ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Θεωρία μηχανικής μάθησης

Γιώργος Στάμου

Καθηγητής Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών ΕΜΠ
Διευθυντής Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης - AILS Lab

Δεδομένα

X : σύνολο στιγμιοτύπων

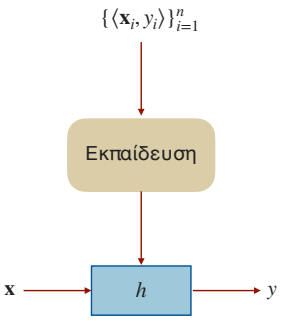
Y : σύνολο ετικετών

$\mathbf{x} \sim \mathscr{D}(X)$

$f : X \to Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle\}$

$\mathbf{x}_i \sim \mathscr{D}(X)$



Διαδικασία εκμάθησης

Πρόβλημα

Βρες ένα ταξινομητή $h \approx f$

Παρατηρήσεις

Δεν γνωρίζουμε την $\mathscr{D}(X)$

Δεν γνωρίζουμε την f

Υποθέτουμε ότι $\mathbf{x}_i \sim \mathscr{D}(X)$

Μέσω του $\mathbb{D}$ γνωρίζουμε ένα μέρος του ταξινομητή f και της κατανομής $\mathscr{D}$

Πρέπει να βρούμε μία καλή εκτίμηση για το κατά πόσο $h \approx f$

ΘΕΩΡΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ - ΣΦΑΛΜΑ ΜΑΘΗΣΗΣ

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

$\mathbf{x} \sim \mathcal{D}(X)$

$f : X \to Y$: ιδανικός ταξινομητής

Έστω $f : X \to \{+1, -1\}$

Σύνολο δεδομένων $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle\}$

$\mathbf{x}_i \sim \mathcal{D}(X)$

Σφάλμα (error)

Άγνωστο το f

$$L_f(h) = \frac{\left| x \in X : h(x) \neq f(x) \right|}{|X|}$$

Άγνωστα τα $f, \mathcal{D}$

$$L_{(\mathcal{D},f)}(h) = \Pr_{\mathbf{x} \sim \mathcal{D}(X)} \left(h(x) \neq f(x) \right)$$

Η πιθανότητα να επιλέξω ένα τυχαίο δείγμα $x \in X$ για το οποίο $h(x) \neq f(x)$

Πρόβλημα

Βρες ένα ταξινομητή $h \approx f$

$$L_{\mathbb{D}}(h) = \frac{\left| i \in \mathbb{N}_n : h(\mathbf{x}_i) \neq y_i \right|}{|n|}$$

Εμπειρικό σφάλμα - σφάλμα εκμάθησης
(empirical error - training error)

ΘΕΩΡΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ - ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ

Παρατηρήσεις

- Το **εμπειρικό σφάλμα** *μπορεί να υπολογιστεί*, συνεπώς αυτό χρησιμοποιείται στη διαδικασία εκμάθησης (empirical risk minimisation - ERM)
- Το **πραγματικό σφάλμα** κρίνει τελικά την λειτουργία του ταξινομητή
- Δεν αποκλείεται να έχουμε *χαμηλό εμπειρικό σφάλμα και υψηλό πραγματικό σφάλμα* (φαινόμενο υπερπροσαρμογής - overfitting)

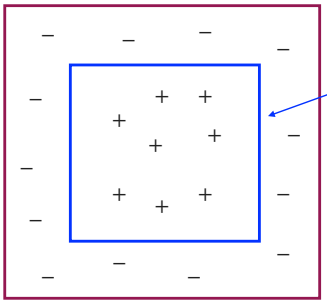
$$L_{\mathbb{D}}(h) = \frac{\left| i \in \mathbb{N}_n : h(\mathbf{x}_i) \neq y_i \right|}{|n|} \qquad L_{(\mathcal{D},f)}(h) = \Pr_{\mathbf{x} \sim \mathcal{D}(X)} \left(h(x) \neq f(x) \right)$$

Επιλέγοντας τον ταξινομητή: $h^0_{\mathbb{D}}(x) = y_i$, αν $x \in \mathbb{D}$
$h^0_{\mathbb{D}}(x) = 0$, αν $x \notin \mathbb{D}$

Δεν είναι δύσκολο να δούμε ότι:

$L_{\mathbb{D}}(h^0_{\mathbb{D}}) = 0$ $\qquad$ $L_{\mathcal{D}}(h^0_{\mathbb{D}}, f) = 0.5$

Σύνολο στιγμιοτύπων X
(έστω εμβαδόν τετραγώνου 2)



Ιδανικός ταξινομητής f
(έστω εμβαδόν τετραγώνου 1)

Έστω $\mathcal{D}$ ομοιόμορφη

Συνεπώς, κατά τη διαδικασία εκμάθησης πρέπει να επιλέξουμε τον ταξινομητή χρησιμοποιώντας το εμπειρικό σφάλμα, βρίσκοντας ταυτόχρονα έναν τρόπο *να αποφύγουμε την υπερπροσαργογή*, καταλήγοντας σε όσο το δυνατόν μικρότερο πραγματικό σφάλμα

