ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Τεχνητή Νοημοσύνη

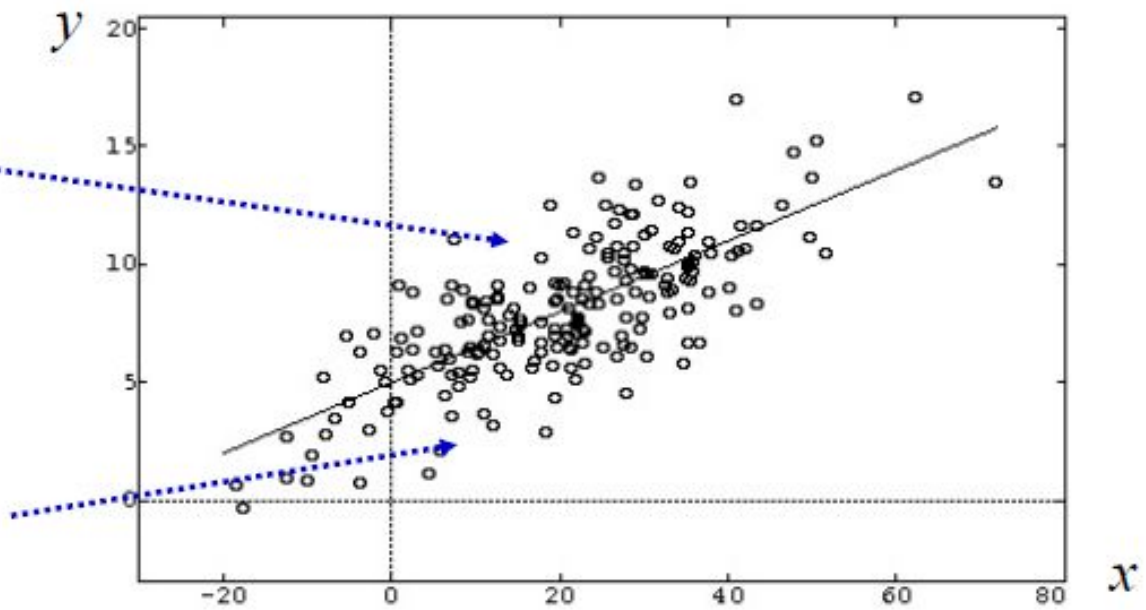
8ο φροντιστήριο (2023-24)

Επιμέλεια: Σοφία Ελευθερίου,
Φοίβος Χαραλαμπάκος

Γραμμική παλινδρόμηση

Τα σημεία πάνω από τη γραμμή της $f(x)$ έχουν: $y > w_1 x + w_0$

Τα σημεία κάτω από τη γραμμή της $f(x)$ έχουν: $y < w_1 x + w_0$



Πηγή εικόνας: http://en.wikipedia.org/wiki/Linear_regression

- Θέλουμε να **μάθουμε την $f(x)$** από ένα **δείγμα** (τελείες).
- Περιοριζόμαστε σε **γραμμικές υποθέσεις** (συναρτήσεις):

$$y = f_{w_1, w_0}(x) = w_1 x + w_0$$

- Άρα ψάχνουμε τα **καλύτερα**: w_1, w_0

Γραμμική παλινδρόμηση – συνέχεια

- Αν έχουμε **δύο ιδιότητες** x_1, x_2, οι γραμμικές μας υποθέσεις αντιστοιχούν σε **επίπεδα** του τριδιάστατου χώρου:

$$y = f_{w_2, w_1, w_0}(x_1, x_2) = w_2 x_2 + w_1 x_1 + w_0$$

- Γενικότερα, αν έχουμε ιδιότητες $x_1, x_2, \ldots, x_n$, οι γραμμικές μας υποθέσεις αντιστοιχούν σε **υπερ-επίπεδα** του $(n+1)$-διάστατου χώρου:

$$y = f_{w_n, \ldots, w_0}(x_1, \ldots, x_n) = w_n x_n + \ldots w_1 x_1 + w_0$$

$$= \sum_{l=0}^{n} w_l x_l = \langle w_0, w_1, \ldots, w_n \rangle \cdot \langle x_0, x_1, \ldots x_n \rangle$$

$$f_{\vec{w}}(\vec{x}) = \vec{w} \cdot \vec{x} = W^T X$$

Θεωρούμε ότι πάντα $x_0 = 1$.

Αν θεωρήσουμε κάθε διάνυσμα ως πίνακα μιας στήλης.

και ψάχνουμε το **καλύτερο** $\vec{w}$.

Συνάρτηση αξιολόγησης

- Ο **χώρος αναζήτησης** περιλαμβάνει τα δυνατά $\vec{w}$.
- Για να **αξιολογήσουμε** κάθε **κατάσταση** $\vec{w}$, θα χρησιμοποιήσουμε τη συνάρτηση:

$$E(\vec{w}) = \frac{1}{2}\sum_{i=1}^{m}[f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}]^2$$

όπου:

