Μέθοδοι βελτιστοποίησης

Μέθοδοι βαθμίδας - *Gradient Schemes*

Α. Ν. Γιαννακόπουλος

Ο.Π.Α

Εαρινό Εξάμηνο 2026

1. Βασικές έννοιες

2. *Gradient* σχήματα
   - Προσέγγιση του $Df(x)$
   - *Steepest descent* για τετραγωνικές συναρτήσεις
   - Σύγκλιση της μεθόδου *gradient descent*
   - Παραλλαγές της μεθόδου *gradient descent*
   - Εφαρμογή 1: *Ridge regression – Tikhonov regularization*
   - Εφαρμογή 2: *SVM*
   - Εφαρμογή 3: *Recommender systems* και παραγοντοποίηση πινάκων

Η βαθμίδα (gradient) και οι ιδιότητες της

Έστω $f : D \subset \mathbb{R}^d \to \mathbb{R}$ μια συναρτηση d-μεταβλητών.

Ορισμός

Το διάνυσμα $Df(x) = (\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_1}(x))$ *ονομαζεται η βαθμίδα της* f *στο σημείο* $x \in D$.

Παράδειγμα

Δίνεται η συνάρτηση $f(x) = (\sum_{i=1}^{d}(x_i - x_{i,0})^2)^{1/2}$ *που εκφράζει την απόσταση του σημείου* $x = (x_1, \cdots, x_d)$ *απο το σημείο* $x_0 = (x_{1,0}, \cdots, x_{d,0})$. *Η βαθμίδα της* f *στο σημείο* x *δίνεται απο το διάνυσμα*

$$Df(x) = \frac{1}{f(x)}(x_1 - x_{1,0}, \cdots, x_d - x_{d,0}).$$

Η βαθμίδα μπορει να εκφράσει την μεταβολή της συναρτησης f σε καποια συγκεκριμένη κατευθυνση $h \in \mathbb{R}^d$, μεσω της κατευθυνόμενης παραγώγου,

$$D_h f(x) = \frac{d}{d\epsilon} f(x + \epsilon h)\Big|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(f(x + \epsilon h) - f(x)) = \langle Df(x), h \rangle$$

Απο την ανισότητα Cauchy-Schwarz βλεπουμε πως η

- $D_h f(x)$ παίρνει την ελαχιστη τιμή της για $h = -\frac{Df(x_0)}{\|Df(x_0)\|}$ – διεύθυνση μεγιστης μειωσης της f τοπικά στο x,
- $D_h f(x)$ παίρνει την μεγιστη τιμή της για $h = \frac{Df(x_0)}{\|Df(x_0)\|}$ – διεύθυνση μεγιστης αυξησης της f τοπικά στο x,

Η βαθμίδα συνδέεται με την επίλυση προβλημάτων βελτιστοποίησης.

Θεώρημα

Έστω $f : D \subset \mathbb{R}^d \to \mathbb{R}$, όπου D ανοιχτό, μια C^1 συναρτηση.
Το σημείο $x^* \in D$ είναι ένα τοπικό ακρότατο της f αν και μονο αν ειναι κρίσιμο σημείο δηλαδη,

$$Df(x^*) = 0$$

Για τον χαρακτηρισμό των τοπικών ακροτάτων x^*, χρειαζόμαστε παραπάνω πληροφορία.

Για συναρτήσεις οι οποίες ειναι κυρτές τα τοπικα ακρότατα αντιστοιχούν σε ολικά ελαχιστα

Αλλιώς θα πρέπει να χρησιμοποιήσουμε τις συνθήκες δευτερης τάξης που έχουν νόημα για συναρτήσεις πιο λείες και που χρησιμοποιούν την δευτερη παράγωγο - πινακας του Hess

Ορισμός

Αν η $f : D \subset \mathbb{R}^d \to \mathbb{R}$ *είναι* C^3 *ο πίνακας του Hess στο σημείο* x *είναι ο συμμετρικός πίνακας*

$$D^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\cdots,d}.$$

Για τέτοιες συναρτήσεις ισχυει το θέωρημα του Taylor

$$f(x+\epsilon h) = f(x) + \epsilon\langle Df(x), h\rangle + \frac{1}{2}\epsilon^2\langle h, D^2 f(x)h\rangle + O(\epsilon^2),$$

συνεπώς,

- x^* τοπικό ελάχιστο αν $D^2 f(x^*)$ θετικά ημι-ορισμενος
- x^* τοπικό μεγιιστο αν $-D^2 f(x^*)$ θετικά ημι-ορισμενος
- x^* σαγματικό σε καθε άλλη περίπτωση

Gradient σχηματα

Οι μέθοδοι αυτοί ειναι μέθοδοι βελτιστοποίησης οι οποίοι βασιζονται σε επαναληπτικά σχήματα που ξεκινώντας απο κάποια αρχικη εκτίμηση x_0 για το τοπικό ακρότατο, την βελτιώνουν σταδιακά με νέες εκτιμήσεις x_k, μεχρις ότου καταλήξουν σε μια επαρκώς καλή εκτίμηση του.

