Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Τεχνητή Νοημοσύνη, 2024–25
Διδάσκων: Ι. Ανδρουτσόπουλος

Ασκήσεις μελέτης της 18ης διάλεξης

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 του ταξινομητή λογιστικής παλινδρόμησης, όταν χρησιμοποιείται (batch) ανάβαση κλίσης, δίνεται από τον ακόλουθο τύπο:

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

όπου log ο φυσικός λογάριθμος και:

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

και:

$$P(c_+|\vec{x};\vec{w}) = 1/\left(1 + e^{-\vec{w}\vec{x}}\right) \; (2)$$
$$P(c_-|\vec{x};\vec{w}) = e^{-\vec{w}\vec{x}}/\left(1 + e^{-\vec{w}\vec{x}}\right) \; (3)$$

Αντικαθιστώντας τις (2), (3) στην (1):

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

Έχουμε επίσης:

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

Άρα, για $l \in \{0, \dots, n\}$:

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

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

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

Συνεπώς, ο κανόνας ενημέρωσης βαρών θα είναι:

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

18.2. Ποια ακριβώς μορφή παίρνει ο κανόνας ενημέρωσης βαρών της άσκησης 18.1 αν στη λογαριθμική δεσμευμένη πιθανοφάνεια $l(\vec{w})$ 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:

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

Σύμφωνα με την άσκηση 18.1:

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

Επομένως:

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

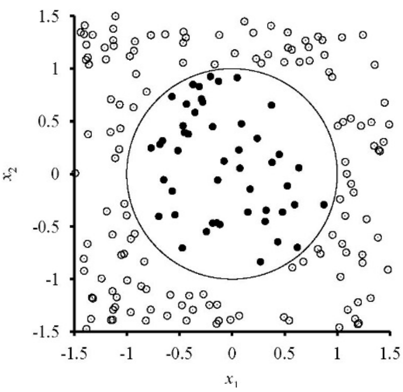
Άρα ο κανόνας ενημέρωσης βαρών γίνεται:

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

ή ισοδύναμα:

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

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 <u>ίδια</u> παραδείγματα που χρησιμοποιήσαμε για την εκπαίδευσή του. Θα κατατάξει όλα τα παραδείγματα εκπαίδευσης στις σωστές κατηγορίες; Αν ναι, γιατί; Αν όχι, γιατί και τι



Σχήμα από το βιβλίο των Russel και Norvig.

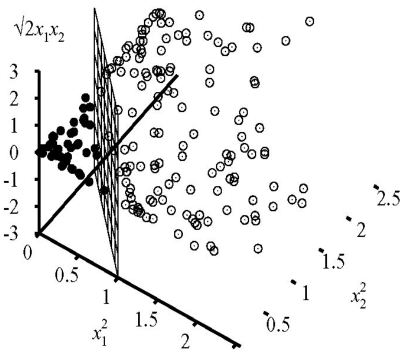
θα μπορούσαμε να κάνουμε, για να βοηθήσουμε τον ταξινομητή λογιστικής παλινδρόμησης να τα κατατάξει σωστά;

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.

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

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

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), 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.

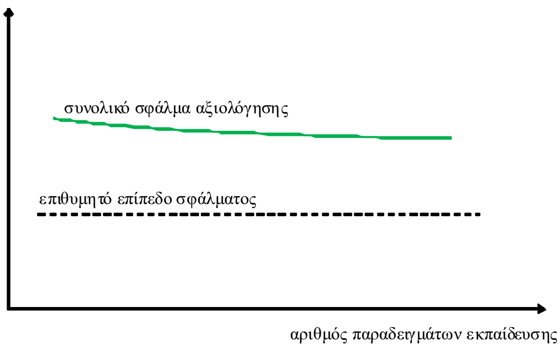


Σχήμα από το βιβλίο των Russel και Norvig.

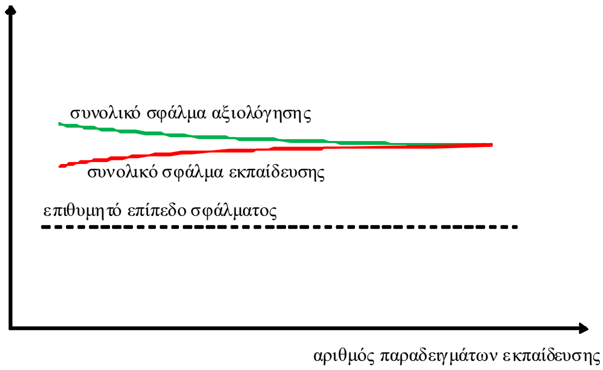
18.4. Ένας φοιτητής εκπαιδεύει έναν ταξινομητή λογιστικής παλινδρόμησης με (batch) ανάβαση κλίσης. Προκειμένου η εκπαίδευση να ολοκληρώνεται πιο γρήγορα, αύξησε πολύ την τιμή της σταθεράς η του κανόνα ενημέρωσης βαρών, ελπίζοντας ότι έτσι θα εκτελούνταν λιγότερα βήματα κατά την ανάβαση κλίσης. Παρατήρησε 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) $\vec{w}$ 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.

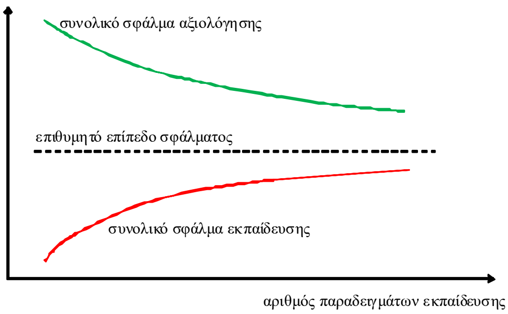
18.5. Ένας συνάδελφός σας αναπτύσσει ένα σύστημα που χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση. Το συνολικό σφάλμα στα δεδομένα αξιολόγησης, όμως, παραμένει αρκετά υψηλότερα από το επιθυμητό επίπεδο. Για να μειώσει το σφάλμα αξιολόγησης, σκέφτεται να προσθέσει περισσότερα δεδομένα



εκπαίδευσης, η κατασκευή των οποίων, όμως, είναι πολύ χρονοβόρα. Τι θα του συνιστούσατε να κάνει πριν επενδύσει χρόνο στην κατασκευή νέων παραδειγμάτων εκπαίδευσης;



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.



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.