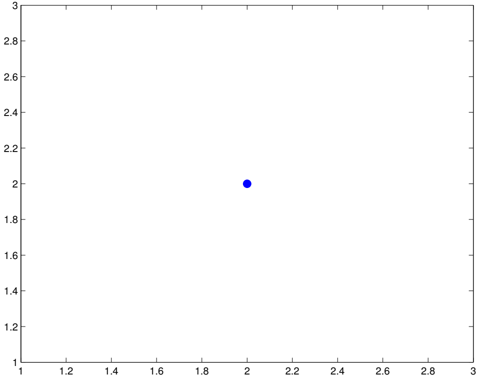
Εκπαίδευση με μέγιστη πιθανόφανεια

$$\mathbf{w} = (X^T X)^{-1} X^T t$$

- Η λύση είναι μοναδική όταν ο $D+1 \times D+1$ πίνακας $(X^T X)$ είναι full rank
- Αν ο αριθμός των δεδομένων είναι μικρός, και συγκεκριμένα όταν $N < D+1$, το σύστημα έχει άπειρες λύσεις
- $\Rightarrow$ αυτό υποννοεί υπερεκπαίδευση μια και θα μπορούμε να παρεμβάλλουμε τα δεδομένα εκπαίδευσης με άπειρους τρόπους

Εκπαίδευση με μέγιστη πιθανόφανεια

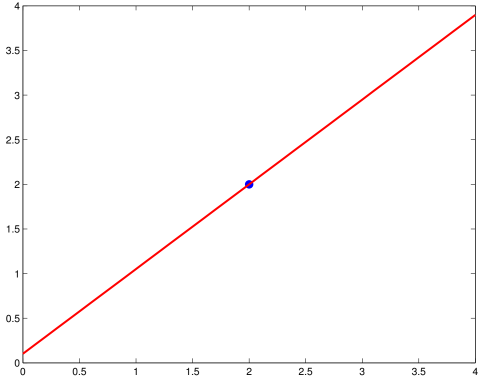
Παράδειγμα άπειρων λύσεων



Έστω $D = 1$, δηλ. το γραμμικό μοντέλο έχει τη μορφή $y(x, \mathbf{w}) = w_0 + w_1 x$. Είπαμε ότι για αριθμό δεδομένων $N < D + 1 = 2$ (δηλ. στην προκειμένη περίπτωση για ένα μόνο δεδομένο $N = 1$!) έχουμε άπειρες λύσεις για τις παραμέτρους (w_0, w_1)

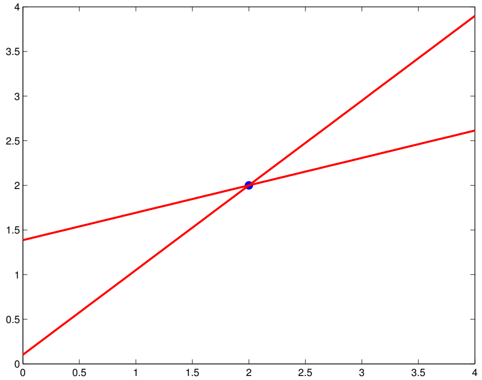
- τι σημαίνει αυτό διαισθητικά;

Παράδειγμα άπειρων λύσεων



- ⇒ σημαίνει ότι υπάρχουν άπειρες ευθείες γραμμές που μπορούν να περάσουν από ένα μόνο σημείο/δεδομένο

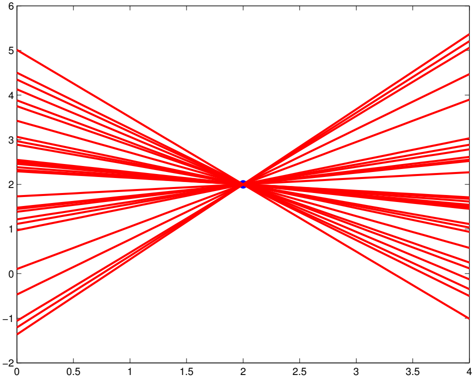
Παράδειγμα άπειρων λύσεων



- ⇒ σημαίνει ότι υπάρχουν άπειρες ευθείες γραμμές που μπορούν να περάσουν από ένα μόνο σημείο/δεδομένο

Εκπαίδευση με μέγιστη πιθανόφανεια

Παράδειγμα άπειρων λύσεων



- $\Rightarrow$ σημαίνει ότι υπάρχουν άπειρες ευθείες γραμμές που μπορούν να περάσουν από ένα μόνο σημείο/δεδομένο

Εκπαίδευση με μέγιστη πιθανόφανεια

$$\mathbf{w} = (X^T X)^{-1} X^T t$$

- Θα θέλαμε να κανονικοποίησουμε την εκπαίδευση ώστε το παραπάνω γραμμικό σύστημα να έχει πάντα μοναδική λύση
- Θα θέλαμε επίσης και μια πιθανοτική ερμηνεία αυτής της κανονικοποίησης
    - μια και το μοντέλο μας είναι πλέον πιθανοτικό...

Κανονικοποίηση με Bayesian Maximum Aposterior μέθοδο

- Το αρχικό μοντέλο είχε από κοινού κατανομή δεδομένων

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

- Ορίζουμε μια κατανομή ως προς τις παραμέτρους $\mathbf{w}$, την οποία υποθέτουμε κανονική

$$p(\mathbf{w}) = \prod_{i=0}^{D} p(w_i) = \prod_{i=0}^{D} \left(\frac{\alpha}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\alpha}{2} w_i^2\right\} = \left(\frac{\alpha}{2\pi}\right)^{\frac{D+1}{2}} \exp\left\{-\frac{\alpha}{2}||\mathbf{w}||^2\right\}$$

Εκπαίδευση με κανονικοποίηση

Κανονικοποίηση με Bayesian Maximum Aposterior μέθοδο

- Πιθανοφάνεια

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

- Εκ των προτέρων κατανομή

$$p(\mathbf{w}) = \prod_{i=0}^{D} p(w_i) = \prod_{i=0}^{D} \left(\frac{\alpha}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\alpha}{2} w_i^2\right\} = \left(\frac{\alpha}{2\pi}\right)^{\frac{D+1}{2}} \exp\left\{-\frac{\alpha}{2}||\mathbf{w}||^2\right\}$$