Η βελτίωση γίνεται σε κάθε επανάληψη k μετακινούμενοι προς μια διευθυνση h_k κατα στην οποία επιτυγχάνεται η μεγιστη μείωση (ή αύξηση) της τιμής της συναρτησης f.

Για την ελαχιστοποίηση της συνάρτησης f, μία κατάλληλη διεύθυνση μειωσης τοπικα στην θέση x_k είναι η διεύθυνση $-\alpha_k Df(x_k)$ για κατάλληλη επιλογή του $\alpha_k > 0$.

$$x_{k+1} = x_k - \alpha_k Df(x_k), \;\; \alpha_k > 0$$

Η κατάλληλη επιλογή του $a_k > 0$ θα είναι έτσι ώστε να έχουμε τοπικα την μικρότερη δυνατή τιμή για την συναρτηση f,

$$x_{k+1} = x_k - \alpha_k Df(x_k), \quad \alpha_k > 0$$
$$\alpha_k = \arg\min_{a>0} f(x_k - aDf(x_k))$$

Το πρόβλημα βελτιστοποίησης για την εύρεση του α_k είναι ένα μονοδιάστατο πρόβλημα βελτιστοποίησης το οποίο μπορεί να επιλυθεί σχετικα ευκολα.

Προσέγγιση του $Df(x)$

Με e_j συμβολίζουμε το μοναδιαίο διάνυσμα την i διευθυνση.

Τις περισσότερες φορές η προσέγγιση του $Df(x)$ γίνεται με ένα σχήμα διαφορών.

$$\frac{\partial f}{\partial x_i}(x) = \frac{1}{\epsilon}(f(x + \epsilon e_i) - f(x))$$

ή με ένα σχημα κεντρικών διαφορών

$$\frac{\partial f}{\partial x_i}(x) = \frac{1}{2\epsilon}(f(x + \epsilon e_i) - f(x - \epsilon e_i))$$

για αρκετά μικρό $\epsilon > 0$.

Εύρεση του α_k

Η ευρεση του α_k ειναι ένα μονοδιάστατο προβλημα βελτιστοποίησης για την συναρτηση μίας μεταβλητής

$$a \mapsto \varphi(a) := f(x_k - aDf(x_k)).$$

Αυτό μπορεί να λυθεί με διάφορες μεθόδους π.χ.

- Η μεθοδος της διχοτόμου (bisection)
- Η μέθοδος της χρυσής τομής (golden section)
- Η μεθοδος της εφαπτομένης
- Η μέθοδος Newton- Raphson

Bisection

Για την ελαχιστοποίηση της φ:

1. Επιλέγουμε αρχικα ένα διάστημα $[a, b]$ τέτοιο ώστε $\varphi'(a)\varphi'(b) < 0$.
2. Μετά παίρνουμε τα διαστηματα $[a, \frac{a+b}{2}]$ και $[\frac{a+b}{2}, b]$.
   - Αν $\varphi'(a)\varphi'(\frac{a+b}{2}) < 0$ το σημειο μηδενισμού της φ' βρίσκεται στο $[a, \frac{a+b}{2}]$.
     Θέτουμε $a = a$ και $b = \frac{a+b}{2}$ και ξαναπάμε στο βήμα 2
   - Αν $\varphi'(\frac{a+b}{2})\varphi'(b) < 0$ το σημειο μηδενισμού της φ' βρίσκεται στο $[\frac{a+b}{2}, b]$
     Θέτουμε $a = \frac{a+b}{2}$ και $b = b$ και ξαναπάμε στο βήμα 2
3. Συνεχίζουμε μεχρις ότου $b - a < \epsilon$

Η μεθοδος της χρυσής τομής - Golden section

Για να βρούμε ένα ελάχιστο της φ στο διάστημα $[a, b]$

1. Παίρνουμε το καινούργιο διάστημα $[x_1, x_2] \subset [a, b]$ με

$$x_1 = a + r(b - a),$$
$$x_2 = b - r(b - a),$$
$$r = \frac{3 - \sqrt{5}}{2} \simeq 0.3819$$

   - Αν $\varphi(x_1) < \varphi(x_2)$ το ελάχιστο θα πρέπει να βρισκεται σε κάποιο $a^* \in [a, x_2]$.
     Θέτουμε $a = a$ και $b = x_2$ και ξαναπάμε στο βήμα 1.
   - Αν $\varphi(x_1) \geq \varphi(x_2)$ το ελάχιστο θα πρέπει να βρισκεται σε κάποιο $a^* \in [x_1, b]$.
     Θέτουμε $a = x_1$ και $b = b$ και ξαναπάμε στο βήμα 1.

2. Επαναλαμβάνουμε μεχρις ότου $b - a < \epsilon$.

