Μέθοδοι βελτιστοποίησης

Newton Schemes

Α. Ν. Γιαννακόπουλος

Ο.Π.Α

Εαρινό Εξάμηνο 2026

Η βασικη ιδέα της μεθόδου Newton

Έστω $f : D \subset \mathbb{R}^d \to \mathbb{R}$ μια συναρτηση d-μεταβλητών η οποία ειναι επαρκώς λεια.

Η μέθοδος του Newton βασιζεται στον ανάπτυγμα Taylor

$$f(x) = f(x_0) + \langle Df(x_0), x - x_0) + \frac{1}{2}\langle x - x_0, D^2 f(x_0)(x - x_0)\rangle + O(\|x - x_0\|^3).$$

Συνεπώς κοντά στο σημείο $x_0 \in D$ προσεγγίζουμε την f με την τετραγωνική συνάρτηση

$$g(x) = f(x_0) + \langle Df(x_0), x - x_0) + \frac{1}{2}\langle x - x_0, D^2 f(x_0)(x - x_0)\rangle$$

και ελαχιστοποιούμε αυτή αντί της f!

Η συνθήκη πρώτης τάξης για το παραπάνω πρόβλημα είναι η

$$Dg(x) = Df(x_0) + D^2 f(x_0)(x - x_0) = 0,$$

η οποία μας δινει το ελαχιστο της g, $\hat{x}^*$ σαν λύση του συστηματος γραμμικών εξισώσεων

$$Df(x_0) + D^2 f(x_0)(\hat{x}^* - x_0) = 0,$$

ή ισοδύναμα (αν αντιστρέφεται ο πίνακας του Hess)

$$\hat{x}^* = x_0 - [D^2 f(x_0)]^{-1} Df(x_0).$$

Πόσο κοντά ειναι όμως το $\hat{x}^*$ στο πραγματικό ελάχιστο x^* της f;

Στην πραγματικότητα όμως η παραπάνω τετραγωνική προσέγγιση είναι τοπική.

Η ιδέα λοιπόν ειναι να πάρουμε μια ακολουθία προσεγγίσεων (x_k) για το x^* και μετά να εφαρμοσουμε την παραπάνω τετραγωνική προσέγγιση κάθε φορά για $x_0 = x_k$ για να καθορίσουμε την επόμενη προσέγγιση x_{k+1}.

Αυτο μας οδηγεί στο σχημα του Newton δηλ. στην επαναληπτική μέθοδο μέσω της ακολουθίας

$$x_{k+1} = x_k - [D^2 f(x_k)]^{-1} Df(x_k)$$

Αλγόριθμος (Newton)

- *Ξεκινάμε με μία αρχική προσέγγιση* x_1 *για το ελάχιστο και ένα περιθώριο σφάλματος* ϵ
- *Σε κάθε βήμα* k*, δεδομενου του* x_k
  - *Λύνουμε το γραμμικό συστημα*

$$D^2 f(x_k) p_k = Df(x_k)$$

  - *Θέτουμε*

$$x_{k+1} = x_k - p_k$$

- *Επαναλαμβάνουμε μέχρις ότου* $\|x_{k+1} - x - k\| < \epsilon$ *και επιστρέφουμε το τελευταίο* x_{k+1}.