- Αν εφαρμόζαμε με απόλυτη ευλάβεια την θεωρία πιθανοτήτων, το ζητούμενο της μάθησης θα ήταν η εύρεση της εκ των υστέρων κατανομής με το θεώρημα του Bayes

$$p(\mathbf{w}|\mathbf{t}, X, \beta) = \frac{p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w})}{\int p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w}) d\mathbf{w}} = \frac{p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w})}{p(\mathbf{t}|X, \beta)}$$

Εκπαίδευση με κανονικοποίηση

Κανονικοποίηση με Bayesian Maximum Aposterior μέθοδο

$$p(\mathbf{w}|\mathbf{t}, X, \beta) = \frac{p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w})}{p(\mathbf{t}|X, \beta)}$$

- Ωστόσο ο ακριβής υπολογισμός της $p(\mathbf{w}|\mathbf{t}, X, \beta)$ στις περισσότερες περιπτώσεις είναι αδύνατος (αν και στην συγκεκριμένη περίπτωση μπορεί να γίνει αναλυτικά)
- Οπότε στην πράξη θα θέλαμε να εκφράσουμε μια σύνοψη της εκ των υστέρων κατανομής και συγκεκριμένα να υπολογίσουμε εκείνο το $\mathbf{w}$ που μεγιστοποιεί την

$$p(\mathbf{w}|\mathbf{t}, X, \beta)$$

- ή ισοδύναμα μεγιστοποιεί την ποσότητα

$$p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w})$$

Κανονικοποίηση με Bayesian Maximum Aposterior μέθοδο

- Θα θέλαμε να μεγιστοποιήσουμε την

$$p(\mathbf{t}|X, \mathbf{w}, \beta) \times p(\mathbf{w})$$

- Παίρνωντας λογάριθμο και μεγιστοποιώντας προκύπτει η κανονικοποιημένη συνάρτηση ελαχίστων τετραγώνων (την οποία ελαχιστοποιούμε)

$$E(\mathbf{w}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2}_{\text{data term}} + \underbrace{\lambda\frac{||\mathbf{w}||^2}{2}}_{\text{regularization term}}$$

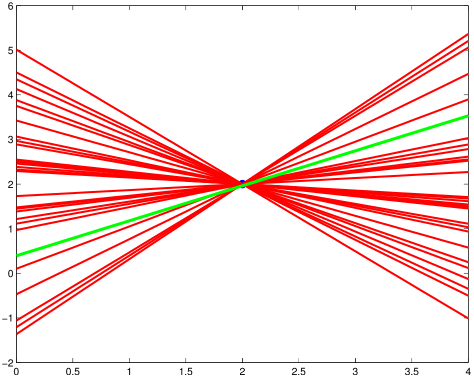
όπου $\lambda = \frac{\alpha}{\beta}$

- Η λύση της ελαχιστοποίησης δίνει

$$\mathbf{w} = (X^T X + \lambda I)^{-1} X^T t$$

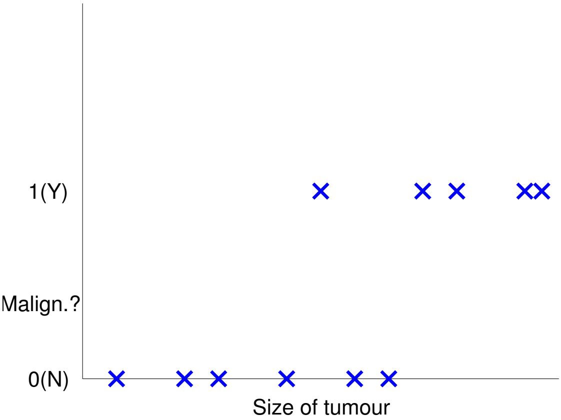
που είναι πάντα μοναδική

Παράδειγμα κανονικοποιημένης λύσης



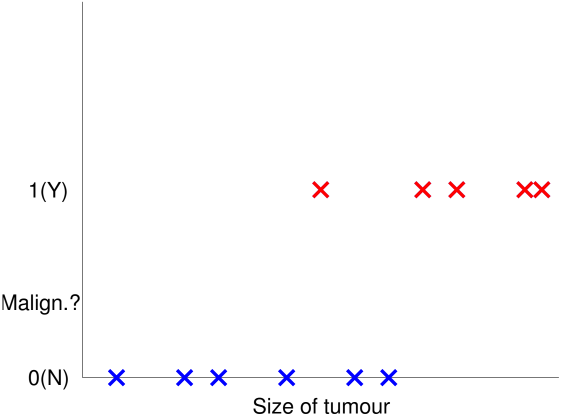
Με την πράσινη γραμμή φαίνεται η μοναδική λύση που παίρνουμε επιλέγοντας $\lambda = 0.1$

Κατηγοριοποίηση



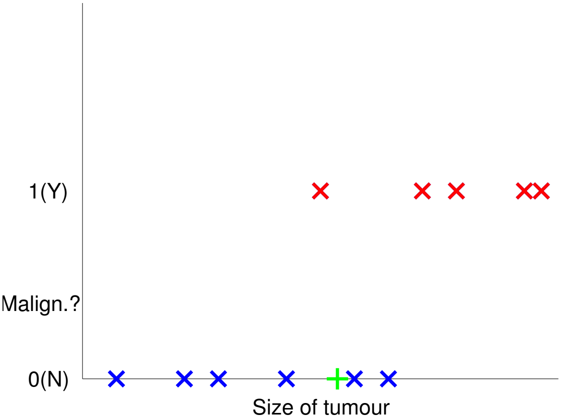
Έστω δεδομένα ασθενών ώστε στον οριζόντιο άξονα (δεδομένα εισόδου) δίνεται το μέγεθος ενός «ενδεχομένως» καρκινικού όγκου και στο κάθετο άξονα έχουμε δύο τιμές 0 ή 1 (δεδομένα εξόδου) για τις δύο περιπτώσεις, δηλ. καλοήθης ή κακοήθης

