Αναγνώριση Προτύπων

ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ 2

Στοιχεία φοιτητή
Ονοματεπώνυμο: Κωνσταντίνος Τσόπελας
Αριθμός Μητρώου: 03400198

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
ΔΠΜΣ ΕΔΕΜΜ
18 Ιανουαρίου 2023

Άσκηση 2.1

1. Μετονομάζω $f_L(x) = f(x) = f^L(f^{L-1}(f^{L-2}(\dots(f^1(x))\dots)))$, και πάω για επαγωγή.

Βλέπουμε έυκολα ότι:

$$f_L(x) = f^L(f_{L-1}(x))$$

και, με αυτό σαν εργαλείο, έχω:

- $f_1(x) = f^1(x) = \sigma(W_1 x + b_1) = W_1 x + b_1$, αφού έχουμε πάρει σ την ταυτοτική.
- Αν $f_{L-1}(x) = W^{L-1}x + b^{L-1}$ (γραμμική), τότε (με σ ταυτοτική):

  $$f_L(x) = W_L(W^{L-1}x + b^{L-1}) + b_L = (W_L W^{L-1})x + (W_L b^{L-1} + b_L)$$

  που είναι και πάλι μια γραμμική.

Άρα, κάθε συνάρτηση της μορφής που μας δίνεται είναι τελικά ισοδύναμη με μία γραμμική, δηλαδή ένα δίκτυο ενός επιπέδου.

2. Με απλή αντικατάσταση των δεδομένων, λαμβάνουμε:

$$\begin{aligned}\tilde{f}(x_1, x_2) = \begin{bmatrix}\mu & \mu & -\mu & -\mu\end{bmatrix} \sigma \left(\lambda \begin{bmatrix} x1 + x2 \\ -x_1 - x_2 \\ x_1 - x_2 \\ -x_1 + x_2 \end{bmatrix} \right) \\ = \mu\sigma(\lambda(x_1 + x_2)) \\ + \mu\sigma(\lambda(-x_1 - x_2)) \\ - \mu\sigma(\lambda(x_1 - x_2)) \\ - \mu\sigma(\lambda(-x_1 + x_2))\end{aligned}$$

3. Προσέγγιση Taylor δεύτερης τάξης γύρω από το 0 της σ:

$$\sigma(x) = \sigma(0) + \dot{\sigma}(0)x + \frac{\ddot{\sigma}(0)}{2}x^2 + O(x^3) \tag{1}$$

όπου ο συμβολισμός $g(x) = O(f(x))$ σημαίνει κάποια απροσδιόριστη συνάρτηση η οποία κυριαρχείται από το x^3 σε κάποια περιοχή του 0, δηλαδή για x αρκετά κοντά στο 0, ισχύει $|g(x)| \leq M|f(x)|$ για κάποια θετική σταθερά M.

Κατόπιν, αντικαθιστώντας την παραπάνω έκφραση στο αποτέλεσμα του προηγούμενου ερωτήματος, λαμβάνουμε:

$$\begin{aligned}\tilde{f}(x) = \mu[\dot{\sigma}(0)\lambda(x_1 + x_2) + \frac{\ddot{\sigma}(0)}{2}\lambda^2(x_1 + x_2)^2 + O(\lambda^3(x_1 + x_2)^3) \\ + \dot{\sigma}(0)\lambda(-x_1 - x_2) + \frac{\ddot{\sigma}(0)}{2}\lambda^2(-x_1 - x_2)^2 + O(\lambda^3(-x_1 - x_2)^3) \\ - \dot{\sigma}(0)\lambda(x_1 - x_2) - \frac{\ddot{\sigma}(0)}{2}\lambda^2(x_1 - x_2)^2 - O(\lambda^3(x_1 - x_2)^3) \\ - \dot{\sigma}(0)\lambda(-x_1 + x_2) - \frac{\ddot{\sigma}(0)}{2}\lambda^2(-x_1 + x_2)^2 - O(\lambda^3(-x_1 + x_2)^3)]\end{aligned}$$

Οι πρωτοβάθμιοι όροι αλληλοαναιρούνται. Οι δευτεροβάθμιοι ομαδοποιούνται λόγω του τετραγώνου, και τους τριτοβάθμιους μπορούμε να τους συνοψίσουμε απλώς σαν $O(\lambda^3)$, καθώς θα πάρουμε όριο καθώς το λ πάει στο 0, με τα x_1, x_2 σταθερά, και, όπως θα δούμε στη συνέχεια, ο όρος αυτός ούτως ή άλλως θα μηδενιστεί.