$(\vec{x}^{(i)}, y^{(i)})$ τα **παραδείγματα εκπαίδευσης** (δείγμα),

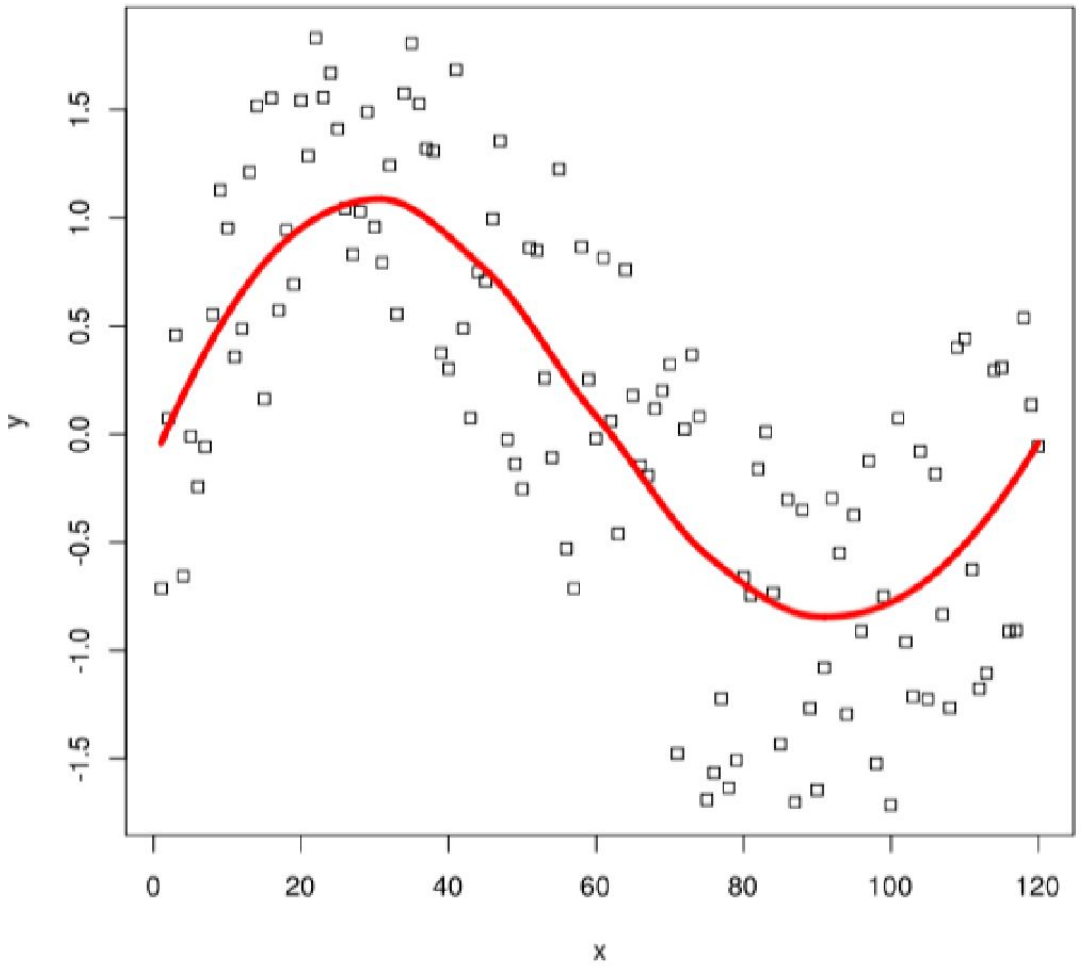
$y^{(i)}$ η **ορθή απόκριση** για είσοδο $\vec{x}^{(i)}$.

- «Γραμμική παλινδρόμηση **ελαχίστων τετραγώνων**».
  - Αξιολογούμε αθροίζοντας τα **τετράγωνα** των **διαφορών** των **αποκρίσεων** από τις **επιθυμητές** τιμές.

Άσκηση 17.1.

Οι τελείες του σχήματος στα δεξιά παριστάνουν έναν δείγμα που προέρχεται από πληθυσμό ο οποίος ακολουθεί στην πραγματικότητα την άγνωστη συνάρτηση y = f(x), της οποίας η γραφική παράσταση είναι η συνεχής καμπύλη[1]. Λόγω θορύβου κατά τις μετρήσεις της δειγματοληψίας, όμως, τα σημεία του δείγματος δεν βρίσκονται ακριβώς πάνω στη συνεχή καμπύλη. Θέλουμε να μάθουμε από το δείγμα μια συνάρτηση y = h(x), που να προσεγγίζει κατά το δυνατόν περισσότερο την f(x).

Α) Εξηγήστε γιατί η χρήση γραμμικής παλινδρόμησης ελαχίστων τετραγώνων δεν θα οδηγούσε σε ικανοποιητική h(x).

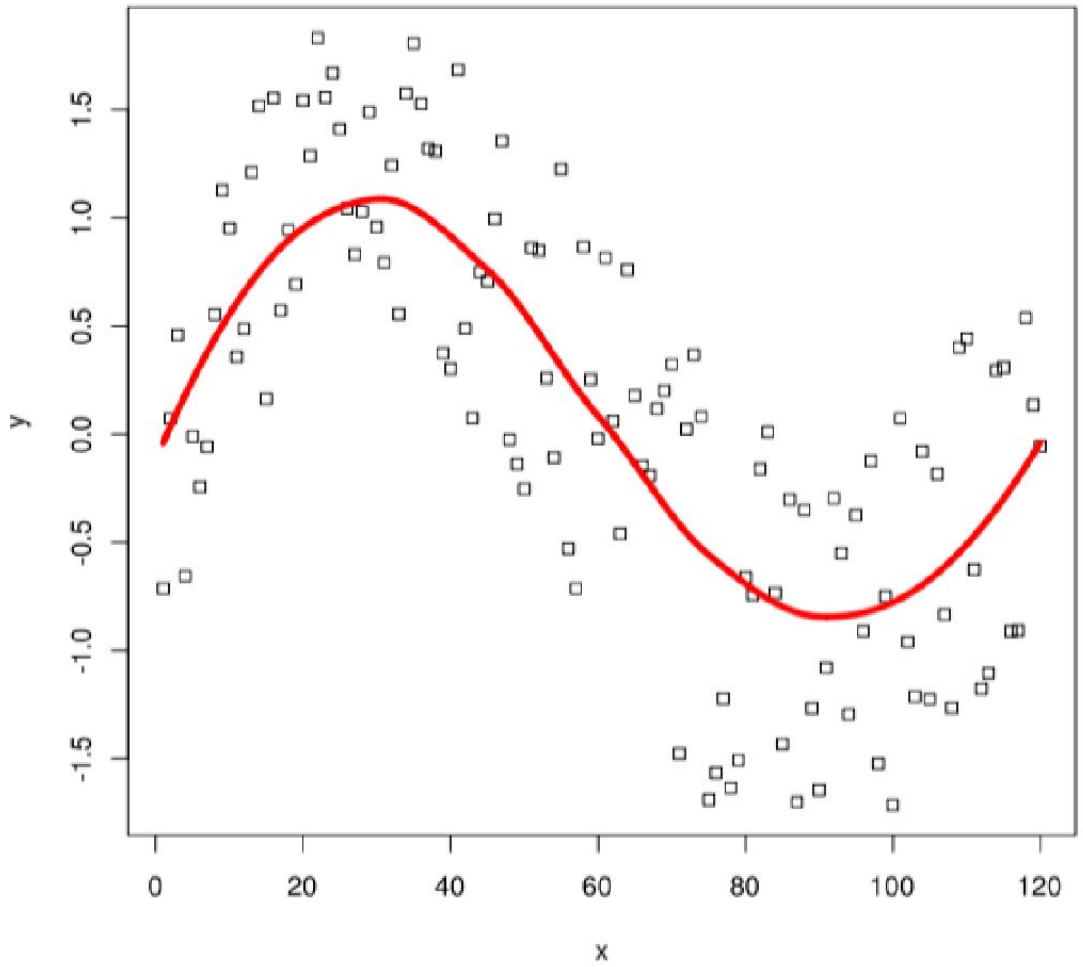


Απάντηση: Η γραμμική παλινδρόμηση ελαχίστων τετραγώνων μαθαίνει ευθείες γραμμές (ή γενικότερα επίπεδα ή υπερ-επίπεδα). Στη συγκεκριμένη περίπτωση η καμπύλη της y= f(x) προφανώς δεν μπορεί να προσεγγιστεί καλά από μια μόνο ευθεία γραμμή.

[1]Το σχήμα προέρχεται από την ιστοσελίδα http://en.wikipedia.org/wiki/Local_regression.