Κατηγοριοποίηση



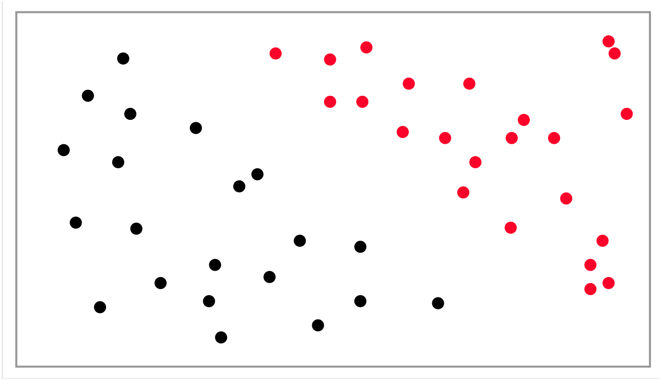
Κάθε δεδομένο ανήκει σε μια κατηγορία, δηλ στην κατηγορία 0 ή την κατηγορία 1 (δηλ. κάθε $t_n \in \{0, 1\}$)

Κατηγοριοποίηση



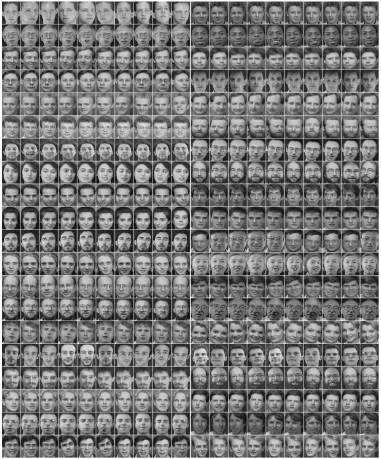
Πρόβλημα κατηγοριοποίησης: Για ένα νέο όγκο (+ στο σχήμα) πως μπορούμε να εκτιμήσουμε την πιθανότητα να είναι καλοήθης ή κακοήθης;

Κατηγοριοποίηση



Τα δεδονένα εισόδου θα μπορούσε να είναι δισδιάστατα

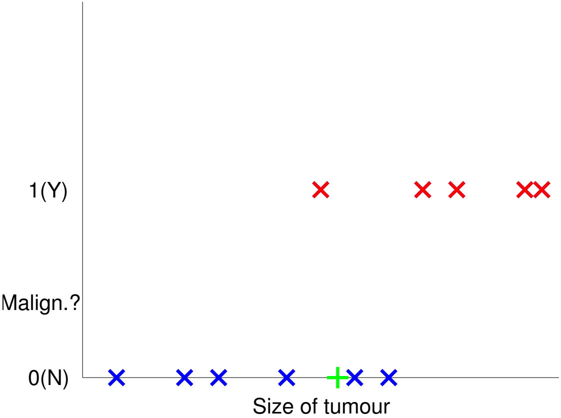
Κατηγοριοποίηση



ή πολυδιάστατα

(όπως στο παράδειγμα που θα θέλαμε να κατηγοριοποιούμε εικόνες ανάλογα με το φύλο)

Κατηγοριοποίηση

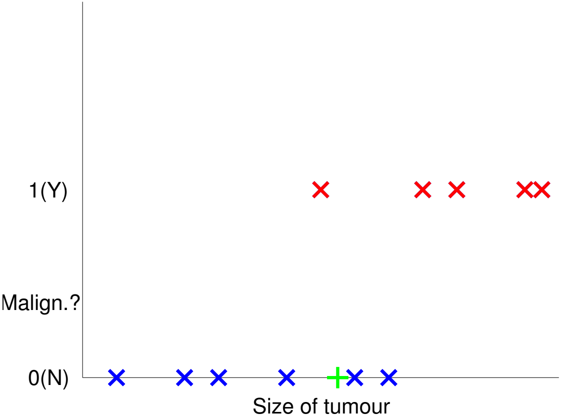


Για ένα νέο όγκο πως μπορούμε να εκτιμήσουμε την πιθανότητα να είναι καλοήθης ή κακοήθης;

Χρειαζόμαστε ένα μοντέλο που να εκτιμά την πιθανότητα να είναι κακοήθης (δηλ. την πιθανότητα του ενδεχομένου $t = 1$ δοθέντος του x)

$$p(t = 1|x) \quad \text{όπου } x = \text{'size of tumour'}$$

Κατηγοριοποίηση



$$p(t = 1|x) \quad \text{όπου } x \; = \; \text{'size of tumour'}$$

Προφανώς η πιθανότητα να έχουμε καλοήθη όγκο είναι

$$p(t = 0|x) = 1 - p(t = 1|x)$$

(οπότε χρειάζεται να μοντελοποιήσουμε μόνο την $p(y = 1|x)$!)

Γραμμική λογιστική παλινδρόμηση

Αφού $p(t=1|x)$ είναι μια πιθανότητα, δηλ. $p(t=1|x) \in [0,1]$, αν εισάγουμε ένα μοντέλο $p(t=1|x,\mathbf{w})$ με παραμέτρους $\mathbf{w}$ θα πρέπει να ικανοποιεί

$$p(t=1|x,\mathbf{w}) \in [0,1]$$

Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας μια συνάρτηση $y(x,\mathbf{w})$ (link function) που παίρνει συνεχή τιμές, π.χ.

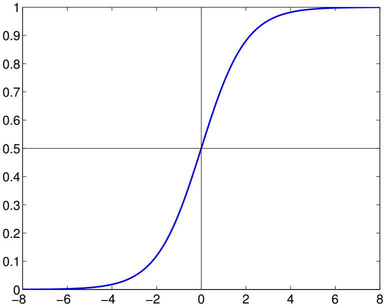
$$y(x,\mathbf{w}) = w_0 + w_1 x, \quad \mathbf{w} = (w_0, w_1)$$

την οποία περνάμε μέσα από τη σιγμοειδή (ονομάζεται και λογιστική συνάρτηση)

$$p(t=1|x,\mathbf{w}) = \frac{1}{1+e^{-y(x,\mathbf{w})}} = \frac{1}{1+e^{-w_0-w_1 x}}$$

(*Το μοντέλο ονομάζεται γραμμική λογιστική παλινδρόμηση λόγω ότι η συνάρτηση που μπαίνει ως είσοδο στην σιγμοειδή είναι γραμμική!*)

Γραμμική λογιστική παλινδρόμηση



Σιγμοειδής συνάρτηση

$$\frac{1}{1+e^{-z}}, \quad z \in \mathbb{R}$$

- Παίρνει τιμές από μηδέν έως 1
- Όταν $z = 0$ ισούται με 0.5
- Όταν $z < 0$ η τιμή της είναι μικρότερη του 0.5
- Όταν $z > 0$ η τιμή της είναι μεγαλύτερη του 0.5