Η μεθοδος Newton - Raphson

Η μεθοδος αυτή βασίζεται στο ανάπτυγμα Taylor για την συνάρτηση φ,

$$\varphi(a) \simeq \varphi(a_0) + \varphi'(a_0)(a - a_0) + \frac{1}{2}\varphi''(a_0)(a - a_0)^2$$

Ελαχιστοποιουμε τώρα αυτή την τετραγωνική προσέγγιση κοντά στο a_0 αντί της ίδιας της φ, και βρισκουμε ελάχιστο στο

$$a^* = a_0 - \frac{\varphi'(a_0)}{\varphi''(a_0)}$$

Αλλάζοντας καθε φορά το a_0 γύρω απο το οποίο παιρνουμε την τετραγωνική προσέγγιση, βασισμενοι στην παραπάνω ιδέα παιρνουμε το επαναληπτικό σχήμα

$$a_{k+1} = a_k - \frac{\varphi'(a_k)}{\varphi''(a_k)},$$

το οποίο υπο συνθήκες συγκλίνει στο $a^* = \arg\min_a \varphi(a)$.

Η μεθοδος της εφαπτομενης

Η μέθοδος της εφαπτομενης ειναι κατα καποιο τρόπο συναφής με την μέθοδο Newton - Raphson μόνο που η δεύτερη παράγωγος προσεγγίζεται απο την εφαπτομένη.

Προσομοιάζει με την μέθοδο Newton- Raphson για την λύση της $\varphi'(a) = 0$ ενώ συνδυαζει και κάποια στοιχεία απο την μέθοδο της διχοτόμου.

Σύμφωνα με την μέθοδο αυτή η τοπική προσέγγιση για την ευρεση του ελαχιστου γίνεται απο το επαναληπτικό σχήμα

$$a_{k+1} = a_k - \varphi'(a_k)\left(\frac{a_k - a_{k-1}}{\varphi'(a_k) - \varphi'(a_{k-1})}\right)$$

Steepest descent για τετραγωνικές συναρτήσεις

Ας πάρουμε την τετραγωνική συνάρτηση της μορφής

$$f(x) = \frac{1}{2}\langle x, Qx\rangle - \langle b, x\rangle$$

με $Q \in \mathbb{R}^{d\times d}$ συμμετρικό και θετικα ορισμενο πίνακα και $b \in \mathbb{R}^d$.

Έχουμε ότι

$$Df(x) = Qx - b$$

Η συναρτηση f είναι κυρτή οπότε για την ευρεση του ελαχιστου αρκεί να λύσουμε το γραμμικό σύστημα

$$Qx = b$$

Η μεθοδος steepest descent στην περίπτωση αυτή μπορεί να γραφεί ως εξής

$$x_{k+1} = x_k - \alpha_k(Qx_k - b),$$
$$a_k = \arg\min_{a>0} \varphi(a),$$

$$\varphi(a) = \frac{1}{2}\langle x_k - a(Qx_k - b), Q(x_k - a(Qx_k - b))\rangle - \langle b, x_k - a(Qx_k - b)\rangle$$

ή στην πιο συμπαγή μορφή

$$g_k = Qx_k - b,$$
$$x_{k+1} = x_k - \alpha_k g_k,$$
$$\alpha_k = \arg\min_a \frac{1}{2}\langle x_k - a\ g_k, Q(x_k - ag_k)\rangle - \langle b, x_k - ag_k\rangle$$

Η επιλογή του a_k μπορεί να γίνει αναλυτικα γιατι οδηγεί σε ένα τετραγωνικό πρόβλημα βελτιστοποίησης.

Έχουμε ότι

$$\varphi(a) = \frac{1}{2}\langle g_k, Qg_k\rangle a^2 + (\langle b, g_k\langle -\langle g_k, Qx_k\rangle)a - \langle b, x_k\rangle$$
$$= \frac{1}{2}\langle g_k, Qg_k\rangle a^2 - \langle g_k, g_k\rangle a - \langle b, x_k\rangle,$$

όπου χρησιμοποιήσαμε την συμμετρία του Q και τον ορισμό του g_k. Μπορούμε λοιπόν να βρούμε οτι το ελάχιστο για την συνάρτηση αυτή επιτυγχάνεται για

$$a^* = \frac{\langle g_k, g_k\rangle}{\langle g_k, Qg_k\rangle} > 0$$

$$g_k = Qx_k - b,$$
$$a_k = \frac{\langle g_k, g_k \rangle}{\langle g_k, Qg_k \rangle},$$
$$x_{k+1} = x_k - a_k g_k$$

