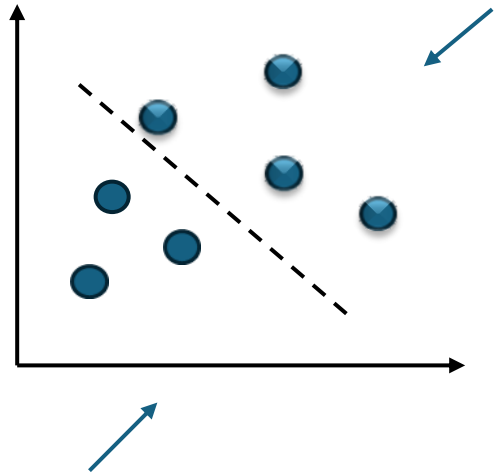
Τεχνητή νοημοσύνη

Φροντιστήριο 9

Ασκήσεις μελέτης της 18ης διάλεξης

Γραμμικοί διαχωριστές



Τα σημεία πάνω από την ευθεία έχουν $\vec{w} \cdot \vec{x} > 0$

Τα σημεία κάτω από την ευθεία έχουν $\vec{w} \cdot \vec{x} < 0$

Για δύο ιδιότητες x_1, x_2 θέλουμε να βρούμε την **ευθεία** που διαχωρίζει τις δύο κατηγορίες.

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

Για περισσότερες ιδιότητες θέλουμε να βρούμε το **υπερ-επίπεδο** που να διαχωρίζει τις δύο κατηγορίες

$$w_0 + w_1 x_1 + .. + w_n x_n = \sum_{l=0}^{n} w_l x_l = \vec{w} \cdot \vec{x} = 0$$

Σιγμοειδής συνάρτηση

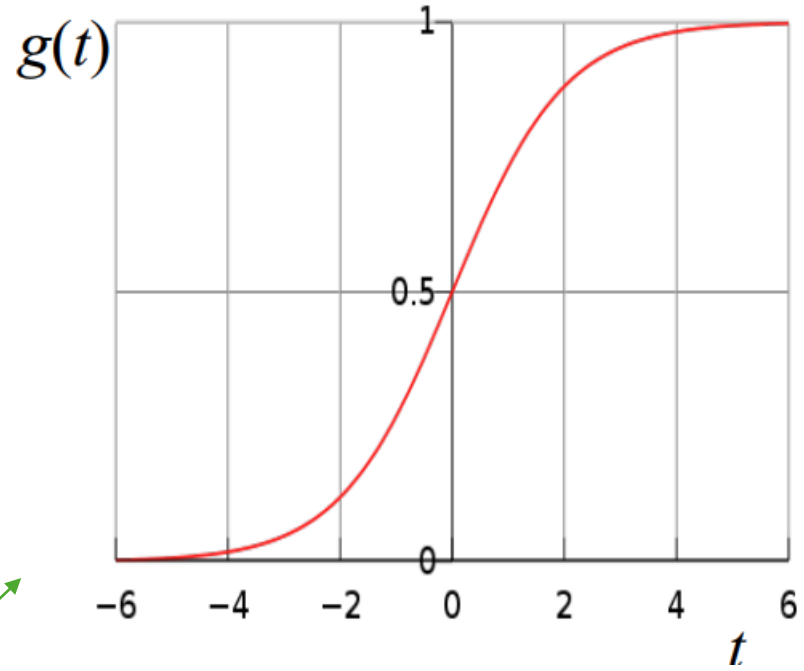
Συχνά θέλουμε ο ταξινομητής να επιστρέφει και ένα **βαθμό βεβαιότητας**.

t = $\vec{w} \cdot \vec{x}$ όπου t η προσημασμένη απόσταση από το επίπεδο διαχωρισμού.

Και άρα η πιθανότητα του $\vec{x}$ να ανήκει στη θετική κατηγορία είναι:

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

Αντίστοιχα, $P(c_- | \vec{x}) = 1 - P(c_+ | \vec{x})$



Μεταξύ 0 και 1

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

Ταξινομητές λογιστικής παλινδρόμησης

Κατά την εκπαίδευση, επιλέγουν το $\vec{w}$ που κάνει τον ταξινομητή πιο βέβαιο ότι τα παραδείγματα εκπαίδευσης ανήκουν στις σωστές κατηγορίες.