Άσκηση 17.1.

Β) Πώς θα μπορούσαμε να μάθουμε μια πιο ικανοποιητική h(x) χρησιμοποιώντας τον αλγόριθμο των k κοντινότερων γειτόνων (ή μια παραλλαγή του); Τι θα κάναμε κατά το στάδιο της εκπαίδευσης και τι όποτε (κατόπιν) μας δίνουν ένα x για το οποίο πρέπει να επιστρέψουμε το y = h(x);



Απάντηση:

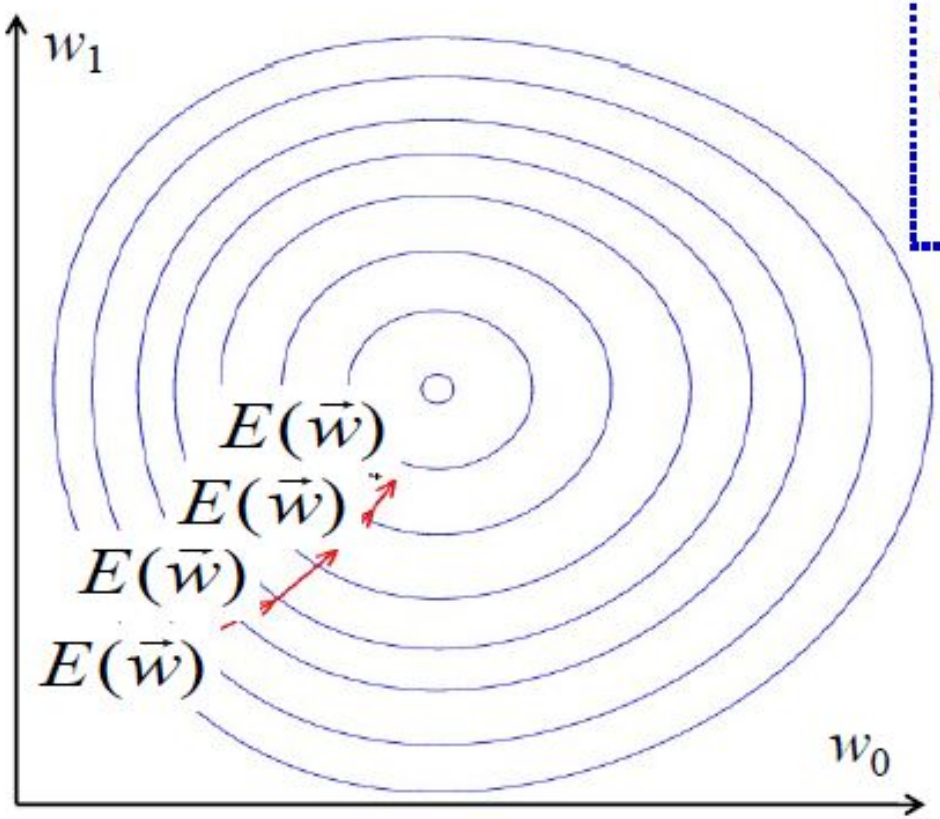
Κατά το στάδιο της εκπαίδευσης απλά αποθηκεύουμε όλες τις συντεταγμένες (x, y) των σημείων του δείγματος.

Κατόπιν, όποτε μας δίνουν ένα νέο x' και μας ζητούν το y' = h(x'), ανακτούμε τα k σημεία του δείγματος (γείτονες) των οποίων οι τιμές x βρίσκονται πιο κοντά στο x' και επιστρέφουμε το μέσο όρο των τιμών y αυτών των k σημείων (των γειτόνων). Μπορούμε επίσης να ζυγίζουμε κατά τον υπολογισμό του μέσου όρου τις y τιμές των γειτόνων, δίνοντας σε κάθε μία y τιμή γείτονα βάρος π.χ. αντιστρόφως ανάλογο της απόστασης της x τιμής του γείτονα από το x'.

Κατάβαση κλίσης (gradient descent)

Ξεκινώ με τυχαία βάρη.
Μετράω σφάλμα $E(\vec{w})$ στα παραδείγματα εκπαίδευσης με τα τρέχοντα βάρη $\vec{w}$. Προς τα πού να μεταβάλω τα βάρη;

$$E(\vec{w}) = \frac{1}{2}\sum_{i=1}^{m}[f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}]^2$$



Η κλίση $\nabla E(\vec{w})$ είναι ένα διάνυσμα που δείχνει προς την κατεύθυνση μεταβολής των βαρών που οδηγεί στη μεγαλύτερη **αύξηση** του $E(\vec{w})$. Το $-\nabla E(\vec{w})$ δείχνει προς την μεγαλύτερη **μείωση**.

Σε κάθε βήμα, τροποποιούμε το $\vec{w}$ κατά η προς την κατεύθυνση που προκαλεί τη μεγαλύτερη μείωση του σφάλματος:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla E(\vec{w})$$

Κατάβαση λόφου με συνάρτηση αξιολόγησης E.

Στοχαστική κατάβαση κλίσης

1. Ξεκίνα με τυχαία βάρη $\vec{w}$.
2. Θέσε $i \leftarrow 1$ και $s \leftarrow 0$. Ανακάτεψε τα παραδείγματα.
3. Υπολόγισε το $E_i(\vec{w}) = \frac{1}{2}[f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}]^2$ **μόνο στο <u>τρέχον</u>** (i-στό) παράδειγμα εκπαίδευσης.
4. $s \leftarrow s + E_i(\vec{w})$
5. Ενημέρωσε τα βάρη: $\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla E_i(\vec{w})$ δηλαδή: $w_l \leftarrow w_l - \eta \cdot [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}] \cdot x_l^{(i)}$
6. Αν υπάρχει $(i+1)$-στό παράδειγμα, θέσε $i \leftarrow i+1$ και πήγαινε στο βήμα 3.
7. Αν το s δεν έχει συγκλίνει και δεν υπερβήκαμε το μέγιστο αριθμό επαναλήψεων, πήγαινε στο βήμα 2.

Προκύπτει υπολογίζοντας τις μερικές παραγώγους...