Σχόλια

1. Ας πάρουμε το γενικό επαναληπτικό σχήμα

$$x_{k+1} = x_k + d_k,$$

όπου d_k ειναι μια επιλεγμενη διεύθυνση στον $\mathbb{R}^d$.
Τόσο η μέθοδοι βαθμίδας όσο και το σχημα του Νευτωνα μπορούν να θεωρηθούν σαν ειδικές περιπτώσεις του γενικού αυτού σχηματος
   - Η μεθοδοι βαθμίδας αν $d_k = -\rho_k Df(x_k))$ – κίνηση στην διεύθυνση της βαθμίδας
   - Η μέθοδος Newton αν $d_k = -[D^2 f(x_k)]^{-1} Df(x_k)$, δηλ. τροποποιειται η διευθυνση της βαθμιδας με τον γραμμικό μετασχηματισμό $d_k = ADf(x_k)$ όπου $A = -[D^2 f(x_k)]^{-1}$.
2. Η επιλογή αυτή οδηγει σε μεγαλύτερη ακριβεια και ταχυτερη συγκλιση στην μεθοδο, όμως σε αντάλλαγμα αυτό έχει την μεγαλύτερη υπολογιστικη πολυπλοκότητα καθώς χρειάζεται σε κάθε βημα η επίλυση ενός γραμμικού συστήματος $d \times d$.

Τροποποιήσεις του σχήματος του Newton

Μια συνηθισμένη τροποποίηση του σχηματος του Newton είναι να κρατησουμε μεν την διεύθυνση $d_k = -[D^2 f(x_k)]^{-1} Df(x_k)$ αλλά να τροποποιήσουμε το μέγεθος της μετακίνησης, δηλ.

$$x_{k+1} = x_k - \rho_k [D^2 f(x_k)]^{-1} Df(x_k), \quad \rho > 0$$

Η επιλογή του ρ_k μπορεί να γίνει ετσι ώστε να έχουμε την μεγαλύτερη δυνατή μείωση της τιμής της συνάρτησης στο νέο σημείο δηλαδή

$$\rho_k = \arg\min_{\alpha>0} f\left(x_k - \alpha [D^2 f(x_k)]^{-1} Df(x_k) \right).$$

Το πρόβλημα βελτιστοποίησης για την εύρεση του ρ_k είναι ένα μονοδιάστατο πρόβλημα βελτιστοποίησης το οποίο μπορεί να επιλυθεί σχετικα ευκολα (π.χ. με τις μεθόδους που ειδαμε στο ανάλογο προβλημα για τις μεθόδους βαθμίδας)

Προσέγγιση του $D^2 f(x)$

Μία άλλη παραλλαγή είναι η προσέγγιση του πίνακα του Hess $D^2 f(x)$ με ένα σχήμα διαφορών παρόμοιο με αυτό που χρησιμοποιούμε για τις πρώτες παραγώγους

$$\frac{\partial^2 f}{\partial x_j \partial x_i} \simeq \frac{1}{\epsilon}(Df_i(x + \epsilon e_j) - Df_i(x))$$

ή με το κεντρικό σχήμα διαφορών

$$\frac{\partial^2 f}{\partial x_j \partial x_i} \simeq \frac{1}{\epsilon}(Df_i(x + \epsilon e_j) - Df_i(x - \epsilon e_j)),$$

για $\epsilon > 0$ αρκετά μικρό.

Στο παραπάνω η προσέγγιση του $Df(x)$ γίνεται με ένα σχήμα διαφορών.

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

ή με ένα σχημα κεντρικών διαφορών

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

για αρκετά μικρό $\epsilon > 0$.

Η μέθοδος Levenberg-Marquardt

Η διεύθυνση

$$d_k = -[D^2 f(x_k)]^{-1} Df(x_k),$$

είναι μια καλή διευθυνση μεταβολής αν το x_k είναι κοντά στο ελάχιστο.

Αν όχι, αυτό δεν είναι απαραίτητο.

Η μέθοδος Levenberg - Marquardt αποτελεί παραλλαγή της μεθόδου του Newton στο ότι τροποποιεί αυτή την διεύθυνση ως εξής

$$d_k = -M_k^{-1} Df(x_k),$$
$$M_k = D^2 f(x_k) + \mu_k I, \ \mu_k \in \mathbb{R}$$

Η διαταραχή της $D^2 f$ απο τον όρο $\mu_k I$ έχει σημαντική επίδραση στην μέθοδο.