Ιδιότητες της ακολουθίας x_k

- Για κάθε $k \in \mathbb{N}$ έχουμε ότι

$$x_{k+2} - x_{k+1} \perp x_{k+1} - x_k$$

- Για κάθε $k \in \mathbb{N}$ έχουμε ότι

$$f(x_{k+1}) \leq f(x_k)$$

Σύγκλιση της μεθόδου gradient descent

Θεώρημα

Αν η συνάρτηση $f : \mathbb{R}^d \to \mathbb{R}$ είναι κυρτή και παραγωγίσιμη με συνεχή κατά Lipschitz βαθμίδα

$$\|Df(x) - Df(y)\| \leq L \|x - y\|, \ \ \forall x, y \in \mathbb{R}^d,$$

σε k επαναλήψεις της μεθόδου gradient descent με βήμα $a \leq \frac{1}{L}$ θα επιτύχουμε βελτίωση σχετικα με την τιμή της συνάρτησης f ως ακολούθως

$$f(x_k) - f(x^*) \leq \frac{1}{2a}\frac{1}{k}\|x_0 - x^*\|^2.$$

- Η σύγκλιση της μεθόδου είναι αργή και ακόμα και για συναρτήσεις με Lipschitz συνεχή παράγωγο επιτυγχάνεται σε $O(1/\epsilon)$ βήματα για ακρίβεια $|f(x_k) - f(x^*)| \leq \epsilon$.
- Η συγκλιση επιταχύνεται αν η συνάρτηση f είναι ισχυρά κυρτή, δηλ. αν η $f(x) - \frac{\lambda}{2}\|x\|^2$, είναι κυρτή συνάρτηση για κάποιο $\lambda > 0$.

Παραλλαγές της μεθόδου gradient descent

Πολλες φορές η επιλογή του a_k με τον βέλτιστο τροπο μέσω της λύσης του

$$a_k = \arg\min_{a>0} f(x_k - aDf(x_k))$$

μπορει να ειναι χρονοβόρα και να προτιμήσουμε να την αποφύγουμε.

Στις περιπτώσεις αυτές μπορούμε να επιλέξουμε ενα βημα μέσω άλλων κριτηρίων

$$x_{k+1} = x_k - \rho_k Df(x_k), \quad \rho_k > 0,$$

με ρ_k επιλεγμενο βάσει καποιου άλλου κριτήριου.

- Αυθαιρετη επιλογή σταθερού βηματος $\rho_k = \rho$.
- Επιλογή μεταβαλλόμενου βήματος ρ_k βασει διαφόρων κριτηριων.
  - $\rho_k \to 0$ με $\sum_{k=1}^{\infty} \rho_k = \infty$ - ο κανονας αυτός εξασφαλίζει οτι εκει που θα συγκλίνει η μέθοδος θα ισχύει $Df(x^*) = 0$.
  - Επιλογή των ρ_k ετσι ωστε $f(x_k - a_k g_k) < f(x_k)$ με τρόπο ώστε η μειωση να είναι επαρκής, π.χ.

$$f(x_k - a_k g_k) \leq f(x_k) - c\, a_k \langle Df(x_k), Df(x_k)\rangle, \quad \text{Armijo rule,}$$

για κάποιο $c \in (0,1)$, συνήθως μικρό, ή

$$\sigma \leq \frac{f(x_k) - f(x_k - a_k g_k)}{a_k \langle Df(x_k), Df(x_k)\rangle} \leq 1 - \sigma, \quad \text{Goldstein rule,}$$

για κάποιο $\sigma \in (0, 1/2)$.

Coordinate gradient descent

Στην παραλλαγή της gradient descent που ονομαζεται Coordinate gradient descent η ανανέωση του διανύσματος x γίνεται μια συνιστώσα την φορά.

Αυτό αντιστοιχεί σε κάθε βήμα στην επίλυση του μονοδιάστατου προβλήματος ελαχιστοποίησης

$$x_{i,k+1} = \arg\min_{z\in\mathbb{R}} f(x_{1,k+1}, \cdots, x_{i-1,k+1}, z, x_{i+1,k}, \cdots, x_{d,k})$$

Ο αλγοριθμος αυτός είναι ιδιαίτερα δημοφιλής για συναρτήσεις f που έχουν την μορφή

$$f(x) = \sum_{i=1}^{n} f_i(x_i).$$

Εφαρμόζεται σε μεγάλη γκάμα προβλημάτων μηχανικης μάθυσης π.χ. σε προβλήματα εκμάθησης SVM ή σε προβλήματα παραγοντοποίησης

Στην πράξη ο αλγόριθμος τροποποιείται ως

Αλγόριθμος (Coordinate gradient descent)

- *Για κάποια αρχική επιλογή x.*
- *Μέχρις ότου κάποιο κριτήριο σύγκλισης ικανοποιηθεί*
  1. *Για κάθε συντεταγμένη $i \in \{1, \cdots, n\}$.*
  2. *Επιλέγουμε μέγεθος βήματος a.*
  3. *Ανανεώνουμε την αντίστοιχη συντεταγμένη ως*

$$x_i \leftarrow x_i - a\frac{\partial f}{\partial x_i}(x)$$

Το βήμα 1 μπορεί να τροποποιηθεί επιλέγοντας τις συντεταγμενες όχι κυκλικά αλλά με κάποιο άλλο σχημα επιλογής π.χ. τυχαία.

Μία άλλη παραλλαγή της μεθόδου είναι να κάνουμε την ανανέωση κατα blocks συνιστωσών αντί για συνιστώσα προς συνιστώσα.