Άσκηση 17.2.
Γράψτε τους υπολογισμούς με τους οποίους προκύπτει ο κανόνας ενημέρωσης βαρών της γραμμικής παλινδρόμησης ελαχίστων τετραγώνων, όταν χρησιμοποιείται **στοχαστική κατάβαση κλίσης.**

Απάντηση:
Ο κανόνας ενημέρωσης βαρών της γραμμικής παλινδρόμησης ελαχίστων τετραγώνων, όταν χρησιμοποιείται στοχαστική κατάβαση κλίσης, δίνεται από τον τύπο:

$\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla_{\vec{w}} E_i(\vec{w})$, όπου $E_i(\vec{w}) = \frac{1}{2}\left[f_{\vec{w}}\left(\vec{x}^{(i)}\right) - y^{(i)}\right]^2$ και $f_{\vec{w}}\left(\vec{x}^{(i)}\right) = \mathrm{x_n w_n} + \cdots + \mathrm{x_1 w_1} + \mathrm{w_0}$

Επίσης:

$$\nabla_{\vec{w}} E_i(\vec{w}) = < \frac{\partial E_i(\vec{w})}{\partial w_0}, \frac{\partial E_i(\vec{w})}{\partial w_1}, \ldots, \frac{\partial E_i(\vec{w})}{\partial w_l}, \ldots, \frac{\partial E_i(\vec{w})}{\partial w_n} >$$

Για $l \in \{0, \ldots, n\}$:

$$\frac{\partial E_i(\vec{w})}{\partial w_l} = \left[\mathrm{f}_{\vec{w}}\left(\vec{x}^{(i)}\right) - y^{(i)}\right] \cdot x_l^{(i)}$$

Άρα,

$$\nabla_{\vec{w}} E_i(\vec{w}) = \left[\mathrm{f}_{\vec{w}}\left(\vec{x}^{(i)}\right) - y^{(i)}\right] < x_0^{(i)}, x_1^{(i)}, \ldots, x_l^{(i)}, \ldots, x_n^{(i)} > = \left[\mathrm{f}_{\vec{w}}\left(\vec{x}^{(i)}\right) - y^{(i)}\right] \cdot \vec{x}^{(i)}$$

και ο τύπος ενημέρωσης βαρών γίνεται:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \left[\mathrm{f}_{\vec{w}}\left(\vec{x}^{(i)}\right) - y^{(i)}\right] \cdot \vec{x}^{(i)}$$

Άσκηση 17.3.

Καταγράψαμε μεγάλο αριθμό παρτίδων (συνολικά M παρτίδες) μεταξύ παικτών Othello. Σε κάθε παρτίδα καταγράψαμε όλα τα στιγμιότυπα της σκακιέρας (τις θέσεις όλων των πιονιών), ένα στιγμιότυπο αμέσως μετά από κάθε κίνηση παίκτη. Τα στιγμιότυπα όλων των παρτίδων συνολικά ήταν K, ενώ τα διαφορετικά στιγμιότυπα (μετρώντας τα ίδια στιγμιότυπα μόνο μία φορά το καθένα) όλων των παρτίδων συνολικά ήταν L. Για κάθε στιγμιότυπο, καταγράψαμε τις τιμές των ιδιοτήτων $f_1(n)$, $f_2(n)$, $f_3(n)$, όπου τώρα **n** είναι το στιγμιότυπο· εμφανίστηκαν συνολικά N διαφορετικοί συνδυασμοί τιμών των $f_1(n)$, $f_2(n)$, $f_3(n)$ στις M παρτίδες. Για κάθε στιγμιότυπο, καταγράψαμε ακόμη αν η παρτίδα στην οποία εμφανίστηκε το στιγμιότυπο έληξε υπέρ του μαύρου παίκτη, υπέρ του άσπρου ή ισόπαλη.

Εξηγήστε πώς θα μπορούσαμε να εκμεταλλευτούμε τα καταγεγραμμένα στοιχεία των M παρτίδων, για να μάθουμε με γραμμική παλινδρόμηση ελαχίστων τετραγώνων τις καλύτερες δυνατές τιμές των w_0, w_1, w_2, w_3, ώστε για κάθε κόμβο n του δέντρου αναζήτησης, με MiniMax, η $h(n) = w_1 \times f_1(n) + w_2 \times f_2(n) + w_3 \times f_3(n) + w_0$ επιστρέφει μια κατά το δυνατόν ακριβέστερη εκτίμηση του *αναμενόμενου* οφέλους με το οποίο θα τελειώσει το παιχνίδι, αν βρεθεί στην κατάσταση του κόμβου n.

Θεωρήστε ότι το όφελος ενός παιχνιδιού είναι 1 όταν κερδίζει ο Max (μαύρος), –1 όταν κερδίζει ο Min (άσπρος) και 0 όταν το παιχνίδι λήγει ισόπαλο.

Άσκηση 17.3.

Α) Πόσα παραδείγματα (διανύσματα ιδιοτήτων) εκπαίδευσης θα είχαμε στη γραμμική παλινδρόμηση και ποιες θα ήταν οι μεταβλητές της συνάρτησης που θα μάθαινε η γραμμική παλινδρόμηση; Φροντίστε να μην υπάρχουν ασυνεπή παραδείγματα εκπαίδευσης.

Απάντηση:

Θα είχαμε N παραδείγματα εκπαίδευσης, ένα για κάθε διαφορετικό συνδυασμό τιμών των $f_1(n)$, $f_2(n)$, $f_3(n)$ που παρατηρήθηκε στις M παρτίδες. Οι μεταβλητές θα ήταν οι ιδιότητες $f_1(n)$, $f_2(n)$, $f_3(n)$.

Άσκηση 17.3.

Β) Πώς ακριβώς (δώστε μαθηματικό τύπο) θα υπολογίζαμε για κάθε παράδειγμα εκπαίδευσης την «ορθή» (επιθυμητή) απόκριση της συνάρτησης που θα θέλαμε να μάθει η γραμμική παλινδρόμηση;

Απάντηση:

Κάθε παράδειγμα εκπαίδευσης παριστάνει έναν από τους Ν διαφορετικούς συνδυασμούς τιμών των τριών ιδιοτήτων που παρατηρήθηκε στις Μ παρτίδες. Έστω σ ένας από αυτούς τους διαφορετικούς συνδυασμούς και σ_1 , σ_2 , ..., σ_k *τα καταγεγραμμένα στιγμιότυπα (όχι αναγκαστικά διαφορετικά μεταξύ τους) που είχαν το συγκεκριμένο συνδυασμό τιμών του σ στις* $f_1(n)$, $f_2(n)$, $f_3(n)$. *Η επιθυμητή απόκριση για το σ θα ήταν:*

$$\frac{1}{k}\sum_{i=1}^{k} u(\sigma_i) = y^{(\sigma)}$$

όπου $u(\sigma_i)$ *είναι το όφελος με το οποίο τελείωσε η παρτίδα στην οποία καταγράφηκε το στιγμιότυπο .*

Άσκηση 17.3.