Γραμμική λογιστική παλινδρόμηση

$$p(t=1|x,\mathbf{w}) = \frac{1}{1+e^{-w_0-w_1 x}} = \begin{cases} > 0.5 & w_0 + w_1 x > 0 \\ = 0.5 & w_0 + w_1 x = 0 \\ < 0.5 & w_0 + w_1 x < 0 \end{cases}$$

- Το σύνορο απόφασης (decision boundary) ορίζεται ως το x για το οποίο οι δύο κατηγορίες είναι ισοπίθανες, δηλ.

$$p(y=1|x,\mathbf{w}) = p(t=0|x,,\mathbf{w}) = 0.5 \Rightarrow$$

$$w_0 + w_1 x = 0 \;\; \Rightarrow \;\; x = -\frac{w_0}{w_1}$$

- Όταν το δεδομένο εισόδου είναι διάνυσμα, δηλ. $\mathbf{x} = (x_1, \ldots, x_D)$

$$p(t=1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-w_0-\sum_{i=1}^{D} w_i x_i}}$$

και σύνορο απόφασης είναι όλα τα $\mathbf{x}$ για τα οποία

$$w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = 0$$

(γραμμή για $D=2$, επιφάνεια για $D=3\ldots$ κτλ)

Γραμμική λογιστική παλινδρόμηση

Έστω ότι γνωρίζουμε τις τιμές των παραμέτρων $\mathbf{w} = (w_0, w_i, w_2, \ldots, w_D)$ (το πως τις εκτιμούμε θα το δούμε σε λίγο)

Στην πράξη θα χρησιμοποιούμε το μοντέλο μας ως εξής. Για οποιοδήποτε δεδομένο $\mathbf{x}_*$ θα προβλέπουμε/αποφασίζουμε την άγνωστη κατηγορία $t_* \in \{0, 1\}$ βάσει

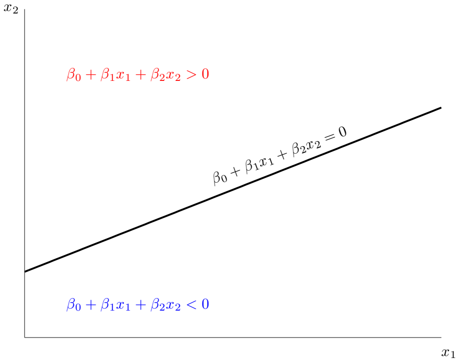
$$\begin{aligned} p(t = 1|x, \mathbf{w}) \geq 0.5 \quad &\Rightarrow \quad t_* = 1 \\ p(t = 1|x, \mathbf{w}) < 0.5 \quad &\Rightarrow \quad t_* = 0 \end{aligned}$$

ή ισοδύναμα

$$\begin{aligned} w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \geq 0 \quad &\Rightarrow \quad t_* = 1 \\ w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D < 0 \quad &\Rightarrow \quad t_* = 0 \end{aligned}$$

(*Οπότε το σύνορο απόφασης* $w_0 + w_i x_i + w_2 x_2 + \ldots + w_D x_D = 0$ *είναι ουσιαστικά εκεί που η απόφαση μας αλλάζει!*)

Γραμμική λογιστική παλινδρόμηση



Η μαύρη γραμμή είναι το σύνορο απόφασης

Κάθε δεδομένο εισόδου $\mathbf{x} = (x_1, x_2)$ που βρίσκεται πάνω από τη διαχωριστική γραμμή θα κατηγοριοποιηθεί στην κατηγορία 1. Αν είναι κάτω από την γραμμή θα κατηγοροποιηθεί στην 0

Γραμμική λογιστική παλινδρόμηση

$$p(t=1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-w_0-\sum_{i=1}^{D} w_i x_i}} = \sigma(w_0 + \sum_{i=1}^{D} w_i x_i)$$

$$p(t=0|\mathbf{x},\mathbf{w}) = 1 - \sigma(w_0 + \sum_{i=1}^{D} w_i x_i)$$

Μπορουμε να γράφουμε το μοντέλο μας πιο συνοπτικά ως εξής. Αν $\tilde{\mathbf{x}} = [1, \mathbf{x}]$

$$w_0 + \sum_{i=1}^{D} w_i x_i = \mathbf{w}^T \tilde{\mathbf{x}}$$

Οπότε

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

$$(\text{δηλ. } p(1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^T\tilde{\mathbf{x}}),\ \ p(0|\mathbf{x},\mathbf{w}) = (1-\sigma(\mathbf{w}^T\tilde{\mathbf{x}})))$$

Εκπαίδευση με μέγιστη πιθανοφάνεια

Θέλουμε να εκτιμήσουμε τα $\mathbf{w} = (w_0, w_1, \ldots, w_D)$

- Δεδομένα $\mathcal{D} = (\mathbf{x}_n, t_n)_{n=1}^N$. Θα γράφουμε $\tilde{\mathbf{x}} = [1, \mathbf{x}] \in \mathbb{R}^{D+1}$ για το δεδομένο προσαυξημένο με ένα επιπλέον στοιχείο το οποίο είναι ίσο με 1. Για ευκολία στην παρουσίαση θα συμβολίζουμε το $\tilde{\mathbf{x}}$ απλά ως $\mathbf{x}$
- Υποθέτουμε ότι κάθε t_n έχει παραχθεί ανεξάρτητα δοθέντος του $\mathbf{x}_n$ έτσι ώστε

$$t_n | \mathbf{x}_n \sim p(t_n | \mathbf{x}_n, \mathbf{w})$$

όπου

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

$$\mathbf{w}^T \mathbf{x}_n = \sum_{i=0}^{D} w_i x_{n,i}$$

- Ποια είναι η από κοινού κατανομή των $\mathbf{t} = (t_n)_{n=1}^N$ δοθέντος των $X = (\mathbf{x}_n)_{n=1}^N$;

Εκπαίδευση με μέγιστη πιθανοφάνεια

- Υποθέτουμε ότι κάθε t_n έχει παραχθεί ανεξάρτητα δοθέντος του $\mathbf{x}_n$ έτσι ώστε

$$t_n|\mathbf{x}_n \sim p(t_n|\mathbf{x}_n, \mathbf{w})$$

  όπου

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

$$\mathbf{w}^T\mathbf{x}_n = \sum_{i=0}^{D} w_i x_{n,i}$$

- Η από κοινού κατανομή είναι

$$\begin{aligned} p(\mathbf{t}|X, \mathbf{w}) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}) \\ &= \prod_{n=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}_n)^{t_n}\left(1 - \sigma(\mathbf{w}^T\mathbf{x}_n)\right)^{1-t_n} \end{aligned}$$