Εφαρμογή 1: Ridge regression - Tikhonov regularization

Η ridge regression προσαρμόζει ένα μοντέλο γραμμικής παλινδρομησης

$$Y = w_1 X_1 + \cdots + w_d X_d,$$

στα δεδομένα $\mathbf{X}_i$, Y_i, λύνοντας το πρόβλημα ελαχιστοποίησης

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(w) = \min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|D\mathbf{w} - Y\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2, \;\; \lambda > 0.$$

Στο όριο $\lambda \to 0$, οδηγούμαστε στην τυπική γραμμική παλινδρομηση.

Η μεθοδος της βαθμίδας μπορεί να χρησιμοποιηθεί για την αριθμητική επίλυση του προβλήματος αυτού.

Ο υπολογισμός της βαθμίδας

Χρησιμοποιώντας τις ιδιότητες του εσωτερικού γινομένου

$$\mathcal{L}(\mathbf{w}+\epsilon\bar{\mathbf{w}})-\mathcal{L}(\mathbf{w})=\epsilon\langle D\mathbf{w}-y,D\bar{\mathbf{w}}\rangle+\epsilon\lambda\langle\mathbf{w},\bar{\mathbf{w}}\rangle+O(\epsilon^2),$$

οπότε

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mathcal{L}(\mathbf{w}+\epsilon\bar{\mathbf{w}})-\mathcal{L}(\mathbf{w})) &= \langle D\mathbf{w}-y,D\bar{\mathbf{w}}\rangle+\lambda\langle\mathbf{w},\bar{\mathbf{w}}\rangle\\ &= \langle D^T(D\mathbf{w}-y),\bar{\mathbf{w}}\rangle+\lambda\langle\mathbf{w},\bar{\mathbf{w}}\rangle\\ &= \langle D^T(D\mathbf{w}-y)+\lambda\mathbf{w},\bar{\mathbf{w}}\rangle\\ &=: \langle D\mathcal{L}(\mathbf{w}),\bar{\mathbf{w}}\rangle\end{aligned}$$

οπότε

$$D\mathcal{L}(\mathbf{w})=D^T(D\mathbf{w}-y)+\lambda\mathbf{w}=(D^TD+\lambda I)\mathbf{w}-D^Ty$$

Η συνθήκη πρώτης τάξης γίνεται

$$D\mathcal{L}(\mathbf{w})=0 \iff (D^TD+\lambda I)\mathbf{w}=D^Ty$$

Η μέθοδος της βαθμίδας οδηγεί στο επαναληπτικό σχήμα

$$\mathbf{w}_{k+1} = \mathbf{w}_k - a_k D\mathcal{L}(\mathbf{w}_k),$$

ή ισοδύναμα

$$\mathbf{w}_{k+1} = \mathbf{w}_k - a_k[(D^T D + \lambda I)\mathbf{w}_k - D^T y].$$

Το παραπάνω σχήμα μπορεί να γραφεί και ως

$$\mathbf{w}_{k+1} = (1 - a_k\lambda)\mathbf{w}_k - a_k D^T(D\mathbf{w}_k - y).$$

Το διάνυσμα $\mathcal{E}_k := D\mathbf{w}_k - y$ αντιστοιχεί στο σφάλμα που παρουσιάζει το μοντέλο μας στην k-επανάληψη της μεθόδου.

Με τον παραπάνω ορισμό η μέθοδος επαναδιατυπώνεται ως

$$\mathbf{w}_{k+1} = (1 - a_k\lambda)\mathbf{w}_k - a_k D^T\mathcal{E}_k.$$

Η παράμετρος $a_k > 0$ ονομαζεται ρυθμός μαθησης (learning rate)

Παρατηρείστε ότι η συνάρτηση σφάλματος (loss function) χωρίς τον όρο κανονικοποίησης μπορεί να γραφει ισοδύναμα με την μορφή

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w}),$$

$$\mathcal{L}_i(\mathbf{w}) = \frac{1}{2}|(D\mathbf{w} - y)_i|^2 = \frac{1}{2}\sum_{j=1}^{d}(X_{ij}w_j - y_i)^2$$

όπου $\mathcal{L}_i(\mathbf{w})$ είναι το σφάλμα του μοντέλου για τα δεδομένα $\mathbf{X}_i, y_i)$.

Η βαθμίδα μπορει να γραφεί σαν το άθροισμα

$$D\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} D\mathcal{L}_i(\mathbf{w}),$$

$$D\mathcal{L}_i = (D\mathbf{w} - y)_i\mathbf{X}_i^T =: E_i\mathbf{X}_i^T$$

με E_i το σφάλμα που παρουσιάζει το μοντέλο στο i - δεδομένο $(\mathbf{X}_i, y_i)$.