Γ) *Ποια θα ήταν η σχέση της συνάρτησης $f_w(x) = w \times x$ που θα μάθαινε η γραμμική παλινδρόμηση με την h(n) που θα χρησιμοποιούσαμε τελικά;*

Απάντηση:

$$h(n) = f_{\vec{w}}(\vec{x}) \text{ , } \mu\varepsilon : \vec{x} = < f_1(n), f_2(n), f_3(n) >$$

$$f_{\vec{w}}(\vec{x}) = f_{w_0,w_1,w_2,w_3}(\langle f_1(n), f_2(n), f_3(n) \rangle) = w_1 \cdot f_1(n) + w_2 \cdot f_2(n) + w_3 \cdot f_3(n) + w_0 = h(n)$$

Ταξινομητές λογιστικής παλινδρόμησης
(logistic regression classifiers)

$$P(c_+ \mid \vec{x}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}},\ P(c_- \mid \vec{x}) = 1 - P(c_+ \mid \vec{x}) = \frac{e^{-\vec{w}\cdot\vec{x}}}{1+e^{-\vec{w}\cdot\vec{x}}}$$

- Κατά την **εκπαίδευση**, επιλέγουν το $\vec{w}$ που κάνει τον ταξινομητή πιο βέβαιο ότι τα **παραδείγματα εκπαίδευσης** ανήκουν στις **σωστές κατηγορίες**.
  - Μεγιστοποιούν τη (δεσμευμένη) «**πιθανοφάνεια**» των παραδειγμάτων.

$$L(\vec{w}) = P(y^{(1)}, \ldots, y^{(m)} \mid \vec{x}^{(1)}, \ldots, \vec{x}^{(m)}; \vec{w})$$

Οι σωστές κατηγορίες των παραδειγμάτων εκπαίδευσης.

Τα παραδείγματα εκπαίδευσης.

Μεγιστοποίηση πιθανοφάνειας

- Θεωρώντας ότι τα παραδείγματα εκπαίδευσης έχουν επιλεγεί από τον **ίδιο πληθυσμό** και είναι **ανεξάρτητα**:

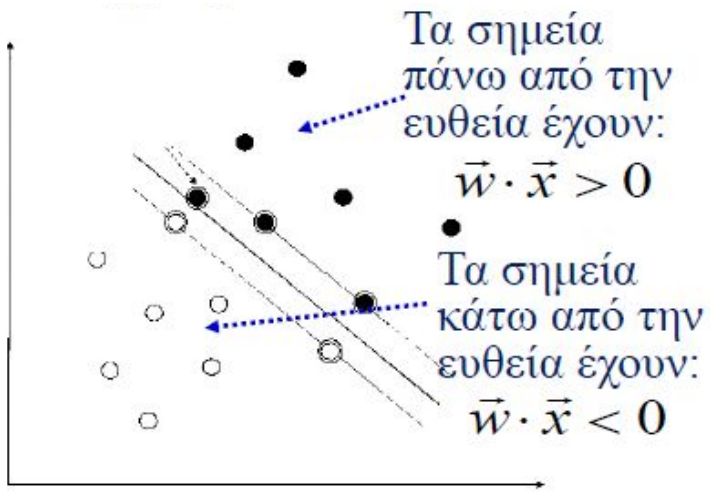
$$L(\vec{w}) = P(y^{(1)}, \ldots, y^{(m)} \mid \vec{x}^{(1)}, \ldots, \vec{x}^{(m)}; \vec{w})$$

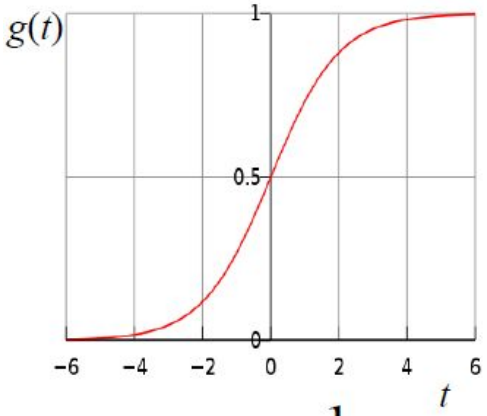
$$= \prod_{i=1}^{m} P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w})$$

- Αντί να μεγιστοποιήσουμε την $L(\vec{w})$, βολεύει να μεγιστοποιήσουμε τη (γνησίως αύξουσα):

$$l(\vec{w}) = \log L(\vec{w}) = \sum_{i=1}^{m} \log P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w})$$

  - Είναι **κοίλη συνάρτηση**, δεν υπάρχει κίνδυνος να φτάσω σε τοπικό μέγιστο.





$$g(t)=\frac{1}{1+e^{-t}}$$

Πιθανότητα το $\vec{x}$ να ανήκει στη **θετική** κατηγορία:

$$P(c_+\mid\vec{x})=\frac{1}{1+e^{-\vec{w}\cdot\vec{x}}}$$

Πιθανότητα να ανήκει στην **αρνητική** κατηγορία:

$$P(c_-\mid\vec{x})=1-P(c_+\mid\vec{x})$$

Μεγιστοποιούμε την πιθανοφάνεια των παραδειγμάτων

$$L(\vec{w})=P\left(y^{(1)},\ldots,y^{(m)}\mid\vec{x}^{(1)},\ldots,\vec{x}^{(m)};\vec{w}\right)$$

Θεωρούμε ότι τα παραδείγματα εκπαίδευσης έχουν επιλεγεί από τον ίδιο πληθυσμό και ότι είναι **ανεξάρτητα** οπότε

$$L(\vec{w})=\prod_{i=1}^{m}P\left(y^{(i)}\mid\vec{x}^{(i)};\vec{w}\right)$$

Αν παραστήσουμε τις κατηγορίες με y = 1 (θετική κατηγορία) και y = 0 (αρνητική κατηγορία), τότε:

$$P\left(y^{(i)}\mid\vec{x}^{(i)};\vec{w}\right)=P\left(c_+\mid\vec{x}^{(i)};\vec{w}\right)^{y^{(i)}}\cdot P\left(c_-\mid\vec{x}^{(i)};\vec{w}\right)^{1-y^{(i)}}$$

Τελικά, αντί να μεγιστοποιήσουμε την L(w), μεγιστοποιούμε την

$$l(\vec{w})=\log L(\vec{w})$$

Άσκηση 18.1.
Γράψτε τους υπολογισμούς με τους οποίους προκύπτει ο κανόνας ενημέρωσης βαρών του ταξινομητή λογιστικής παλινδρόμησης, όταν χρησιμοποιείται (batch) ανάβαση κλίσης.