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

Μεγιστοποιούμε την δεσμευμένη πιθανοφάνεια

Σωστές κατηγορίες παραδειγμάτων εκπαίδευσης

Παραδείγματα εκπαίδευσης

Μεγιστοποίηση πιθανοφάνειας

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

$$L(\vec{w}) = P(y^{(1)}, \dots, y^{(m)} | \vec{x}^{(1)}, \dots, \vec{x}^{(m)}; \vec{w}) =$$
$$= \prod_{i=1}^{m} P(y^{(i)} | \vec{x}^{(i)}; \vec{w})$$

Αντί να μεγιστοποιήσουμε την $L(\vec{w})$ **μεγιστοποιούμε τον λογάριθμό** της που είναι γνησίως **αύξουσα και κοίλη** οπότε δεν κινδυνεύομε να παγιδευτούμε σε τοπικό μέγιστο:

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

Μεγιστοποίηση πιθανοφάνειας

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

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

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

Για y = 1 (θετική κατηγορία), ο 2ος όρος εξαφανίζεται.

Για y = 0 (αρνητική), ο 1ος όρος εξαφανίζεται.

Μεγιστοποίηση πιθανοφάνειας

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

Για y = 1 (θετική κατηγορία), ο 2ος όρος εξαφανίζεται.

Για y = 0 (αρνητική), ο 1ος όρος εξαφανίζεται.

$$\text{Άρα, } l(\vec{w}) = \sum_{i=1}^{m} \log \mathrm{P}(c_+ \mid \vec{x}^{(i)};\ \vec{w})^{y(i)} + \log \mathrm{P}(c\_ \mid \vec{x}^{(i)};\ \vec{w})^{(1-y(i))} =$$

$$= \boxed{\sum_{i=1}^{m} y^{(i)} \log \mathrm{P}(c_+ \mid \vec{x}^{(i)};\ \vec{w}) + (1 - y^{(i)}) \log \mathrm{P}(c\_ \mid \vec{x}^{(i)};\ \vec{w})}$$

Μεγιστοποίηση πιθανοφάνειας

Με **ανάβαση κλίσης**, ο κανόνας ενημέρωσης των βαρών δίνεται από τον ακόλουθο τύπο:

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla l(\vec{w})$$

Άσκηση 18.1

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

Άσκηση 18.1

$$\sum_{i=1}^{m} y^{(i)} \log \mathrm{P}(c_+|\vec{x}^{(i)};\ \vec{w}) + (1 - y^{(i)}) \log \mathrm{P}(c\_|\vec{x}^{(i)};\ \vec{w}) \quad (1)$$

Από τη σιγμοειδή συνάρτηση έχουμε:

$$\mathrm{P}(c_+|\vec{x}^{(i)};\ \vec{w}) = \frac{1}{1+ e^{-\vec{w}\cdot\vec{x}}} \quad (2)$$

Και άρα $\mathrm{P}(c\_|\vec{x}^{(i)};\ \vec{w}) = \frac{e^{-\vec{w}\cdot\vec{x}}}{1+ e^{-\vec{w}\cdot\vec{x}}}$ (3)

Άσκηση 18.1

Αντικαθιστούμε τις (2) και (3) στην (1):

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

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

$$\cancel{+ y^{(i)} log\left(1+e^{-\vec{w}\vec{x}^{(i)}}\right)}$$

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

Άσκηση 18.1

Με **ανάβαση κλίσης**, ο κανόνας ενημέρωσης των βαρών δίνεται από τον ακόλουθο τύπο:

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla l(\vec{w})$$

Για το $\nabla l(\vec{w})$ ισχύει:

$$\nabla_{\vec{w}} l(\vec{w}) = \langle \frac{\partial l(\vec{w})}{\partial w_0}, \frac{\partial l(\vec{w})}{\partial w_1}, \cdots, \frac{\partial l(\vec{w})}{\partial w_i}, \cdots, \frac{\partial l(\vec{w})}{\partial w_n} \rangle$$