Το επαναλήπτικό σχήμα βαθμίδας μπορεί να γραφει

$$\mathbf{w}_{k+1} = \mathbf{w}_k - a\sum_{i=1}^{n} D\mathcal{L}_i(\mathbf{w}_k)$$

$$= \mathbf{w}_k - a\sum_{i=1}^{n} \mathbf{X}_i^T E_i$$

Stochastic gradient descent

Μπορούμε να ανανεώνουμε το $\mathbf{w}$ χρησιμοποιώντας κάθε φορά μονο ένα όρο του παραπάνω αθροίσματος, δηλ. να σπάσουμε το βήμα απο k στο $k+1$ σε n επιμέρους βήματα

1. Για $i = 1, \cdots, n$ επαναλαμβάνουμε
2. $\mathbf{w} \leftarrow \mathbf{w} - aD\mathcal{L}_i(\mathbf{w})$

Το μπλόκ των βημάτων 1 και 2 επαναμβάνεται όσες φορές χρειαστεί μέχρι ο αλγόριθμος να συγκλίνει όμως κάθε φορά που ξαναξεκινάμε ενδεχομένως να ανακατέψουμε τα δεδομένα στο δείγμα μας.

Ο όρος της κανονικοποίηαης μπορεί εύκολα να ενσωματωθεί στο παραπάνω σχήμα.

Ο αλγόριθμος Stochastic Gradient Descent ανανεώνει τα βάρη του μοντέλου $\mathbf{w}$, σε κάθε ένα απο τα δεδομένα στο συνολο δεδομένων εκμάθησης (training set), ενδεχομένως καθώς αυτά συλλέγονται και συμπληρώνουν το training set, (online learning).

Ο παραπάνω αλγόριθμος χρησιμοποιείται συχνά σε προβλήματα μηχανικής μαθησης μεγάλης κλίμακας.

Στα προβλήματα αυτά, ο ακριβής υπολογισμός της βαθμίδας χρειάζεται ένα άθροισμα με μεγάλο πλήθος προσθετέων (όσοι και το μέγεθος του δείγματος μάθησης) το οποίο υπολογιστικά μπορεί να ειναι πολύ ακριβό!

Σε αντίθεση, ο αλγοριθμος stochastic gradient descent μπορει να τροποποιηθεί έτσι ώστε να δειγματίζει ένα υποσυνολο των δεδομένων σε κάθε βήμα, πράγμα που τον καθιστά πολύ πιο αποδοτικό.

Εφαρμογή 2: SVM

Το πρόβλημα της κατασκευής ενός SVM που διαχωρίζει σημεία $\mathbf{X}_i \in \mathbb{R}^d$ σε ένα d-διάστατο feature space σε δύο κλάσεις $Y_i = +1$ και $Y_i = -1$ μπορεί να γραφεί σαν ενα πρόβλημα βελτιστοποίησης της μορφής

$$\min_{w\in\mathbb{R}^d} f(w) := \min_{w\in\mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n}(\max(0, 1 - y_i\langle w, \mathbf{X}_i\rangle))^2 + \frac{\lambda}{2}\|w\|^2$$

L^2-hinge loss

Η συναρτηση f μπορεί να γραφεί σαν το άθροισμα

$$f(w) = \sum_{i=1}^{n} f_i(w) + \frac{\lambda}{2}\|w\|^2,$$
$$f_i(w) = \frac{1}{2}(\max(0, 1 - y_i\langle w, \mathbf{X}_i\rangle))^2,$$

Η βαθμίδα της συνάρτησης f δίνεται απο την σχέση

$$Df(w) = \sum_{i=1}^{n} Df_i(w) + \lambda w,$$
$$Df_i(w) = -\max\{0, 1 - y_i\langle w, \mathbf{X}_i\rangle\}y_i\mathbf{X}_i$$

Η μέθοδος της βαθμίδας για τα L^2-hinge loss SVMs παίρνει την μορφή

$$w_{k+1} = w_k - \alpha_k \left(\sum_{i=1}^{n} Df_i(w) + \lambda w \right) ,, \quad w_k \in \mathbb{R}^d,$$
$$Df_i(w) = -\max\{0, 1 - y_i \langle w, \mathbf{X}_i \rangle\} y_i \mathbf{X}_i$$

Θα μπορούσε επίσης να χρησιμοποιηθεί ειτε το coordinate gradient descent σχήμα ή και το stochastic gradient descent σχήμα, βάσει του οποίου για μεγάλα σύνολα δεδομένων λαμβάνεται μια μια η συνεισφορά των δεδομένων στην βαθμίδα (ένας όρος την φορά - online learning)

Το L^2-hinge loss SVMs είναι στενά συνδεδεμένο με το μοντέλο της λογιστικής παλινδρόμησης (logistic regression) με κατάλληλο όρο ομαλοποίησης.

Εφαρμογή 3: Recommender systems και παραγοντοποίηση πινάκων

Το πρόβλημα της κατασκευής του συστήματος προτάσεων (recommender system) μπορεί λοιπόν να γραφεί σαν ενα πρόβλημα βελτιστοποίησης της μορφής

$$\min_{U\in\mathbb{R}^{n\times k},V\in\mathbb{R}^{k\times d}} \frac{1}{2}\sum_{(i,j)\in\mathcal{O}} |s_{ij}-\sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{k} u_{i\ell}^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{k} v_{j\ell}^2,$$

όπου $\mathcal{O}$ είναι το σύνολο των ζευγών (i,j) που περιέχει τις αξιόλογησεις του κάθε χρήστη i για το (υποσύνολο) των αντικειμένων j που έχει αξιολογήσει.