Απάντηση:
Ο κανόνας ενημέρωσης βαρών είναι $\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla l(\vec{w})$

$$l(\vec{w}) = \log L(\vec{w}) = \log \prod_{i=1}^{m} P\left(c_+ \mid \vec{x}^{(i)}; \vec{w}\right)^{y^{(i)}} \cdot P\left(c_- \mid \vec{x}^{(i)}; \vec{w}\right)^{1-y^{(i)}} =$$

$$\sum_{i=1}^{m} y^{(i)} \log P\left(c_+ \mid \vec{x}^{(i)}; \vec{w}\right) + \left(1 - y^{(i)}\right) log\, P\left(c_- \mid \vec{x}^{(i)}; \vec{w}\right)$$

με

$$P\left(c_+ \mid \vec{x}^{(i)}; \vec{w}\right) = 1/\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right)$$

$$P\left(c_- \mid \vec{x}^{(i)}; \vec{w}\right) = e^{-\vec{w}\vec{x}^{(i)}}/\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right)$$

Με αντικατάσταση έχουμε

$$l(\vec{w}) = \sum_{i=1}^{m} -y^{(i)} \log\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right) + \left(1 - y^{(i)}\right)\left(\log e^{-\vec{w}\vec{x}^{(i)}} - \log\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right)\right) =$$

$$= \sum_{i=1}^{m} log\, e^{-\vec{w}\vec{x}^{(i)}} - log\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right) - y^{(i)}\, log\, e^{-\vec{w}\vec{x}^{(i)}}$$

Άσκηση 18.1.

$$= \sum_{i=1}^{m} log\, e^{-\vec{w}\vec{x}^{(i)}} - log\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right) - y^{(i)}\, log\, e^{-\vec{w}\vec{x}^{(i)}} =$$

$$= \sum_{i=1}^{m} -\vec{w}\vec{x}^{(i)} - log\left(1 + e^{-\vec{w}\vec{x}^{(i)}}\right) + y^{(i)}\, \vec{w}\vec{x}^{(i)}$$

Συνεπώς για $l \in \{0,1, \ldots, n\}$

$$\frac{\partial l(\vec{w})}{\partial w_l} = \sum_{i=1}^{m}\left(-1 + y^{(i)} + \frac{e^{-\vec{w}\vec{x}^{(i)}}}{1 + e^{-\vec{w}\vec{x}^{(i)}}}\right) x_l^{(i)} = \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1 + e^{-\vec{w}\vec{x}^{(i)}}}\right) x_l^{(i)}$$

δηλαδή

$$\frac{\partial l(\vec{w})}{\partial w_l} = \sum_{i=1}^{m}\left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right)\right) x_l^{(i)}$$

Συνεπώς, ο κανόνας ενημέρωσης βαρών για κάθε θα είναι:

$$\vec{w}_l \leftarrow \vec{w}_l + \eta \cdot \sum_{i=1}^{m}\left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right)\right) x_l^{(i)}$$

Κανονικοποίηση (regularization)

- Στην πράξη αντί για το:

$$l(\vec{w}) = \sum_{i=1}^{m} \log P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w})$$

συνήθως μεγιστοποιούμε το:

$$l(\vec{w}) - \lambda \cdot \|\vec{w}\|^2 = l(\vec{w}) - \lambda \cdot \sum_{l=0}^{n} w_l^2$$

δηλ. επιβραβεύουμε υποψήφια $\vec{w}$ με πολλά μικρά βάρη.

- Υπάρχει έτσι **μικρότερος κίνδυνος υπερ-εφαρμογής**.
  - Π.χ. αν πολλά βάρη w_l είναι πολύ μικρά, οι αντίστοιχες ιδιότητες ουσιαστικά δεν χρησιμοποιούνται. Με **λιγότερες ιδιότητες** έχουμε **μικρότερο κίνδυνο υπερ-εφαρμογής**.
  - $\lambda > 0$. Η τιμή επιλέγεται με δοκιμές σε δεδομένα επικύρωσης.

Άσκηση 18.2.

Ποια ακριβώς μορφή παίρνει ο κανόνας ενημέρωσης βαρών της άσκησης 18.1 αν στη λογαριθμική δεσμευμένη πιθανοφάνεια l(w) προστεθεί (για να μειωθεί ο κίνδυνος υπερ-εφαρμογής) ο όρος:

$-\lambda \cdot \|\vec{w}\|^2 = -\lambda \cdot \sum_{l=0}^{n} w_l^2$

Απάντηση:

Αφού μεγιστοποιούμε πλέον την ποσότητα $l(\vec{w}) - \lambda \cdot \sum_{l=0}^{n} w_l^2$ *ο κανόνας ενημέρωσης βαρών γίνεται:*

Σύμφωνα με την άσκηση 18.1:

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla \left(l(\vec{w}) - \lambda \cdot \sum_{l=0}^{n} w_l^2 \right)$$

Επομένως:

$$\frac{\partial l(\vec{w})}{\partial w_l} = \sum_{i=1}^{m} \left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right) \right) x_l^{(i)}$$

$$\frac{\partial \left(l(\vec{w}) - \lambda \cdot \sum_{l=0}^{n} w_l^2 \right)}{\partial w_l} = \sum_{i=1}^{m} \left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right) \right) x_l^{(i)} - 2\lambda w_l$$

Άρα ο κανόνας ενημέρωσης βαρών γίνεται:

$$\vec{w}_l \leftarrow \vec{w}_l + \eta \cdot \left(\sum_{i=1}^{m} \left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right) \right) x_l^{(i)} - 2\lambda w_l \right)$$

$$\vec{w}_l \leftarrow (1 - 2\lambda\eta)\vec{w}_l + \eta \cdot \sum_{i=1}^{m} \left(y^{(i)} - P\left(c_+ \,|\vec{x}^{(i)}; \vec{w}\right) \right) x_l^{(i)}$$

Άσκηση 18.3.

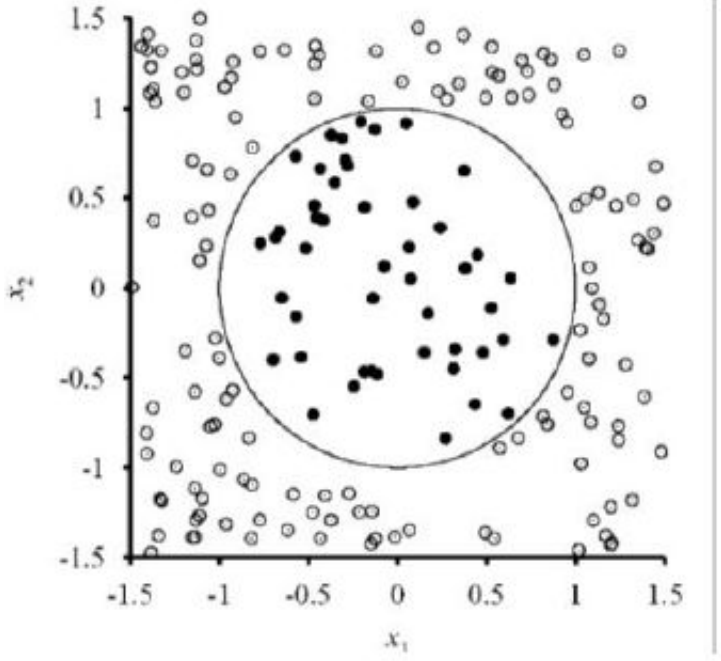
Εκπαιδεύουμε έναν ταξινομητή λογιστικής παλινδρόμησης στα παραδείγματα εκπαίδευσης (τελείες) του σχήματος στα δεξιά. Υπάρχουν δύο κατηγορίες (μαύρη και άσπρη) και δύο ιδιότητες (αντιστοιχούν στους άξονες). Κατόπιν αξιολογούμε τον ταξινομητή στα ίδια παραδείγματα που χρησιμοποιήσαμε για την εκπαίδευσή του. Θα κατατάξει όλα τα παραδείγματα εκπαίδευσης στις σωστές κατηγορίες; Αν ναι, γιατί; Αν όχι, γιατί και τι θα μπορούσαμε να κάνουμε, για να βοηθήσουμε τον ταξινομητή λογιστικής παλινδρόμησης να τα κατατάξει σωστά;