Άσκηση 18.1

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

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

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

$$= \sum_{i=1}^{m} \left[y^{(i)} - P\left(c_+ \middle| \vec{x}^{(i)}\right) \right] x_l^{(i)}$$

Άσκηση 18.1

Άρα, ο κανόνας ενημέρωσης των βαρών θα είναι:

$$w_l \leftarrow w_l + \eta \sum_{i=1}^{m} [y^{(i)} - \mathrm{P}(c_+|\vec{x}^{(i)})]\, x_l^{(i)}$$

Regularization

Υπάρχει έτσι μικρότερος κίνδυνος υπερ-εφαρμογής.

Π.χ. αν πολλά βάρη είναι πολύ μικρά, οι αντίστοιχες ιδιότητες ουσιαστικά δεν χρησιμοποιούνται. Με λιγότερες ιδιότητες έχουμε **μικρότερο κίνδυνο υπερ-εφαρμογής**.

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

L2 Regularization

Άσκηση 18.2

Ποια ακριβώς μορφή παίρνει ο κανόνας ενημέρωσης βαρών της άσκησης 18.1 αν στη λογαριθμική δεσμευμένη πιθανοφάνεια προστεθεί (για να μειωθεί ο κίνδυνος υπερεφαρμογής) ο όρος

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

Άσκηση 18.2

Με **ανάβαση κλίσης**, ο κανόνας ενημέρωσης των βαρών δίνεται από τον ακόλουθο τύπο:

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla l(\vec{w})$$

Προσθέτουμε τον όρο $-\lambda ||\vec{w}||^2 = -\lambda \sum_{i=0}^{n} w_i^2$ και έχουμε:

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

Άσκηση 18.2

Από την 18.1 υπολογίσαμε ότι:

$$\nabla l(\vec{w}) = \sum_{i=1}^{m} [y^{(i)} - P(c_+|\vec{x}^{(i)})]\, x_l^{(i)}$$

Άρα, $\nabla (l(\vec{w}) - \lambda \sum_{l=0}^{n} w_l^2) = \sum_{i=1}^{m} [y^{(i)} - P(c_+|\vec{x}^{(i)})]\, x_l^{(i)} - 2\lambda w_l$

Άρα, ο κανόνας ενημέρωσης των βαρών γίνεται:

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

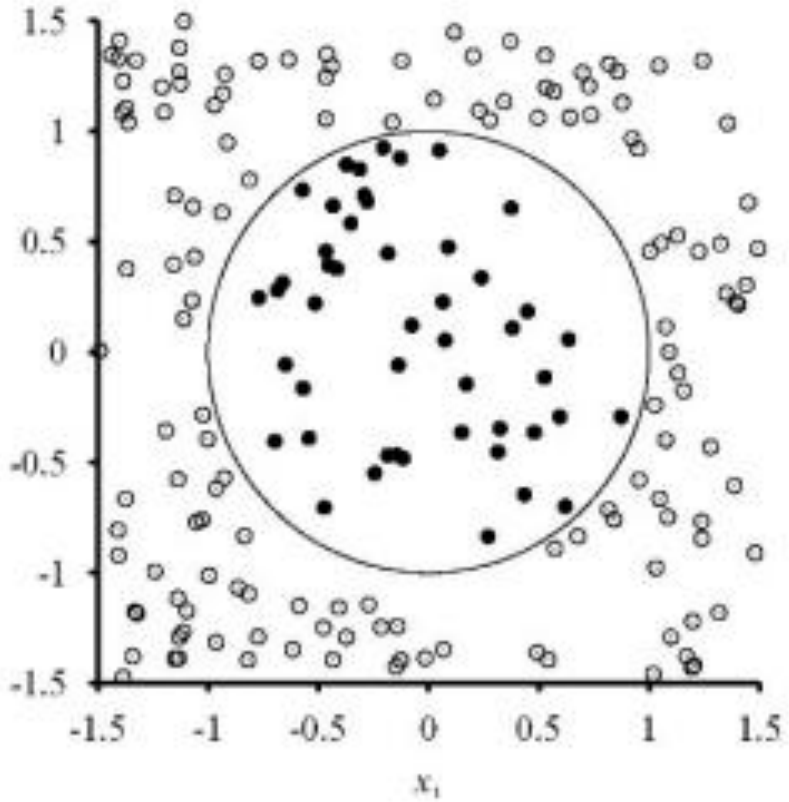
ή αλλιώς $w_l \leftarrow (1 - 2\lambda\eta)\, w_l + \eta \cdot \sum_{i=1}^{m} [y^{(i)} - P(c_+|\vec{x}^{(i)})]\, x_l^{(i)}$