Ο υπολογισμός της βαθμίδας μπορεί να γίνει απλά (ως προς κάθε στοιχείο των πινάκων U και V,

$$\frac{\partial\mathcal{L}}{\partial u_{i\ell}}(U,V) = \underbrace{(s_{ij}-\sum_{\ell=1}^{k} u_{i\ell}v_{j\ell})}_{e_{ij}:=s_{ij}-\hat{s}_{ij}}(-v_{j\ell}) + \lambda u_{i\ell},\quad i\in\{1,\cdots,n\},\ell\in\{1,\cdots,k\},$$

$$\frac{\partial\mathcal{L}}{\partial v_{j\ell}}(U,V) = \underbrace{(s_{ij}-\sum_{\ell=1}^{k} u_{i\ell}v_{j\ell})}_{e_{ij}:=s_{ij}-\hat{s}_{ij}}(-u_{i\ell}) + \lambda v_{j\ell},\quad j\in\{1,\cdots,d\},\ell\in\{1,\cdots,k\},$$

όπου $e_{ij}=s_{ij}-\hat{s}_{ij}$ το σφάλμα του recommender system δηλ. η απόκλιση μεταξύ της πρόβλεψης $\hat{s}_{ij}=\sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}$) απο το σύστημα και της πραγματικής αξιολόγησης s_{ij}.

Η βαθμίδα νπορεί και να γραφει σε πιο συμπαγή μορφή χρησιμοποιώντας τον πίνακα των σφαλμάτων

$$E = (e_{ij}), \quad E \in \mathbb{R}^{n \times k}$$

ως

$$\frac{\partial \mathcal{L}}{\partial U}(U, V) = -EV + \lambda U,$$
$$\frac{\partial \mathcal{L}}{\partial V}(U, V) = -E^T U + \lambda V,$$

Το σχήμα βαθμίδας παίρνει την μορφή

$$U_{k+1} = U_k - a\frac{\partial \mathcal{L}}{\partial U}(U_k, V_k) = (1 - a\lambda)U_k + aEV_k,$$
$$V_{k+1} = V_k - a\frac{\partial \mathcal{L}}{\partial V}(U_k, V_k) = (1 - a\lambda)V_k + aE^T U_k$$

Το παραπάνω σχήμα μπορεί και να υλοποιηθεί με την στρατηγική του stochastic gradient descent

Αν

- $u_i \in \mathbb{R}^k$ το διάνυσμα που περιέχει την i γραμμή του U
- $v_j \in \mathbb{R}^k$ το διάνυσμα που περιέχει την j γραμμή του V

το σχημα SGD τρέχει επάνω σε όλες τις συνιστώσες (με κάποιο τρόπο επιλογής) και ανανεώνει τις εκτιμήσεις συμφωνα με τον κανόνα

$$u_i \leftarrow (1 - a\lambda)u_i + e_{ij}v_j$$
$$u_j \leftarrow (1 - a\lambda)v_j + e_{ij}u_i$$

Το παραπάνω σχήμα μπορεί και να υλοποιηθεί με την στρατηγική του coordinate gradient descent

Στην περίπτωση αυτή θετουμε τις συνιστώσες της βαθμίδας ως προς κάθε $u_{i\ell}$ και κάθε $v_{j\ell}$ ίσες με το 0.

- $u_{i\ell}$: Η συνθήκη $\frac{\partial \mathcal{L}}{\partial u_{i\ell}}(U,V)=0$ αφού προσθέσουμε το $\sum_{j:(i,j)\in\mathcal{O}} u_{i\ell}v_{j\ell}^2$ και στα δυο μέλη, δίνει

$$(\lambda+\sum_{j:(i,j)\in\mathcal{O}} v_{j\ell}^2)u_{i\ell}=\sum_{j:(i,j)\in\mathcal{O}}(e_{ij}+u_{i\ell}v_{j\ell})v_{j\ell}\Longrightarrow u_{i\ell}=\frac{\sum_{j:(i,j)\in\mathcal{O}}(e_{ij}+u_{i\ell}v_{j\ell})v_{j\ell}}{(\lambda+\sum_{j:(i,j)\in\mathcal{O}} v_{j\ell}^2)}$$