- Η επιτυχία της μεθόδου του Newton εξαρτάται σε μεγάλο βαθμό απο το αν ο πίνακας $D^2 f$ είναι θετικά ορισμένος (κυρτότητα).
- Αν ο $D^2 f(x_k)$ δεν είναι θετικά ορισμένος, μπορούμε με κατάλληλη επιλογή του μ_k να κάνουμε τον M_k θετικά ορισμένο οπότε η μέθοδος έχει καλύτερες ιδιότητες σύγκλισης.
- Μια επιλογή μπορεί να είναι $\mu_k > |\lambda_{min}(D^2 f(x_k))|$ όπου $\lambda_{min}(D^2 f(x_k))$ είναι η μικρότερη ιδιοτιμή του πίνακα του Hess
- Στην πράξη η επίλυση του προβλήματος ιδιοτιμών μπορεί να μην χρειάζεται (ή να είναι ακριβή σε υπολογιστικό χρονο) οπότε απλά επιλέγουμε το μ_k αυθαίρετα (π.χ. επαρκώς μεγάλο) τροποποιώντας παράλληλα το μεγεθος του βήματος ρ_k, και κατόπιν προσαρμόζοντας το μ_k (π.χ. υποδιπλασιάζοντας ή διπλασιάζοντας) ανάλογα με το αν $f(x_{k+1}) < f(x_k)$ ή όχι.

Levenberg-Marquardt

Για την ελαχιστοποίηση της f:

1. Επιλέγουμε αρχικα ένα $x_0 \in \mathbb{R}^n$ κάποια ακρίβεια $\epsilon > 0$ και κάποιο $\mu_0 > 0$ μεγάλο.
2. Στο βήμα k.
   - Υπολογίζουμε τον πίνακα $M_k = D^2 f(x_k) + \mu_k I$.
   - Υπολογίζουμε την διευθυνση $d_k = -M_k^{-1} Df(x_k)$
   - Υπολογίζουμε το $\rho_k = \arg\min_{\alpha>0} f(x_k + \alpha d_k)$
   - Θέτουμε $x_{k+1} = x_k + \rho_k d_k$.
   - Αν $f(x_{k+1}) < f(x_k)$ θέτουμε $\mu_{k+1} = \frac{\mu_k}{2}$ αλλιώς $\mu_{k+1} = 2\mu_k$.
   - Αν $\|x_{k+1} - x_k\| < \epsilon$ ή $\|Df(x_{k+1})\| < \epsilon$, επιστρέφουμε το x_{k+1} και σταματάμε.
   - Αν όχι επιστρέφουμε στο βήμα 2.

Εφαρμογή 1: Γραμμική παλινδρόμηση - Ridge regression

Επιστρέφουμε στο πρόβλημα της γραμμικής παλινδρόμησης, δηλ. της ευρεσης ενός γραμμικού μοντέλου το οποίο συνδέει παρατηρήσεις απο χαρακτηριστικα $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})$ με τις αποκρίσεις Y_i για $i = 1, \cdots, n$ δεδομένα.

Υπενθυμίζουμε οτι το μοντέλο το οποίο ψάχνουμε είναι της μορφής

$$Y_i \simeq w_1 X_{i1} + \cdots + w_d X_{id}, \quad i = 1, \cdots, n,$$

και αυτό θα το επιλέξουμε σαν λύση του προβλήματος βελτιστοποίησης

$$\min_w \mathcal{L}(w), \quad \mathcal{L}(w) = \frac{1}{2}\|\mathbf{D}w - Y\|^2 + \frac{\lambda}{2}\|w\|^2,$$

$$\mathbf{D} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \in \mathbb{R}^{n\times d}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d\times 1}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n\times 1}.$$

Έχουμε ήδη υπολογίσει την βαθμίδα της $\mathcal{L}$,

$$D\mathcal{L}(w) = \mathbf{D}^T(\mathbf{D}w - Y) + \lambda w = \mathbf{D}^T\mathcal{E} + \lambda w,$$