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;



Άσκηση 18.3

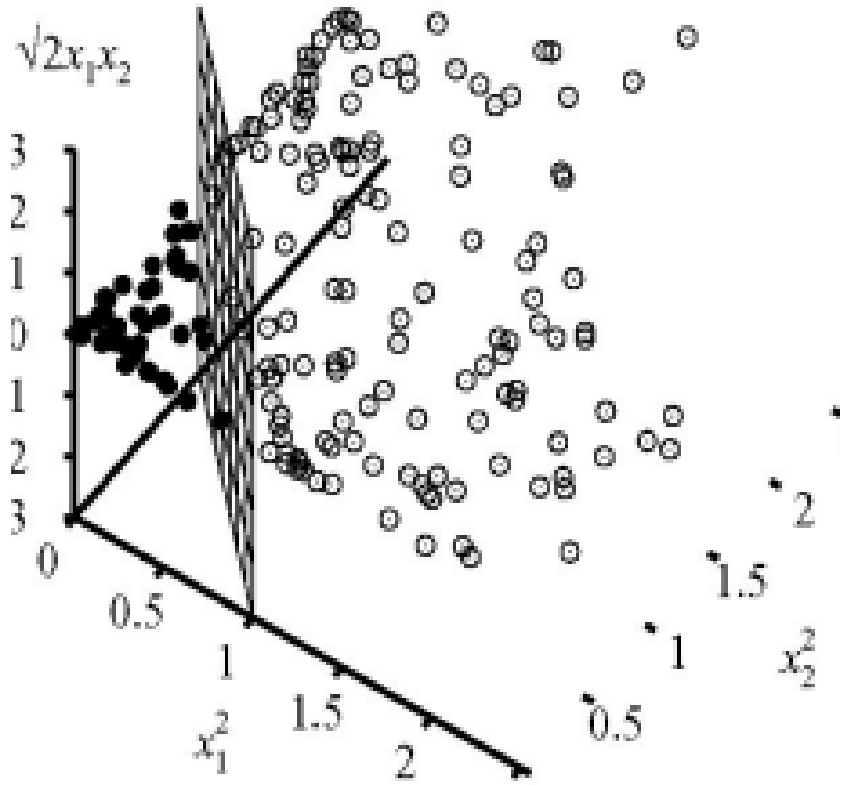
Οι ταξινομητές λογιστικής παλινδρόμησης είναι **γραμμικοί διαχωριστές**, δηλαδή κατά την εκπαίδευση μαθαίνουν μια ευθεία γραμμή, επίπεδο ή γενικότερα υπερ-επίπεδο και κατά την αξιολόγηση κατατάσσουν τις περιπτώσεις (διανύσματα ιδιοτήτων) που τους δίνουμε σε δύο κατηγορίες, ανάλογα με το αν το διάνυσμα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

Άσκηση 18.3

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

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

τα διανύσματα (τελείες) του σχήματος της εκφώνησης διατάσσονται σε ένα νέο τριδιάστατο διανυσματικό χώρο (έχουμε τώρα τρεις ιδιότητες, που αντιστοιχούν στους τρεις άξονες του νέου χώρου), και είναι πλέον γραμμικά διαχωρίσιμα.



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;

Άσκηση 18.4

Η ανάβαση κλίσης αναζητεί στο χώρο των βαρών των ιδιοτήτων το διάνυσμα βαρών (σημείο του χώρου) που μεγιστοποιεί τη δεσμευμένη πιθανοφάνεια των παραδειγμάτων εκπαίδευσης.