Εκπαίδευση με μέγιστη πιθανοφάνεια

Σημειακή εκτίμηση των παραμέτρων $(w_0, w_1, \ldots, w_D)$

- Η από κοινού κατανομή ή πιθανοφάνεια είναι

$$\begin{aligned} p(\mathbf{t}|X, \mathbf{w}) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}) \\ &= \prod_{i=1}^{n} \sigma(\mathbf{w}^T\mathbf{x}_n)^{t_n} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}_n)\right)^{1-t_n} \end{aligned}$$

- Θέλουμε να εκτιμήσουμε $(w_0, w_1, \ldots, w_D)$ μεγιστοποιώντας την πιθανοφάνεια ή ισοδύναμα τον λογάριθμό της

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} t_n \log \sigma(\mathbf{w}^T\mathbf{x}_n) + (1 - t_n) \log \left(1 - \sigma(\mathbf{w}^T\mathbf{x}_n)\right)$$

Εκπαίδευση με μέγιστη πιθανοφάνεια

Εκτιμούμε τις άγνωστες παραμέτρους μεγιστοποιώντας την λογαριθμική πιθανοφάνεια

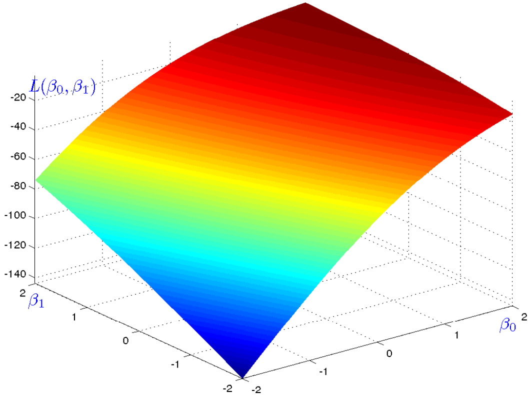
$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} t_n \log \sigma(\mathbf{w}^T \mathbf{x}_n) + (1 - t_n) \log \Big(1 - \sigma(\mathbf{w}^T \mathbf{x}_n)\Big)$$

Η μεγιστοποίηση δεν μπορεί να γίνει αναλυτικά

Θα πρέπει να χρησιμοποιήσουμε αριθμητική βελτιστοποίηση

Ωστόσο η $\mathcal{L}(\mathbf{w})$ είναι κοίλη (στρέφει τα κοιλά προς τα κάτω), οπότε η αριθμητική βελτιστοποίηση μπορεί να βρεί το ολικό μέγιστο

Εκπαίδευση με μέγιστη πιθανοφάνεια



Η $\mathcal{L}(\mathbf{w})$ είναι κοίλη συνάρτηση

Οπότε μπορούμε να βρούμε το ολικό μέγιστο εφαρμόζοντας μια μέθοδο όπως αυτή της ανοδικής κλίσης (gradient ascent)

Εκπαίδευση με μέγιστη πιθανοφάνεια

- Παίρνουμε μερικές παραγώγους

$$\frac{\partial \mathcal{L}(\mathbf{w})}{w_i} = \sum_{n=1}^{N} \Big(t_n - \sigma(\mathbf{w}^T \mathbf{x}_n) \Big) x_{n,i}, \quad i = 0, \ldots, D$$

- Διάνυσμα μερικών παραγώγων

$$\nabla_{\mathcal{L}(\mathbf{w})} = \begin{bmatrix} \frac{\partial \mathcal{L}(\mathbf{w})}{w_0} \\ \frac{\partial \mathcal{L}(\mathbf{w})}{w_1} \\ \ldots \\ \frac{\partial \mathcal{L}(\mathbf{w})}{w_D} \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} \big(t_n - \sigma(\mathbf{w}^T \mathbf{x}_n) \big) x_{n,0} \\ \sum_{n=1}^{N} \big(t_n - \sigma(\mathbf{w}^T \mathbf{x}_n) \big) x_{n,1} \\ \ldots \\ \sum_{n=1}^{N} \big(t_n - \sigma(\mathbf{w}^T \mathbf{x}_n) \big) x_{n,D} \end{bmatrix}$$

Αλγόριθμος ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης

1. Αρχικοποίηση: $k = 1$, $\mathbf{w}^{(k)}$, $\eta > 0$.
2. Ενημέρωση παραμέτρων

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \eta \nabla_{\mathcal{L}(\mathbf{w}^{(k)})}$$

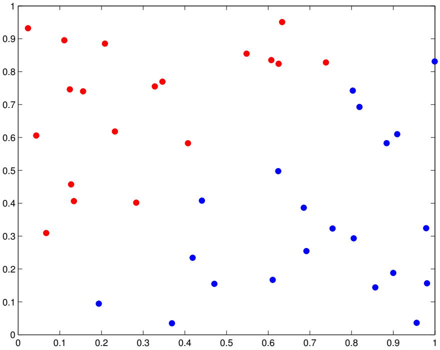
3. $k = k + 1$ και πήγαινε στο βήμα 2, ή τερμάτισε
   Το η ονομάζεται learning rate

Διαισθητικά ο αλγοριθμός υλοποιεί μια συγκεκριμένη στρατηγική για ανέβασμα σε λόφο (hill climbing)

Αλγόριθμος ανοδικής κλίσης

Ας δούμε κάποια παραδείγματα εφαρμογής του αλγορίθμου

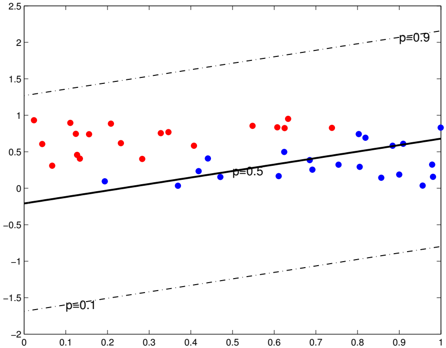
Αλγόριθμος ανοδικής κλίσης



Έχουμε δισδιάστατα δεδομένα εισόδου, δηλ. κάθε
$\mathbf{x}_n = (x_{n,1}, x_{n,2})$