και απο αυτό μπορούμε να υπολογίσουμε τον πίνακα του Hess στην μορφή

$$D^2\mathcal{L}(w) = \mathbf{D}^T\mathbf{D} + \lambda I$$

Η μέθοδος του Newton παίρνει την μορφή

$$w_{k+1} = w_k - (D^2\mathcal{L}(w))^{-1}D\mathcal{L}(w)$$

δηλ.

$$\begin{aligned} w_{k+1} &= w_k - (\mathbf{D}^T\mathbf{D} + \lambda I)^{-1}(\mathbf{D}^T(\mathbf{D}w_k - Y) + \lambda w_k) \\ &= w_k - (\mathbf{D}^T\mathbf{D} + \lambda I)^{-1}(\mathbf{D}^T\mathcal{E}_k + \lambda w_k) \end{aligned}$$

Παρατηρείστε οτι το παραπάνω σχήμα μπορεί να γραφεί ως

$$\begin{aligned} w_{k+1} &= w_k - (\mathbf{D}^T\mathbf{D} + \lambda I)^{-1}((\mathbf{D}^T\mathbf{D} + \lambda I)w_k - \mathbf{D}^T Y) \\ &= w_k - w_k + (\mathbf{D}^T\mathbf{D} + \lambda I)^{-1}\mathbf{D}^T Y \\ &= (\mathbf{D}^T\mathbf{D} + \lambda I)^{-1}\mathbf{D}^T Y \end{aligned}$$

δηλ. η μεθοδος του Newton τερματίζει σε μία επανάληψη και μας δίνει κατευθείαν την αναλυτική λύση για το ridge regression

Εφαρμογή 2: Το μοντέλο Garch(1,1)

Ένα πολύ συνηθισμενο υπόδειγμα για τις λογαριθμικές αποδόσεις των χρηματοοικονομικών αγορών είναι το υπόδειγμα Garch(1,1) σύμφωνα με το οποίο

$$x(t) = \ln(r(t))\,|_{r(1),\cdots,r(t-1)} \sim N(\mu, \sigma^2(t)),$$
$$\sigma^2(t) = \alpha_0 + \alpha_1 x^2(t-1) + \beta_1 \sigma^2(t-1)$$

Τα δεδομένα μας τυπικά είναι τα

$$x(0) = X_0, x(1) = X_1, \cdots, x(n) = X_n,$$

ενώ αυτό που χρειάζεται να προσδιορίσουμε ειναι οι μεταβλητότητες

$$\sigma(0), \sigma(1), \cdots, \sigma(n).$$

Η πιθανοφάνεια του δείγματος $X_0, X_1, \cdots, X_n$ είναι

$$L(\theta; X) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi\sigma^2(i)}} \exp\left(-\frac{X_i^2}{2\sigma^2(i)}\right),$$

$$\sigma^2(t) = \alpha_0 + \alpha_1 X^2(t-1) + \beta_1 \sigma^2(t-1), \quad t = 1, 2, \cdots, n$$

και η εκτίμηση παραμέτρων καταλήγει στο πρόβλημα ελαχιστοποίησης

$$\min_{\theta \in \mathbb{R}^4} \sum_{i=0}^{n} \left[\frac{X_i^2}{2\sigma^2(i)} - \frac{1}{2}\ln(\sigma^2(i))\right],$$

$$\sigma^2(t) = \alpha_0 + \alpha_1 X^2(t-1) + \beta_1 \sigma^2(t-1), \quad t = 1, 2, \cdots, n$$

υπο τους περιορισμούς

$$\sigma(0) > 0, \ \alpha_0 > 0, \alpha_1 > 0, \beta_1 > 0, \ \alpha_1 + \beta_1 < 1.$$

Εφαρμογή 3: Λογιστική παλινδρόμηση

Το πρόβλημα της γραμμικής παλινδρόμησης, είναι αυτό της ευρεσης ενός γραμμικού μοντέλου το οποίο συνδέει παρατηρήσεις απο χαρακτηριστικα $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})$ με τις διακριτές αποκρίσεις $Y_i \in \{0, 1\}$ για $i = 1, \cdots, n$ δεδομένα.