Κατόπιν τούτου, το παραπάνω γίνεται:

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{4\lambda^2\ddot{\sigma}(0)}\left[\frac{\ddot{\sigma}(0)}{2}\lambda^2(2(x_1+x_2)^2 - 2(x_1-x_2)^2) + O(\lambda^3)\right] \\ &= \frac{1}{4}[(x_1+x_2)^2 - (x_1-x_2)^2] + \frac{O(\lambda^3)}{4\lambda^2\ddot{\sigma}(0)}\end{aligned}$$

Παίρνοντας το όριο καθώς το λ πάει στο 0, ο δεύτερος όρος θα μηδενιστεί αφού $|O(\lambda^3)| \leq M|\lambda^3| \Rightarrow |\frac{O(\lambda^3)}{\lambda^2}| \leq M|\lambda|$. Οπότε, καταλήγουμε:

$$\lim_{\lambda\to 0}\tilde{f}(x) = \frac{1}{4}(x_1^2 + x_2^2 + 2x_1x_2 - x_1^2 - x_2^2 + 2x_1x_2) = x_1x_2 = f(x)$$

Άσκηση 2.2

1. Υποθέτουμε (γιατί τι άλλο να κάνουμε) ότι η εκφώνηση εννοεί:

- $L_t = L(o_t, yr_t)$, όπου yr_t η πραγματική έξοδος, τα labels
- $y_t = Vh_t$ η έξοδος του output layer, πριν την τελική ενεργοποίηση.
- o_t η πρόβλεψη του δικτύου (έξοδος μετά την τελική ενεργοποίηση, αυτή που έχουμε συνηθίσει να είναι του τύπου softmax), αυτή δηλαδή που συγκρίνεται τελικά με την πραγματική έξοδο yr_t.

Με αυτές τις παραδοχές, εφαρμόζοντας τον κανόνα της αλυσίδας έχουμε:

$$\frac{\partial L_t}{\partial V} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial y_t}\frac{\partial y_t}{\partial V} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial y_t}\frac{\partial (Vh_t)}{\partial V} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial y_t}h_t$$

Να σημειώσουμε εδώ ότι ο Jacobian $\frac{\partial o_t}{\partial y_t}$ μάλλον θα μπορούσε να γραφτεί και $\text{diag}\{f'(y_t) = f'(Vh_t)\}$, αλλά αφού η εκφώνηση μας ζητάει να συμμετέχουν μόνο οι συγκριμένες μεταβλητές, το αφήνουμε ως έχει.

2. Συμβολίζουμε με $a_t = Wh_{t-1} + Ux_t$. Τότε, με παρόμοιο τρόπο, και σύμφωνα με την εκφώνηση, θα έχουμε:

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial h_t}\frac{\partial h_t}{\partial a_t}\frac{\partial a_t}{\partial V} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial h_t}\,\text{diag}\{f'(Wh_{t-1} + Ux_t\}h_{t-1}$$

όπου η μόνη ``ατασθαλεία'' σε σχέση με την εκφώνηση είναι ο όρος Ux_t, αλλά προφανώς αυτός πρέπει να συμμετέχει, και υποθέσαμε ότι δεν προσμετράται από την εκφώνηση, καθώς αφορά την είσοδο και όχι τις εσωτερικές μεταβλητές του μοντέλου.