Απάντηση:

Οι ταξινομητές λογιστικής παλινδρόμησης είναι γραμμικοί διαχωριστές. Δηλαδή κατά την εκπαίδευση μαθαίνουν μια ευθεία γραμμή, επίπεδο ή γενικότερα υπερ-επίπεδο και κατά την αξιολόγηση κατατάσσουν τις περιπτώσεις (διανύσματα ιδιοτήτων) που τους δίνουμε σε δύο κατηγορίες, ανάλογα με το αν το διάνυσμα κάθε περίπτωσης βρίσκεται πάνω ή κάτω από το υπερ-επίπεδο.

Τα παραδείγματα του σχήματος της εκφώνησης δεν είναι γραμμικά διαχωρίσιμα στο διανυσματικό χώρο του σχήματος (δεν υπάρχει ευθεία γραμμή που να διαχωρίζει τις μαύρες από τις άσπρες περιπτώσεις). Επομένως, αποκλείεται ένας ταξινομητής λογιστικής παλινδρόμησης να καταφέρει να μάθει μια ευθεία που να διαχωρίζει πλήρως τα παραδείγματα εκπαίδευσης των δύο κατηγοριών, αν τα παραδείγματα παριστάνονται όπως στο σχήμα της εκφώνησης.

Άρα αποκλείεται να καταφέρει να κατατάξει όλα τα παραδείγματα σωστά, ακόμα κι αν πρόκειται για τα ίδια παραδείγματα στα οποία εκπαιδεύτηκε.



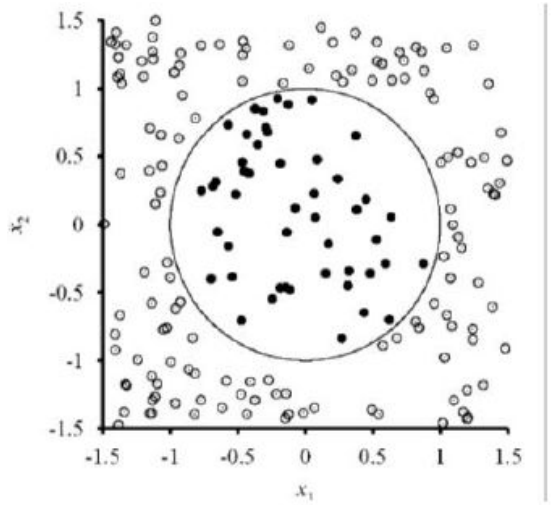
Σχήμα από το βιβλίο των Russel και Norvig.

Άσκηση 18.3.

Απάντηση - συνέχεια:
Αν χρησιμοποιήσουμε, όμως, περισσότερες ιδιότητες, ενδέχεται τα παραδείγματα να γίνουν γραμμικά διαχωρίσιμα. Για παράδειγμα, με το μετασχηματισμό:

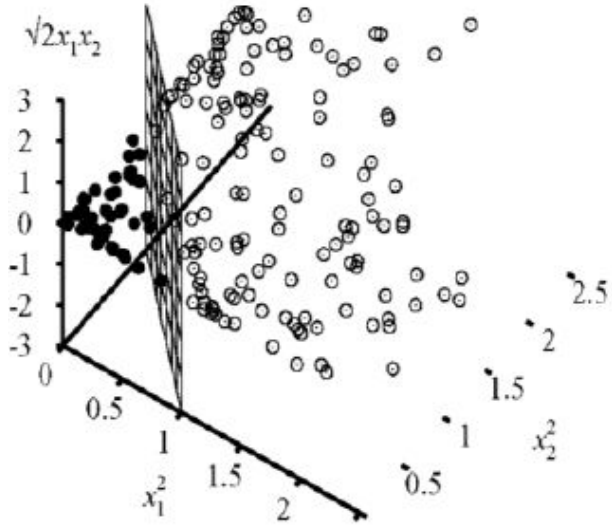
$$\vec{F}(\vec{x}) = < x_1^2, x_2^2, \sqrt{2}x_1x_2 >$$

τα διανύσματα (τελείες) του σχήματος της εκφώνησης διατάσσονται σε ένα νέο τρισδιάστατο διανυσματικό χώρο (έχουμε τώρα τρεις ιδιότητες, που αντιστοιχούν στους τρεις άξονες του νέου χώρου), όπως φαίνεται στο σχήμα στα δεξιά, και είναι πλέον γραμμικά διαχωρίσιμα. Επομένως, ένας ταξινομητής λογιστικής παλινδρόμησης θα μπορούσε να μάθει να κατατάσσει σωστά όλα τα παραδείγματα εκπαίδευσης του αρχικού σχήματος, αρκεί να τα μετασχηματίσουμε πρώτα σύμφωνα με τον παραπάνω μετασχηματισμό



Σχήμα από το βιβλίο των Russel και Norvig.

Αν θέσουμε $X = x_1^2$ και $Y = x_2^2$ και τότε το επίπεδο έχει εξίσωση $X + Y = 1$. Οπότε αν ένα σημείο είναι μέσα στον κύκλο θα είναι $x_1^2 + x_2^2 < 1$ που στο νέο σύστημα θα σημαίνει $X + Y < 1$ και άρα θα είναι πάντα στον ίδιο ημιχώρο.



Σχήμα από το βιβλίο των Russel και Norvig.