Σύμφωνα με το μοντέλο αυτό αν ορίσουμε

$$p_i = P(Y = 1 \mid \mathbf{X} = \mathbf{X}_i),$$

τότε υπάρχει κάποιο διάνυσμα $w = (w_1, \cdots, w_d)$ τέτοιο ώστε

$$\ln\left(\frac{p_i}{1-p_i}\right) = \langle X_i, w\rangle \iff p_i = \frac{1}{1+\exp(-\langle X_i, w\rangle)}, \quad i = 1, \cdots, n$$

Το μοντέλο αυτό κατατάσει αντικείμενα i με χαρακτηριστικά $\mathbf{X}_i$ σε δύο κατηγορίες, την 0 ή την 1 με κάποια πιθανότητα $1 - p_i$ ή p_i αντίστοιχα.

Βρισκει πολλες εφαρμογές στην

- Διαχείριση κινδυνου: Κατάταξη δανειοληπτών στην κατηγορία του αξιόπιστου ή όχι ανάλογα με διάφορα χαρακτηριστικα π.χ. εισόδημα, ηλικία κλπ.
- Ιατρική: Κατάταξη ασθενών σε κατηγορίες ανάλογα με χαρακτηριστικά π.χ. υπερτασικοί ή όχι ανάλογα με βάρος, ηλικία, διατροφικές συνήθειες κλπ.
- Βιοστατιστική, ψυχομετρια, γενετική, κοινωνικές επιστήμες κ.α.

Οριζουμε τα

$$\mathbf{D} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \in \mathbb{R}^{n\times d}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d\times 1}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n\times 1}, \quad p = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \in \mathbb{R}^{n\times 1}.$$

Το μοντέλο λοιπόν παίρνει την συμπαγή μορφη

$$\ln\left(\frac{p}{1-p}\right) = \langle X, w\rangle \iff p = \frac{1}{1+\exp(-\langle X, w\rangle)}$$

Η πιθανότητα παρατήρησης του δεδομένου i, δηλ. του ζεύγους $(\mathbf{X}_i, Y_i)$ είναι

$$p_i^{Y_i}(1-p_i)^{1-Y_i}, \quad p_i = \frac{1}{1+\exp(-\langle X_i, w\rangle)}, \quad i = 1, \cdots, n$$

Η πιθανοφάνεια των δεδομένων $(\mathbf{X}_i, Y_i)$ ειναι

$$L(w) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \quad p_i = \frac{1}{1+\exp(-\langle X_i, w\rangle)},$$

οπότε η αρνητική λογαριθμική πιθανοφάνεια γίνεται

$$\mathcal{L}(w) = -\ln L(w) = -\sum_{i=1}^{n} Y_i \ln p_i + (1-Y_i)\ln(1-p_i)), \quad p_i = \frac{1}{1+\exp(-\langle X_i, w\rangle)}$$

Το κατάλληλο μοντέλο θα επιλεγεί απο την λύση του προβλήματος βελτιστοποίησης

$$\min_w \mathcal{L}(w)$$

Η βελτιστοποίηση μπορεί να γίνει χρησιμοποιώντας την μέθοδο του Newton

Αναφορές

S. Boyd and Vanderberghe, Convex optimization, Cambridge University Press

J. Nocedal, Numerical Optimization, Springer.

D. Kravvaritis and A. N. Yannacopoulos, Variational Methods in Nonlinear Analysis with applications in Optimization and PDEs. De Gruyter, Chapters 4 and 5.

S. K. Mishra and B. Ram, Introduction to unconstrained optimization with R, Springer

C C Aggarwal, Linear Algebra and Optimization for Machine learning, Springer