Κατά τα άλλα, και πάλι, ο όρος $\frac{\partial o_t}{\partial h_t}$ όπως είδαμε και πριν μπορεί να αναπτυχθεί και περαιτέρω, αλλά αφού η εκφώνηση απαιτεί να μην συμμετέχει ο πίνακας V, για παράδειγμα, το αφήσαμε ως έχει.

3. Θεωρώντας $f(x) = x$, θα έχουμε βέβαια ότι $f'(x) = 1$. Κάπως έτσι, στην παραπάνω σχέση που βγάλαμε για την παράγωγο ως προς W, θα εξαφανιστούν οι όροι που έχουν να κάνουν με την f'. Με κάποιες επιπλέον πράξεις παρόμοιες με του πρώτου ερωτήματος, θα έχουμε:

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial y_t}\frac{\partial (y_t = Vh_t)}{\partial h_t} I h_{t-1} = \frac{\partial L_t}{\partial o_t}(\text{diag}\{f'(y_t)\} = I) V h_{t-1} = \frac{\partial L_t}{\partial o_t} V h_{t-1}$$

Έχοντας, τώρα, ``πετάξει'' την f, μπορούμε να γράψουμε για το h_t:

$$h_t = W h_{t-1} + U x_t = W^2 h_{t-2} + W U x_{t-1} + U x_t = \cdots = W^t h_0 + \sum_{\tau=0}^{t-1} W^\tau U x_{t-\tau}$$

Γνωρίζουμε, όμως, από τη γραμμική άλγεβρα ότι η ποσότητα W^t είτε συγκλίνει προς το μηδενικό πίνακα είτε αποκλίνει και απειρίζεται, αναλόγως αν η φασματική ακτίνα του W (το μέγιστο μέτρο ιδιοτιμής του) είναι μικρότερο ή μεγαλύτερο από το 1, αντίστοιχα.

Από τη στιγμή, λοιπόν, που αυτή η ποσότητα συμμετέχει, τελικά, στην παράγωγο $\frac{\partial L_t}{\partial W}$, καταλαβαίνουμε ότι, με κατάλληλα μικρές ή αργά μεταβαλλόμενες τιμές για τα x_t και αντίστοιχα για την επιλεγμένη συνάρτηση κόστους, μπορεί κάλλιστα η παράγωγος αυτή να τείνει, για μεγάλα t, προς το 0 ή το άπειρο.

Άσκηση 2.3

1. Θυμόμαστε από τη θεωρία ότι, όπως και στον απλό υπολογισμό του SVM, έτσι και σε αυτόν που έχουμε εδώ, με τα slack variables, η τελική έκφραση για τα βάρη (πλην του bias) συναρτήσει των πολλαπλασιαστών Lagrange και των διανυσμάτων εισόδου (στην πραγματικότητα μόνο των support vectors, αλλά δεν έχει διαφορά) είναι:

$$\vec{w} = \sum_{i=1}^{n} a_i y_i \vec{x_i}$$

Παρατηρούμε, τώρα, ότι όλα τα σημεία που μας δίνονται είναι πάνω στην ευθεία $x^1 = x^2$, οπότε μπορούν να γραφτούν ως $\vec{x_i} = c_i < 1, 1 >$. Κατόπιν τούτου, είναι πολύ εύκολο να δει κανείς ότι και για τα βάρη θα ισχύει το ίδιο, αφού προκύπτουν ως γραμμικός συνδυασμός των $\vec{x_i}$:

$$\vec{w} = \sum_{i=1}^{n} a_i y_i c_i < 1, 1 >= (\sum_{i=1}^{n} a_i y_i c_i) < 1, 1 >= B < 1, 1 >=< B, B >$$

όπου B κάποια συνολική σταθερά.