Άσκηση 18.4.
Ένας φοιτητής εκπαιδεύει έναν ταξινομητή λογιστικής παλινδρόμησης με (batch) ανάβαση κλίσης. Προκειμένου η εκπαίδευση να ολοκληρώνεται πιο γρήγορα, αύξησε πολύ την τιμή της σταθεράς **η** του κανόνα ενημέρωσης βαρών, ελπίζοντας ότι έτσι θα εκτελούνταν λιγότερα βήματα κατά την ανάβαση κλίσης. Παρατήρησε όμως ότι ο αλγόριθμος δεν τερμάτιζε πλέον· αντίθετα το διάνυσμα βαρών κατέληγε να ταλαντεύεται γύρω από μια τιμή. Γιατί συνέβη αυτό;

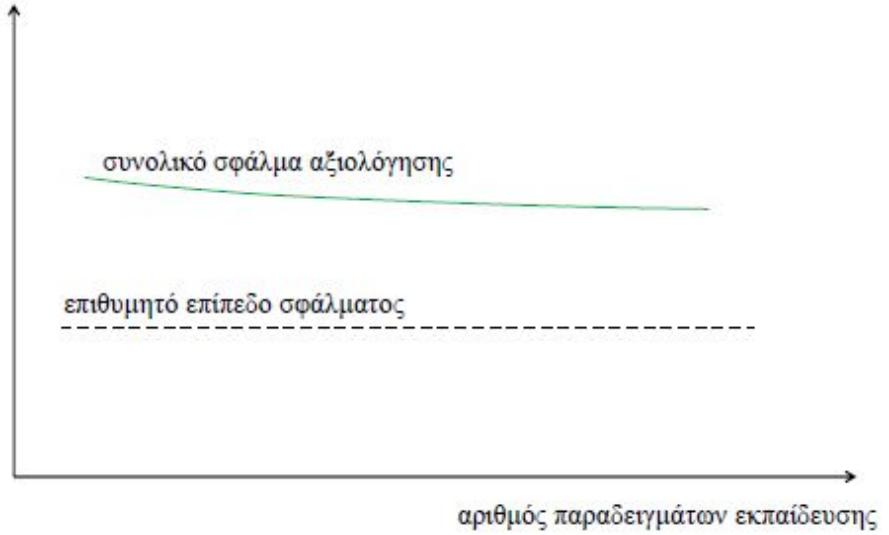
Απάντηση:
Η ανάβαση κλίσης αναζητεί στο χώρο των βαρών των ιδιοτήτων, το διάνυσμα βαρών (σημείο του χώρου) που μεγιστοποιεί τη δεσμευμένη πιθανοφάνεια των παραδειγμάτων εκπαίδευσης (βλ. περιγραφή των ταξινομητών λογιστικής παλινδρόμησης στις διαφάνειες της 18ης διάλεξης).

Σε κάθε επανάληψη της ανάβασης κλίσης, κάνει ένα βήμα στο χώρο των βαρών προς την κατεύθυνση που οδηγεί στην πιο απότομη αύξηση της δεσμευμένης πιθανοφάνειας.

Με πολύ μεγάλο **η**, το βήμα είναι πολύ μεγάλο και ενδέχεται καθώς πλησιάζει το βέλτιστο σημείο (την κορυφή) να περνάει από πάνω του (να πηγαίνει από την άλλη πλευρά της κορυφής), οπότε αναγκάζεται στο επόμενο βήμα να επιστρέψει προς το βέλτιστο σημείο, αλλά περνάει πάλι από πάνω του κ.ο.κ.

Άσκηση 18.5.

Ένας συνάδελφός σας αναπτύσσει ένα σύστημα που χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση. Το συνολικό σφάλμα στα δεδομένα αξιολόγησης, όμως, παραμένει αρκετά υψηλότερο από το επιθυμητό επίπεδο. Για να μειώσει το σφάλμα αξιολόγησης, σκέφτεται να προσθέσει περισσότερα δεδομένα εκπαίδευσης, η κατασκευή των οποίων, όμως, είναι πολύ χρονοβόρα. Τι θα του συνιστούσατε να κάνει πριν επενδύσει χρόνο στην κατασκευή νέων παραδειγμάτων εκπαίδευσης;



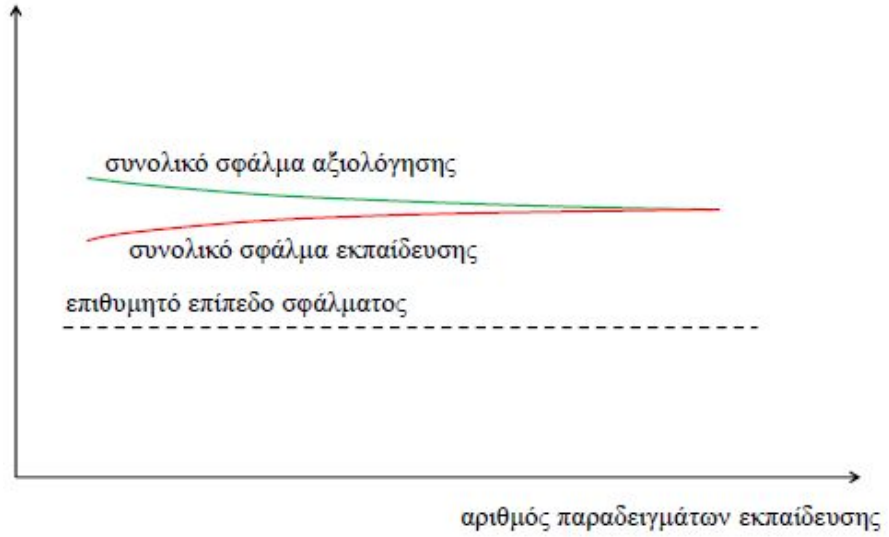
Απάντηση:

Θα του συνιστούσαμε να προσθέσει στο ίδιο διάγραμμα τη γραφική παράσταση του συνολικού σφάλματος στα δεδομένα εκπαίδευσης.

Το σφάλμα στα δεδομένα εκπαίδευσης είναι συνήθως χαμηλότερο από ό,τι το σφάλμα στα δεδομένα αξιολόγησης (για τον ίδιο αριθμό παραδειγμάτων) και αυξάνεται όσο προστίθενται δεδομένα εκπαίδευσης.

Επομένως, αν η καμπύλη του σφάλματος εκπαίδευσης (κόκκινη) βρίσκεται ήδη ψηλότερα από το επιθυμητό επίπεδο σφάλματος, είναι απίθανο η προσθήκη παραδειγμάτων εκπαίδευσης να ρίξει το συνολικό σφάλμα αξιολόγησης στο επιθυμητό επίπεδο.

Επίσης, αν οι δύο καμπύλες έχουν συγκλίνει, το σφάλμα αξιολόγησης θα μειωθεί ελάχιστα προσθέτοντας παραδείγματα εκπαίδευσης.



Άσκηση 18.5.

Απάντηση συνέχεια:

Αντίθετα, αν η καμπύλη του σφάλματος εκπαίδευσης (κόκκινη) είναι ακόμα κάτω από το επιθυμητό επίπεδο, η προσθήκη παραδειγμάτων εκπαίδευσης ενδέχεται να ρίξει το σφάλμα αξιολόγησης (πράσινη καμπύλη) στο επιθυμητό επίπεδο, ιδιαίτερα αν οι δύο καμπύλες απέχουν ακόμα αρκετά και έχουν ακόμα απότομη κλίση.