- $v_{j\ell}$: Η συνθήκη $\frac{\partial \mathcal{L}}{\partial v_{j\ell}}(U,V)=0$ αφού προσθέσουμε το $\sum_{i:(i,j)\in\mathcal{O}} v_{j\ell}u_{i\ell}^2$ και στα δυο μέλη, δίνει

$$(\lambda+\sum_{i:(i,j)\in\mathcal{O}} u_{i\ell}^2)v_{j\ell}=\sum_{i:(i,j)\in\mathcal{O}}(e_{ij}+v_{j\ell}u_{i\ell})u_{i\ell}\Longrightarrow v_{j\ell}=\frac{\sum_{i:(i,j)\in\mathcal{O}}(e_{ij}+v_{j\ell}u_{i\ell})u_{i\ell}}{(\lambda+\sum_{i:(i,j)\in\mathcal{O}} u_{i\ell}^2)}$$

Οι παραπάνω σχέσεις οδηγούν σε ένα (επαναληπτικό) σχήμα σταθερού σημείου της μορφής

$$u_{i\ell}\leftarrow\frac{\sum_{j:(i,j)\in\mathcal{O}}(e_{ij}+u_{i\ell}v_{j\ell})v_{j\ell}}{(\lambda+\sum_{j:(i,j)\in\mathcal{O}} v_{j\ell}^2)}$$
$$v_{j\ell}\leftarrow\frac{\sum_{i:(i,j)\in\mathcal{O}}(e_{ij}+v_{j\ell}u_{i\ell})u_{i\ell}}{(\lambda+\sum_{i:(i,j)\in\mathcal{O}} u_{i\ell}^2)}$$

Αλγόριθμος (Recommender system : Coordinate descent)

- *Ξεκινάμε με κάποια αρχική επιλογή των πινάκων U, V.*
- *Μέχρις ότου κάποιο κριτήριο σύγκλισης ικανοποιηθεί επαναλαμβάνουμε για κάθε συνιστώσα κυκλικά*

$$u_{i\ell} \leftarrow \frac{\sum_{j:(i,j)\in\mathcal{O}}(e_{ij} + u_{i\ell}v_{j\ell})v_{j\ell}}{(\lambda + \sum_{j:(i,j)\in\mathcal{O}} v_{j\ell}^2)}, \; i \in \{1, \cdots, n\}, \ell \in \{1, \cdots, k\}$$

$$v_{j\ell} \leftarrow \frac{\sum_{i:(i,j)\in\mathcal{O}}(e_{ij} + v_{j\ell}u_{i\ell})u_{i\ell}}{(\lambda + \sum_{i:(i,j)\in\mathcal{O}} u_{i\ell}^2)}, \; j \in \{1, \cdots, d\}, \ell \in \{1, \cdots, k\}$$

Μία εναλλακτική προσέγγιση του ίδιου προβλήματος είναι τα εναλλασόμενα ελάχιστα τετράγωνα (alternating least squares)

Στην προσέγγιση αυτή ξεκινάμε απο την παρατήρηση ότι

- αν γνωρίζαμε τον πίνακα V το πρόβλημα μας γίνεται ένα πρόβλημα ελαχίστων τετραγώνων για τον πίνακα U,
- αν γνωρίζαμε τον πίνακα U το πρόβλημα μας γίνεται ένα πρόβλημα ελαχίστων τετραγώνων για τον πίνακα V

Τα επιμέρους προβλήματα ελαχίστων τετραγώνων γνωρίζουμε πως να τα διαχειριστούμε, και μάλιστα είναι διαχωρίσιμα.

Συνεπώς, μέχρι να επιτευχθεί κάποιο κριτήριο σύγκλισης επαναλαμβάνουμε τα παρακάτω βήματα

1. Κρατώντας το U σταθερό λύνουμε ως προς το V το πρόβλημα ελαχίστων τετραγώνων

$$\min_{V\in\mathbb{R}^{k\times d}} \frac{1}{2}\sum_{(i,j)\in\mathcal{O}} |s_{ij} - \sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{k} v_{j\ell}^2$$

$$= \min_{V\in\mathbb{R}^{k\times d}} \frac{1}{2}\sum_{i:(i,j)\in\mathcal{O}} \left(\sum_{j:(i,j)\in\mathcal{O}} |s_{ij} - \sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 \right) + \sum_{i=1} \left(\frac{\lambda}{2}\sum_{j=1}^{d} v \right),$$

   Καθορίζουμε τον πίνακα V ανά γραμμή $v_i = (v_{i1}, \cdots, v_{id})$.

2. Κρατώντας το V σταθερό λύνουμε ως προς το U το πρόβλημα ελαχίστων τετραγώνων

$$\min_{U\in\mathbb{R}^{n\times k}} \frac{1}{2}\sum_{(i,j)\in\mathcal{O}} |s_{ij} - \sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{k} u_{i\ell}^2$$

$$= \min_{U\in\mathbb{R}^{n\times k}} \frac{1}{2}\sum_{i:(i,j)\in\mathcal{O}} \left(\sum_{j:(i,j)\in\mathcal{O}} |s_{ij} - \sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 \right) + \sum_{i=1}^{n} \left(\frac{\lambda}{2}\sum_{j=1}^{d} u_I \right),$$

   Καθορίζουμε τον πίνακα U ανά στήλη $u_j = (v_{i1}, \cdots, v_{id})$, ;