2. Θυμόμαστε, τώρα, ότι, όπως και στον απλό υπολογισμό του SVM, έτσι και εδώ οι συντελεστές Lagrange που θα χρησιμοποιηθούν για τον υπολογισμό των βαρών θα προκύψουν από την μεγιστοποίηση της ελαχιστοποιημένης ως προς $\vec{w}, b, \vec{\xi}$ λανγκρανζιανής, η οποία θυμόμαστε βέβαια ότι προκύπτει ίση με:

$$\mathcal{L}(\vec{a}; \vec{w}^*, b^*, \vec{\xi}^*) = -\frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} a_i$$

Σημ: οι συντελεστές Lagrange β_i δεν συμμετέχουν στην παράσταση, γιατί έτσι κι αλλιώς είναι πλήρως προσδιορισμένοι, άρρηκτα συνδεδμένοι με τους a_i, αφού μία από τις συνθήκες στασιμότητας (συγκεκριμένα, για τα ξ) μας δίνει ότι $a_i + b_i = C$.

Για τη μεγιστοποίηση, οπότε, αυτής της ποσότητας, μπορεί να πάρει κανείς τις παραγώγους ως προς τα a_i, οπότε και εύκολα προκύπτει το ακόλουθο τελικό γραμμικό σύστημα για τους πολλαπλασιαστές Lagrange:

$$\sum_{j=1}^{n}(y_i y_j x_i^T x_j)a_j = 1 \text{ for all } i = 1, \ldots, n$$

Για την μη γραμμική περίπτωση με το kernel, το μόνο που αλλάζει σε όλα αυτά είναι το εσωτερικό γινόμενο:

$$\sum_{j=1}^{n}(y_i y_j K(x_i, x_j))a_j = 1 \text{ for all } i = 1, \ldots, n$$

Με βάση αυτή την εξίσωση, λοιπόν, που είναι αναγκαία ώστε να μας προκύψουν πράγματι βέλτιστα w, θα δείξουμε το ζητούμενο.

Συγκεκριμένα, έστω ότι κάποιο $a_i = 0$. Τότε, όπως ξέρουμε ισχύει:

$$a_i = 0, a_j \leq C \forall j \neq i \Rightarrow \sum_i a_i \leq C(n-1) < 1 \text{ από το δεδομένο της εκφώνησης}$$

Επιπλέον:

$$\begin{aligned} |y_i y_j K(x_i, x_j)| &= |K(x_i, x_j)|, \text{ αφού τα } y_i \text{ είναι 1 ή -1} \\ &< 1 \text{ από τα δεδομένα της άσκησης} \end{aligned}$$

Κατόπιν τούτου, και αφού τα a_i είναι μη αρνητικά, θα ισχύει:

$$\begin{aligned} & y_i y_j K(x_i, x_j) \leq |y_i y_j K(x_i, x_j)| < 1 \\ \Rightarrow & (y_i y_j K(x_i, x_j))a_i < a_i \\ \Rightarrow & \sum_j (y_i y_j K(x_i, x_j))a_i < \sum_j a_i < 1 \end{aligned}$$

Άρα, η παραπάνω αναγκαία συνθήκη δεν μπορεί να ικανοποιείται, και καταλήξαμε σε άτοπο. Συνεπώς, δεν μπορεί κανένα $a_i = 0$!

3. Όντας στις 2 διαστάσεις, μπορούμε απλά να γράψουμε:

$$\begin{aligned} K(u, v) &= u^T v + 4(u^T v)^2 \\ &= u_1 v_1 + u_2 v_2 + 4u_1^2 v_1^2 + 4u_2^2 v_2^2 + 8u_1 v_1 u_2 v_2 \\ &= (u_1, u_2, 2u_1^2, 2u_2^2, 2\sqrt{2}u_1 u_2) \cdot (v_1, v_2, 2v_1^2, 2v_2^2, 2\sqrt{2}v_1 v_2) \end{aligned}$$