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

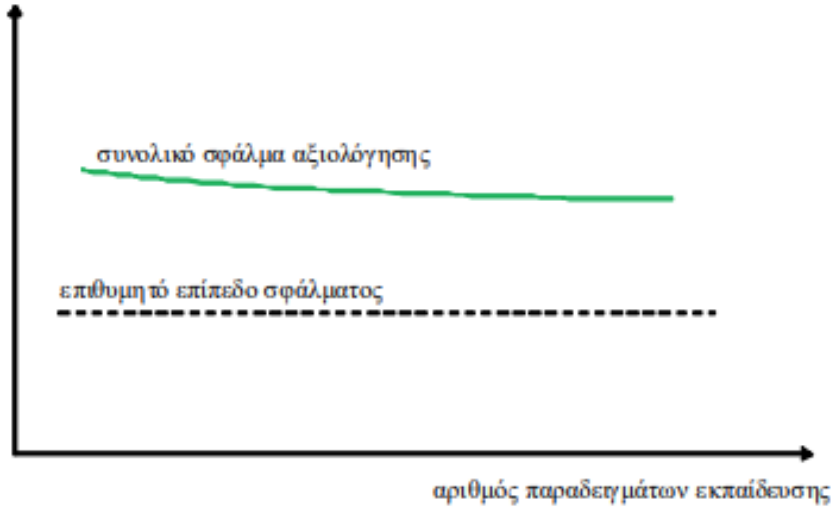
Άσκηση 18.5

Ένας συνάδελφός σας αναπτύσσει ένα σύστημα που χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση.

Το συνολικό σφάλμα στα δεδομένα αξιολόγησης, όμως, παραμένει αρκετά υψηλότερα από το επιθυμητό επίπεδο.

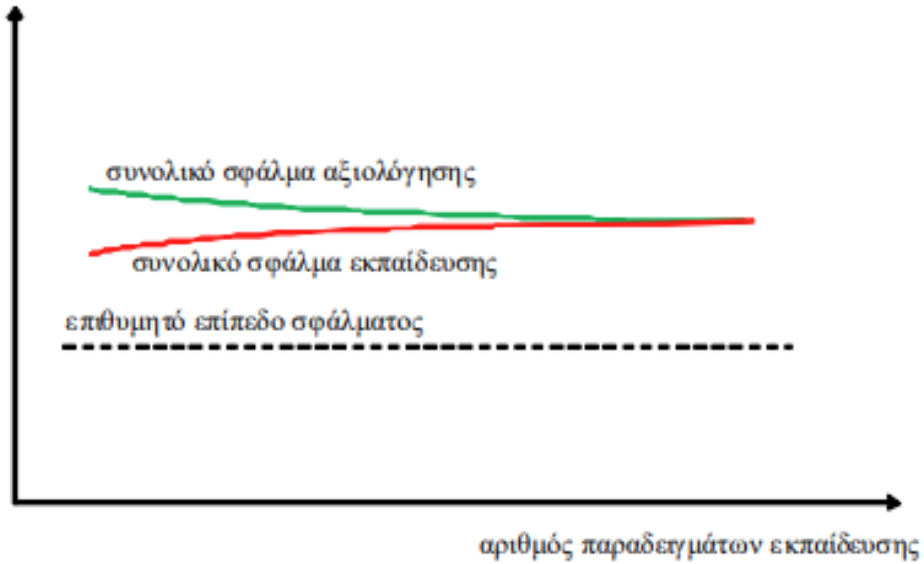
Για να μειώσει το σφάλμα αξιολόγησης, σκέφτεται να προσθέσει περισσότερα δεδομένα εκπαίδευσης, η κατασκευή των οποίων, όμως, είναι πολύ χρονοβόρα.

Τι θα του συνιστούσατε να κάνει πριν επενδύσει χρόνο στην κατασκευή νέων παραδειγμάτων εκπαίδευσης;



Άσκηση 18.5

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

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.