Κόκκινες κουκίδες αντιστοιχούν σε δεδομένα της κατηγορίας 1
ενώ μπλε κουκίδες αντιστοιχούν σε δεδομένα της κατηγορίας 0

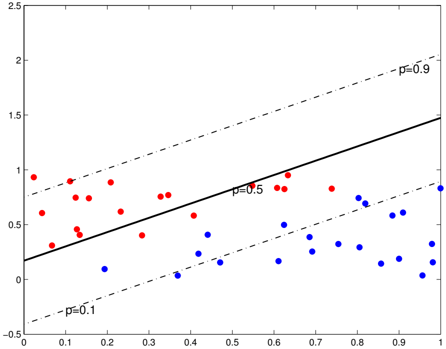
Αλγόριθμος ανοδικής κλίσης



Σχήμα: Η γραμμή $p = 0.5$ αντιστοιχεί στο σύνορο απόφασης $\sigma(w_0 + w_1x_1 + w_2x_2) = 0.5 \;\Rightarrow w_0 + w_1x_1 + w_2x_2 = 0$. Ομοίως οι γραμμές $p = 0.9$ και $p = 0.1$ αντιστοιχούν σε $\sigma(w_0 + w_1x_1 + w_2x_2) = 0.9$ και $\sigma(w_0 + w_1x_1 + w_2x_2) = 0.1$.

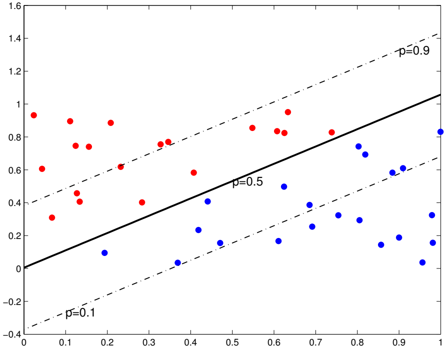
Αρχικοποίηση ($k = 1$) του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



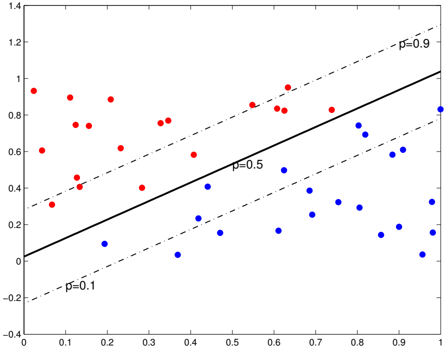
Επανάληψη $k = 10$ του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



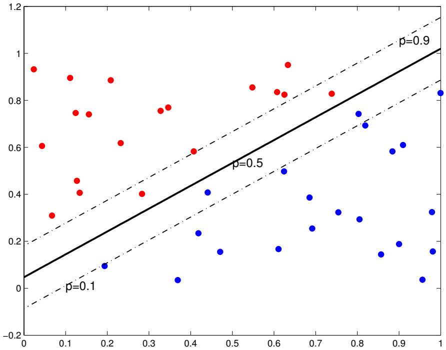
Επανάληψη $k = 20$ του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



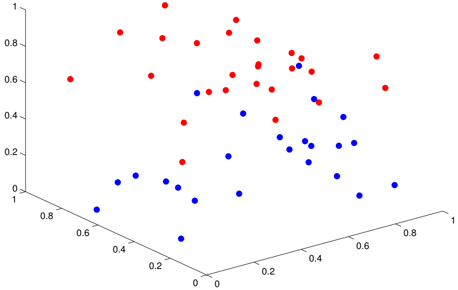
Επανάληψη $k = 50$ του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



Επανάληψη $k = 300$ του αλγορίθμου της ανοδικής κλίσης

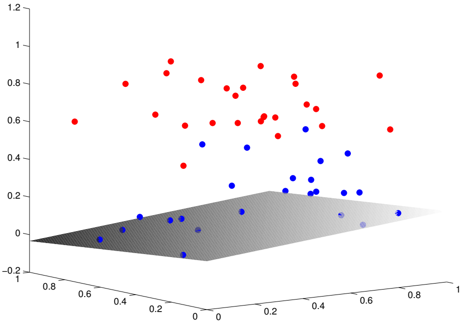
Αλγόριθμος ανοδικής κλίσης



Έχουμε τρισδιάστατα δεδομένα εισόδου, δηλ. κάθε $\mathbf{x}_n = (x_{n,1}, x_{n,2}, x_{n,3})$

Κόκκινες κουκίδες αντιστοιχούν σε δεδομένα της κατηγορίας 1 ενώ μπλε κουκίδες αντιστοιχούν σε δεδομένα της κατηγορίας 0

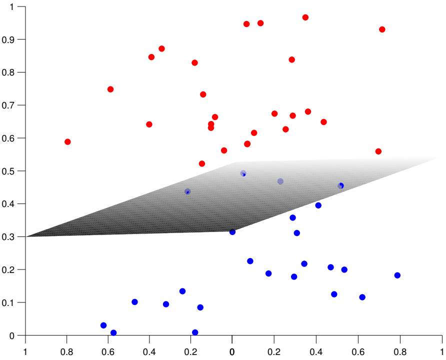
Αλγόριθμος ανοδικής κλίσης



Σχήμα: Η επιφάνεια αντιστοιχεί στο σύνορο απόφασης $\sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_3) = 0.5 \;\Rightarrow w_0 + w_1x_1 + w_2x_2 + w_3x_3 = 0.$

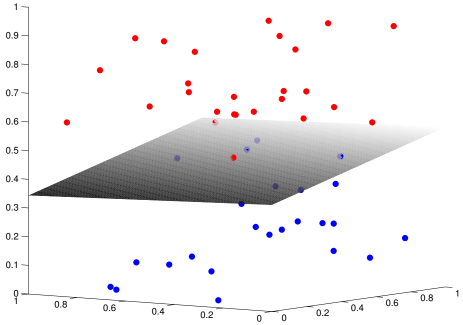
Αρχικοποίηση ($k = 1$) του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