Δηλαδή η ζητούμενη συνάρτηση φ είναι η:

$$\phi(u = (u_1, u_2)) = (u_1, u_2, 2u_1^2, 2u_2^2, 2\sqrt{2}u_1 u_2)$$

Άσκηση 2.4

Έχουμε:

$$\begin{aligned}\mu_i &= E[y|\omega_i] = E[w^T x|\omega_i] = w^T\vec{\mu_i}, i = 1, 2\\ \sigma_i^2 &= E[(y-\mu_i)^2|\omega_i] = E[(w^T x - w^T\vec{\mu_i})^2|\omega_i] = E[(w^T(x-\vec{\mu_i}))^2|\omega_i]\\ &= E[w^T(x-\vec{\mu_i})(x-\vec{\mu_i})^T w|\omega_i] = w^T\Sigma_i w\end{aligned}$$

όπου κατά βάση αυτό που χρησιμοποιήσαμε είναι η γραμμικότητα της μέσης τιμής.

Κατόπιν τούτου, μπορούμε να ξαναγράψουμε το προς μεγιστοποίηση κριτήριο ως:

$$J_1(w) = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} = \frac{(w^T(\vec{\mu_1}-\vec{\mu_2}))^2}{w^T(\Sigma_1+\Sigma_2)w} = \frac{w^T(\vec{\mu_1}-\vec{\mu_2})(\vec{\mu_1}-\vec{\mu_2})^T w}{w^T(\Sigma_1+\Sigma_2)w}$$

το οποίο είναι γραμμένο μόνο σε όρους w συν τις εξ αρχής γνωστές σταθερές. Μπορούμε, λοιπόν, για ευκολία σε αυτό το σημείο να κάνουμε λίγο abstraction, και να πούμε ότι το πρόβλημα που μας μένει να επιλύσουμε είναι η μεγιστοποίηση της ποσότητας:

$$\frac{x^T a a^T x}{x^T B x} \text{ over } x$$

όπου γνωρίζουμε, επιπλέον, ότι ο πίνακας B είναι συμμετρικός και θετικά ορισμένος (υποθέτουμε non-degenerate gaussian σε όλες τις κατευθύνσεις).

Αυτό το κλάσμα, όμως, είναι το γνωστό generalized Rayleigh quotient (την οποία αντιστοιχία καταφέραμε να βρούμε μετά από **υπέρμετρα** μεγάλο παίδεμα και ψάξιμο - κυρίως για το σωστό search query), όπως αυτό παρουσιάζεται, για παράδειγμα, στην αντίστοιχη σελίδα της Wikipedia.

Το απλό Rayleigh quotient είναι το ακόλουθο:

$$R(M, x) = \frac{x^T M x}{x^T x}$$

Αυτό δεν αντιστοιχίζεται άμεσα στη δική μας περίπτωση, που είναι το ακόλουθο generalized Rayleigh quotient:

$$R(aa^T, B, x) = \frac{x^T a a^T x}{x^T B x} \text{ over } x$$

Είναι, όμως, γνωστό και μπορεί εύκολα να δει κανείς ότι το γενικευμένο μπορεί να μετασχηματιστεί σε ένα αντίστοιχο απλό μέσω του μετασχηματισμού $x \mapsto C^T x$, όπου $B = CC^T$ είναι η (μονοσήμαντη) παραγοντοποίηση Cholesky του **θετικά ορισμένου** πίνακα B. Πράγματι:

$$\begin{aligned}R(aa^T, B, x) &= \frac{x^T a a^T x}{x^T B x} = \frac{x^T C C^{-1} a a^T (C^T)^{-1} C^T x}{x^T C C^T x}\\ &= \frac{(C^T x)^T C^{-1} a a^T (C^T)^{-1} (C^T x)}{(C^T x)^T (C^T x)} = R(C^{-1} a a^T (C^T)^{-1}, C^T x)\end{aligned}$$

Τώρα, το αποτέλεσμα που θα χρησιμοποιήσουμε εδώ είναι ότι το Rayleigh quotient $R(M, x)$ έχει ως μέγιστη τιμή την μέγιστη ιδιοτιμή του M, και ως σημείο μεγίστου το αντίστοιχο ιδιοδιάνυσμα (π.χ. 1, 2).

