Θεωρία και Δοκιμή SVM **Αθανάσιος Μητσέλος**
ΣΗΜΜΥ Ημερομηνία Ανάθεσης: 06/12/16
ΕΜΠ Τρέχουσα Ημερομηνία: 16/07/17

1 Εισαγωγή στη γραμμική ταξινόμηση

Στα προβλήματα γραμμικού διαχωρισμού, η λογιστική παλινδρόμηση (logistic regression) και η γραμμική μηχανή υποστήριξης διανυσμάτων (linear SVM) είναι τα δύο πιο ευρέως διαδεδομένα μοντέλα. Μπορούμε να εκτιμήσουμε την παράμετρο του μοντέλου, αναπαριστώντας την ως $\mathbf{w}$, λύνοντας το άνευ περιορισμών πρόβλημα βελτιστοποίησης

$$min_{\mathbf{w}} f(\mathbf{w}). \quad (1)$$

Οι υπάρχοντες άνευ περιορισμών μέθοδοι ελαχιστοποίησης μπορούν να εφαρμοστούν επιτυχώς, παρόλο που απαιτούνται κάποιες διορθώσεις για να αντιμετωπιστούν προβλήματα με μεγάλο όγκο δεδομένα. Γενικώς, αυτές οι μέθοδοι παράγουν μια ακολουθία $\{\mathbf{w}^k\}_{k=0}^{\infty}$, η οποία συγκλίνει στην βέλτιστη λύση. Στην k επανάληψη, ευρίσκεται ο φθίνουσας κατεύθυνσης παράγοντας s^k, από το $\mathbf{w}^k$ της k επανάληψης. Εν συνεχεία αποφασίζεται το μέγεθος του βήματος $a_k > 0$ με απώτερο σκοπό την εύρεση του παράγοντα $\mathbf{w}^{k+1}$ της επόμενης επανάληψης:

$$\mathbf{w}^{k+1} = \mathbf{w}^k + a_k s^k. \quad (2)$$

Η εύρεση των σημαντικών αυτών παραμέτρων, η κατεύθυνση s^k και η επιλογή του μεγέθους του βήματος a_k, έχουν ήδη μελετηθεί εκτενώς στη βιβλιογραφία. Για παράδειγμα, η μέθοδος της πλέον απότομης μετάβασης (gradient descent) και η μέθοδος του Newton είναι ευρέως χρησιμοποιούμενες τεχνικές για την εύρεση του s^k. Για την απόφαση του μεγέθους a_k, έχουμε τις μεθόδους της αναζήτησης γραμμής (line search) και τη μέθοδο του εύρους εμπιστοσύνης (trust region).

2 Η μέθοδος Newton και η επιλογή βήματος

Δεδομένου ενός σετ εκπαίδευσης $(\mathbf{x}_i, y_i)$, $i = 1, ..., l$, όπου $\mathbf{x}_i \in \Re^n$ είναι ένα χαρακτηριστικό διάνυσμα και $y_i = \pm 1$ είναι οι ετικέτες, ένας γραμμικός ταξινομητής βρίσκει ένα διάνυσμα βαρών $\mathbf{w} \in \Re^n$ επιλύοντας το ακόλουθο πρόβλημα:

$$min_{\mathbf{w}} f(\mathbf{w}) \equiv \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi(y_i \mathbf{w}^T x_i), \quad (3)$$

όπου $\mathbf{w}^T\mathbf{w}/2$ είναι ο όρος κανονικοποίησης, $\xi(y_i\mathbf{w}^T x_i)$ είναι η συνάρτηση απωλειών (loss function) και $C > 0$ είναι η παράμετρος κανονικοποίησης. Θεωρούμε τις συναρτήσεις στη λογιστική παλινδρόμηση και στην L2:

$$\xi_{LR}(y\mathbf{w}^T\mathbf{x}) = log(1 + exp(-y\mathbf{w}^T\mathbf{x})) \quad (4)$$
$$\xi_{L2}(y\mathbf{w}^T\mathbf{x}) = (max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2 \quad (5)$$