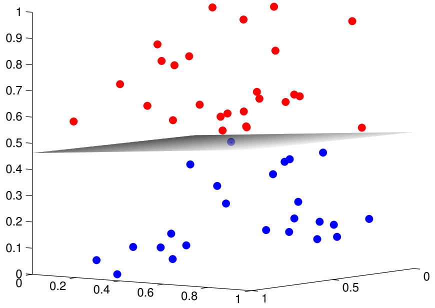
Επανάληψη $k = 20$ του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



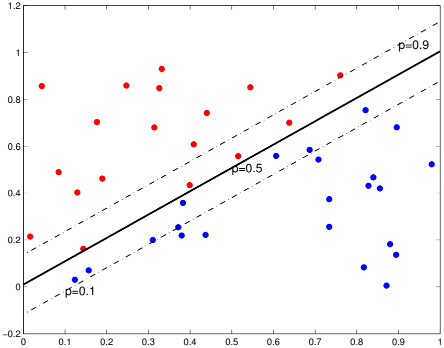
Επανάληψη $k = 50$ του αλγορίθμου της ανοδικής κλίσης

Αλγόριθμος ανοδικής κλίσης



Επανάληψη $k = 300$ του αλγορίθμου της ανοδικής κλίσης

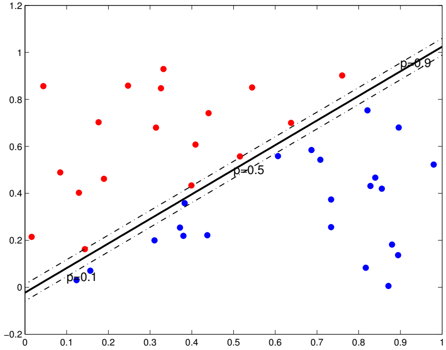
Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση



Έστω ότι τρέχουμε τον αλγόριθμο ανοδικής κλίσης για $k = 300$ επαναλήψεις με ένα συγκεκριμένο learning rate και βρίσκουμε την παραπάνω λύση με παραμέτρους:
$(w_0, w_1, w_2) = (0.1701, 17.2141, -17.3075)$

Έχει συγκλίνει ο αλγόριθμος;

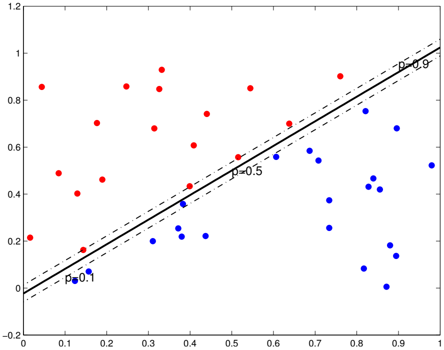
Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση



Αν συνεχίσουμε τις επαναλήψεις και φτάσουμε στις $k = 10000$ επαναλήψεις ο αλγόριθμος βρίσκει παραμέτρους:
$(w_0, w_1, w_2) = (-1.4190, 64.3228, -61.4238)$

Αν εκτελούσαμε και αλλές επαναλήψεις οι παράμετροι σε απόλυτη τιμή θα λάμβαναν ακόμα μεγαλύτερες τιμές

Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση



Οι παραμέτροι (w_0, w_1, w_2) οδηγούνται σε πολύ ακραίες τιμές κοντά στο συν/πλην άπειρο. Αυτό συμβαίνει διότι τα δεδομένα είναι γραμμικά διαχωρίσιμα, δηλ. υπάρχουν (w_0, w_1, w_2) τέτοια ώστε

$$w_0 + w_1 x_{n,1} + w_2 x_{n,2} > 0, \quad \forall n, \; t_n = 1$$

$$w_0 + w_1 x_{n,1} + w_2 x_{n,2} < 0, \quad \forall n, \; t_n = 0$$

Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση

$$w_0 + w_1 x_{n,1} + w_2 x_{n,2} > 0, \quad \forall n, \ t_n = 1$$
$$w_0 + w_1 x_{n,1} + w_2 x_{n,2} < 0, \quad \forall n, \ t_n = 0$$

Αν $a > 0$ (π.χ. a ένας μεγάλος θετικός), ισχύει

$$a w_0 + a w_1 x_{n,1} + a w_2 x_{n,2} > 0, \quad \forall n, \ t_n = 1$$
$$a w_0 + a w_1 x_{n,1} + a w_2 x_{n,2} < 0, \quad \forall n, \ t_n = 0$$

Οπότε για $(w_0^*, w_1^*, w_2^*) = (a w_0, a w_1, a w_2)$ τα δεδομένα συνεχίζουν να είναι γραμμικά διαχωρίσιμα, ωστόσο οι σιγμοειδή συναρτήσεις θα παίρνουν πιο ακραίες τιμές (δηλ. τιμές κοντά στο 0 ή 1) και η συνάρτηση κόστους αυξάνει

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} t_n \log \sigma(\mathbf{w}^T \mathbf{x}_n) + (1 - t_n) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_n)\right) \rightarrow 0$$

οδηγώντας το μοντέλο στην υπερεκπαίδευση $\Rightarrow$ που στην προκειμένη περίπτωση σημαίνει ότι γινόμαστε υπερβολικά βέβαιοι για τις προβλέψεις μας

Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση

Κανονικοποίηση με Bayesian Maximum Aposterior μέθοδο

- Το αρχικό μοντέλο είχε από κοινού κατανομή

$$p(\mathbf{t}|X,\mathbf{w}) = \prod_{n=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}_n)^{t_n}\left(1-\sigma(\mathbf{w}^T\mathbf{x}_n)\right)^{1-t_n}$$

- Ορίζουμε μια κατανομή ως προς τις παραμέτρους $\mathbf{w}$, την οποία υποθέτουμε κανονική

$$p(\mathbf{w}) = \prod_{i=0}^{D} p(w_i) = \prod_{i=0}^{D}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}w_i^2\right\}$$

- Θα θέλαμε να μεγιστοποιήσουμε την

$$p(\mathbf{t}|X,\mathbf{w}) \times p(\mathbf{w})$$

- ή ισοδύναμα τον λογάριθμο της

$$E(\mathbf{w}) = \sum_{n=1}^{N} t_n \log\sigma(\mathbf{w}^T\mathbf{x}_n) + (1-t_n)\log\left(1-\sigma(\mathbf{w}^T\mathbf{x}_n)\right) - \frac{1}{2}\lambda||\mathbf{w}||^2$$

Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση

$$E(\mathbf{w}) = \sum_{n=1}^{N} t_n \log \sigma(\mathbf{w}^T \mathbf{x}_n) + (1 - t_n) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_n)\right) - \frac{1}{2}\lambda ||\mathbf{w}||^2$$

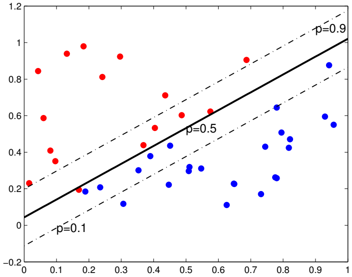
Αυτή η συνάρτηση κόστους παραμένει κοίλη

- διότι αποτελεί το άθροισμα δύο κοίλων συναρτήσεων

Η μεγιστοποίηση έχει μοναδικό μέγιστο το οποίο συμβαίνει πάντα για πεπερασμένη τιμή του $\mathbf{w}$

- δηλ. ακόμα και στην περίπτωση που τα δεδομένα είναι γραμμικά διαχωρίσιμα το $\mathbf{w}$ δεν θα οδηγηθεί στο άπειρο

Υπερεκπαίδευση και κανονικοποίηση στην λογιστική γραμμική παλινδρόμηση



Στο σχήμα απεικονίζεται το μέγιστο της

$$E(\mathbf{w}) = \sum_{n=1}^{N} t_n \log \sigma(\mathbf{w}^T \mathbf{x}_n) + (1 - t_n) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_n)\right) - \frac{1}{2}\lambda ||\mathbf{w}||^2$$

για $\lambda = 0.01$

Λογιστική παλινδρόμηση για πολλές κατηγορίες

Μπορούμε να γενικεύσουμε την μεθοδολογία για περισσότερες από δύο κατηγορίες

- Έστω ότι έχουμε ζεύγη δεδομένων $(\mathbf{x}_n, \mathbf{t}_n)$ όπου τώρα το δεδομένο εξόδου $\mathbf{t}_n$ υποδεικνύει την κατηγορία του $\mathbf{x}_n$ βάσει της 1-of-K κωδικοποίησης από ένα σύνολο $K > 2$ δυνατών κατηγοριών
- 1-of-K κωδικοποίηση σημαίνει ότι κάθε $\mathbf{t}_n$ είναι ένα δυαδικό διάνυσμα διάστασης K με όλα τα στοιχεία του ίσα με μηδέν εκτός από ένα μοναδικό στοιχείο που είναι ίσο με 1, δηλ. ισχύει

$$t_{nk} \in \{0, 1\}, \quad \sum_{k=1}^{K} t_{nk} = 1,$$

- Η θέση εντός του διανύσματος $\mathbf{t}_n$ στην οποία βρίσκεται το μοναδικό 1 υποδεικνύει την κατηγορία του $\mathbf{x}_n$

Μοντέλο

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

όπου

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

μοντελοποιεί την πιθανότητα το δεδομένο $\mathbf{x}_n$ να ανήκει στην κατηγορία k

- παρατήρησε ότι ισχύει $\sum_{k=1}^{K} y_{nk} = 1$

Η συνάρτηση $\frac{e^{\mathbf{w}_k^T \mathbf{x}_n}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T \mathbf{x}_n}}$ ονομάζεται softmax λόγω ότι «επιλέγει» κατά κάποιο τρόπο τη μέγιστη τιμή $\mathbf{w}_k^T \mathbf{x}_n$, όπου $k = 1, \ldots, K$.

Συγκεκριμένα έχει την ιδιότητα ότι

$$\mathbf{w}_k^T \mathbf{x}_n > \mathbf{w}_j^T \mathbf{x}_n, \; j \neq k \Rightarrow y_{nk} > y_{nj}, \; j \neq k$$

Λογιστική παλινδρόμηση για πολλές κατηγορίες

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

Η συνάρτηση softmax μπορεί να γραφεί ως η σιγμοειδής όταν $K = 2$

Λογιστική παλινδρόμηση για πολλές κατηγορίες

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

μοντελοποιεί την πιθανότητα το δεδομένο $\mathbf{x}_n$ να ανήκει στην κατηγορία k. Έχει την ιδιότητα ότι

$$\mathbf{w}_k^T \mathbf{x}_n > \mathbf{w}_j^T \mathbf{x}_n,\ j \neq k \Rightarrow y_{nk} > y_{jk},\ j \neq k$$

Όποτε όλο το σύνολο $\mathbb{R}^D$ χωρίζεται σε K υποσύνολα έτσι ώστε

$$\mathbb{R}_1^D \cup \mathbb{R}_2^D \cup \ldots \mathbb{R}_K^D = \mathbb{R}^D$$

Εντός του $\mathbb{R}_k^D$ αποφασίσουμε την κατηγορία k αφού ισχύει

$$\mathbf{w}_k^T \mathbf{x} > \mathbf{w}_j^T \mathbf{x},\ j \neq k, \quad \mathbf{x} \in \mathbb{R}_k^D$$

Τα σύνορα απόφασης ορίζονται για κάθε ζεύγος κατηγοριών k και j και είναι εκείνα τα $\mathbf{x}$ όπου

$$\mathbf{w}_k^T \mathbf{x} = \mathbf{w}_j^T \mathbf{x}$$

Μοντέλο

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

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 πιθανοφάνεια ή την πιθανοφάνεια μαζί με τον όρο κανονικοποίησης (δηλ. όπως και στην περίπτωση των δύο κατηγοριών)

Λογιστική παλινδρόμηση για πολλές κατηγορίες

Π.χ. Εκπαίδευση μέσω μέγιστης πιθανοφάνειας

- Πιθανοφάνεια

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

όπου $\mathbf{T}$ όλα τα δεδομένα εξόδου (labels) και $\mathbf{W}$ οι παράμετροι

- Λογαριθμική πιθανοφάνεια

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

- Μεγιστοποιούμε την παραπάνω ποσότητα χρησιμοποιώντας κάποια μέθοδο βελτιστοποίησης (όπως αυτή της ανοδικής κλίσης)

Επίλογος

- Διάβασμα για το σπίτι: . Bishop: sections 3.1 μέχρι σελίδα 145, σελίδες 152,153 4.1 και 4.3 ως 4.3.4 (μέχρι σελίδα 210)
- Επόμενα μαθήματα: Μη γραμμικά μοντέλα και νευρωνικά δίκτυα