Για να τα βρούμε, παρατηρούμε ότι ο πίνακας $C^{-1}aa^T(C^T)^{-1} = (C^{-1}a)(C^{-1}a)$ είναι της κλασικής μορφής ``διάνυσμα επί τον εαυτό του'', και άρα έχει rank 1 (προφανώς, αφού όλες οι γραμμές π.χ. είναι πολλαπλάσια του ίδιου διανύσματος), το οποίο με τη σειρά του σημαίνει ότι έχει $n-1$ φορές ιδιοτιμή το 0, και μόνο μία μη μηδενική ιδιοτιμή, η οποία βρίσκεται εύκολα, αφού:

$$(vv^T)v = v(v^Tv) = \|v\|_2^2 v$$

δηλαδή, κάθε πίνακας της μορφής vv έχει μοναδική μη μηδενική ιδιοτιμή το $\|v\|_2^2$, με αντίστοιχο ιδιοδιάνυσμα το v.

Άρα, το μοναδικό ζευγάρι το οποίο ψάχνουμε, και το οποίο μεγιστοποιεί το Rayleigh quotient, στην συγκεκριμένη περίπτωση, είναι ιδιοτιμή:

$$\|C^{-1}a\|_2^2 = (C^{-1}a)^T(C^{-1}a) = a^T(C^T)^{-1}C^{-1}a = a^T(CC^T)^{-1}a = a^TBa$$

και ιδιοδιάνυσμα:

$$C^{-1}a$$

Τέλος, δεδομένου ότι έχουμε εφαρμόσει έναν μετασχηματισμό στα x, για το σημείο μεγίστου x^* το οποίο ψάχνουμε θα ισχύει:

$$C^Tx^* = C^{-1}a \Rightarrow x^* = (C^T)^{-1}C^{-1}a = (CC^T)^{-1}a = B^{-1}a$$

Επιστρέφοντας, οπότε, στο αρχικό πρόβλημα, αφού $a = \vec{\mu_1} - \vec{\mu_2}$ και $B = \Sigma_1 + \Sigma_2$, θα έχουμε:

$$w^* = (\Sigma_1 + \Sigma_2)^{-1}(\vec{\mu_1} - \vec{\mu_2})$$

Άσκηση 2.5

1. Μας ζητείται η πιθανότητα $P(F = 0|D = 0)$. Δεδομένου ότι η αναφορά του οδηγού εξαρτάται μόνο από την ένδειξη του δείκτη, και από τίποτα άλλο, καθώς και ότι φυσικά δεν εξαρτάται και τίποτα άλλο, π.χ. η ένδειξη του δείκτη, από το τι λέει ο οδηγός, στην πραγματικότητα μπορούμε να μοντελοποιήσουμε τα δεδομένα αυτά ως έναν νέο κόμβο στο δίκτυο, με μία μόνο επιπλέον εισερχόμενη ακμή από το G σε αυτόν. Έστω D ο νέος κόμβος, που είναι βέβαια η ίδια η μεταβλητή που μας δίνεται από την εκφώνηση για τον οδηγό.

Θα έχουμε, δηλαδή, με βάση αυτά ότι η συνολική από κοινού θα παραγοντοποιείται ως:

$$P(B, F, G, D) = P(B)P(F)P(G|B, F)P(D|G)$$

Κατόπιν τούτου, μπορούμε να αρχίσουμε να υπολογίζουμε το ζητούμενο:

$$P(F = 0|D = 0) = \frac{P(F = 0, D = 0)}{P(D = 0)}$$

Το μόνο που μένει είναι να υπολογίσουμε αριθμητή και παρονομαστή, που το κάνουμε βέβαια με ευκολία, αφού πρόκειται για τις περιθώριες κατανομές και θα προκύψουν από την άθροιση της από κοινού:

$$P(F = 0, D = 0) = \sum_{B=0,1}\sum_{G=0,1} P(B)P(F = 0)P(G|B, F = 0)P(D = 0|G) = \cdots = 0.1246$$

και παρομοίως,