Η μέθοδος Newton επιλύει το πρόβλημα βελτιστοποίησης εφαρμόζοντας επαναληπτικά κανόνες ανανέωσης όπως στη σχέση (2). Σε κάθε επανάληψη, αποκτάται μια Newton κατεύθυνση s^k ελαχιστοποιώντας την τετραγωνική εκτίμηση

$$f(\mathbf{w}^k + s) - f(\mathbf{w}^k) \approx q_k(s) \equiv \nabla f(\mathbf{w}^k)^T s + \tfrac{1}{2} s^T \nabla^2 f(\mathbf{w}^k) s \quad (6)$$

όπου $\nabla f(\mathbf{w}^k)$ και $\nabla^2 f(\mathbf{w}^k)$ είναι η κλίση και ο Hessian, αντιστοίχως. Εδώ πρέπει να επισημανθεί πως η συνάρτηση L2 δεν είναι διπλά διαφορίσιμη, αλλά μπορούμε να θεωρήσουμε το γενικευμένο Hessian[1]. Με τον όρο κανονικοποίησης $\mathbf{w}^T\mathbf{w}/2$ και την καμπυλότητα των αντικειμενικών συναρτήσεων (4)-(5), ο Hessian πίνακας είναι θετικά ορισμένος, έτσι ώστε να καθορίζεται το s^k λύνοντας το ακόλουθο γραμμικό σύστημα

$$\nabla^2 f(\mathbf{w}^k)s = -\nabla f(\mathbf{w}^k). \quad (7)$$

Σημειώνεται πως η κλίση και ο Hessian του f(w) είναι αντιστοίχως

$$\nabla f(\mathbf{w}) = \mathbf{w} + C\sum_{i=1}^{l} \xi'(y_i\mathbf{w}^T\mathbf{x}_i)y_i x_i, \nabla^2 f(\mathbf{w}) = I + CX^T DX, \quad (8)$$

όπου D είναι ο διαγώνιος πίνακας με

$$D_{ii} = \xi''(y_i\mathbf{w}^T\mathbf{x}_i), \quad (9)$$

είναι ο ταυτοτικός πίνακας και $X = [x_1, ..., x_l]^T$ είναι ο πίνακας δεδομένων. Η ακριβής λύση της (7) είναι πολύ ακριβή υπολογιστικά για μεγάλα όγκο δεδομένων, έτσι χρησιμοποιείται ευρέως η παρούσα απλουστευμένη μέθοδος Newton για τη λύση της (7). Τυπικά χρησιμοποιείται μια επαναληπτική μέθοδος όπως η μέθοδος των συζυγή παραγώγων (conjugate gradient)[2, 3]. Η μέθοδος CG περιλαμβάνει μια ακολουθία γινομένου Hessian διανυσμάτων, αλλά για μεγάλο αριθμό χαρακτηριστικών, ο όρος $\nabla^2 f(\mathbf{w}^k) \in \Re^{nxn}$ είναι πολύ μεγάλος για να αποθηκευτεί. Παλαιότερες προσεγγίσεις [4, 5] έχουν δείξει πως η ιδιαίτερη δομή της (8) επιτρέπει τον υπολογισμό γινομένων των Hessian διανυσμάτων χωρίς απόλυτο ορισμό του Hessian πίνακα:

$$\nabla^2 f(\mathbf{w})s = (I + CX^T DX)s = s + CX^T(D(Xs)). \quad (10)$$

Συνημμένα

Αναφορές

[1] O. L. Mangasarian. *A finite method for classification*. Optim. Methods Soft., 17(5):913-929,2002

[2] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[3] M. R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*. Journal of Research of the National Bureau of Standards, 49:409–436, 1952.

[4] S. S. Keerthi and D. DeCoste. *A modified finite Newton method for fast solution of large scale linear SVMs*. JMLR, 6:341–361, 2005.

[5] C.-J. Lin, R. C. Weng, and S. S. Keerthi. *Trust region Newton method for large-scale logistic regression*. JMLR, 9:627–650, 2008.