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

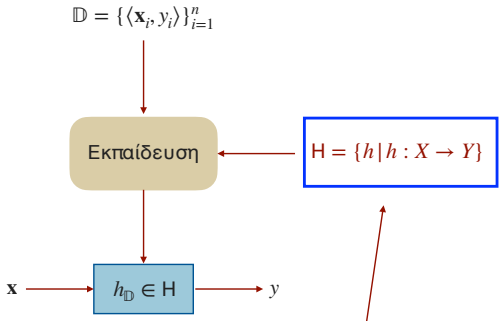
$\mathbf{x} \sim \mathcal{D}(X)$

$f : X \to Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle\}$

$\mathbf{x}_i \sim \mathcal{D}(X)$

H = $\{h \mid h : X \to Y\}$: σύνολο ταξινομητών $X \to Y$



Πρόβλημα

Βρες ένα ταξινομητή $h \approx f$

$h_{\mathbb{D}} \in \arg\min_{h \in \mathrm{H}} \{L_{\mathbb{D}}(h)\}$

Η επιλογή από ένα συγκεκριμένο σύνολο ταξινομητών *πιθανά* μας διαφαλίζει ότι η διαδικασία εκμάθησης δεν θα μπορέσει να καταλήξει σε έναν ταξινομητή τύπου $h_{\mathbb{D}}^{0}$

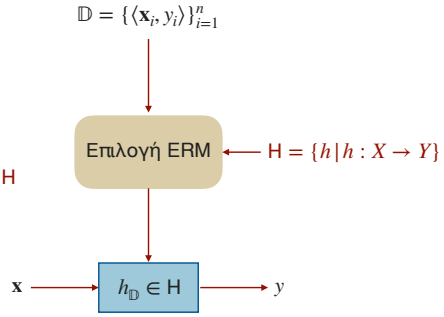
Επαγωγική μεροληψία

H = $\{h \mid h : X \to Y\}$: σύνολο ταξινομητών $X \to Y$

$h_{\mathbb{D}} \in \arg\min_{h \in \mathrm{H}} \{L_{\mathbb{D}}(h)\}$

Η επιλογή της μορφής του h είναι προκαθορισμένη, καθορίζεται από το σύνολο H

Συνεπώς, κάθε περιορισμός του συνόλου H λειτουργεί μεροληπτικά για την μορφή που τελικά θα έχει η h



Το φαινόμενο αυτό αναφέρεται ως **επαγωγική μεροληψία** (inductive bias) και γενικά αυξάνει την πιθανότητα σφάλματος

Για να μην επηρεάζεται σημαντικά το σφάλμα από την επαγωγική μεροληψία, πρέπει η επιλογή της μορφής των υποθέσεων να βασίζεται σε πρότερη γνώση για το πρόβλημα: για παράδειγμα, για την εύρεση ενός ταξινομητή που διακρίνει τους "σκύλους μέσου μεγέθους" θα ήταν λογικό να χρησιμοποιήσουμε παράλληλα-στους-άξονες παραλληλόγραμμα

Όσο πιο μεροληπτική είναι η επιλογή του H τόσο περισσότερο προστατεύει από την υπερπροσαρμογή με πιθανότητα να μεγαλώσει το σφάλμα (ανάλογα με το πρόβλημα και την αντίστοιχη επιλογή του H)

Θεωρούμε ότι το μέγεθος του H είναι φραγμένο, δηλαδή το H είναι πεπερασμένο

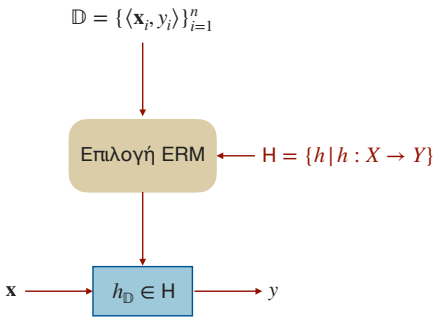
Ισχυρισμός

Για ένα ικανοποιητικά μεγάλο σύνολο δεδομένων εκπαίδευσης $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ (n αρκετά μεγάλο)

η διαδικασία επιλογής $h_{\mathbb{D}} \in \arg\min_{h \in \mathrm{H}} \{L_{\mathbb{D}}(h)\}$ δεν οδηγεί σε υπερπροσαρμογή

Βήματα για την απόδειξη

- **Βήμα 1.** Θα διατυπώσουμε ένα σύνολο από *ρεαλιστικές* υποθέσεις (assumptions) και παρατηρήσεις που αφορούν τη διαδικασία κατασκευής του συνόλου των δεδομένων εκπαίδευσης
- **Βήμα 2.** Θα συσχετίσουμε τον αριθμό των παραδειγμάτων που συμπεριλαμβάνονται στο σύνολο δεδομένων εκπαίδευσης με την πιθανότητα υπερπροσαρμογής, τυποποιώντας σταδιακά τις εμπλεκόμενες έννοιες
- **Βήμα 3.** Θα διατυπώσουμε τυπικά τον ισχυρισμό λαμβάνοντας υπόψην τις υποθέσεις που αφορούν το σύνολο δεδομένων εκπαίδευσης και τη συσχέτιση του μεγέθους του με την υπερπροσαρμογή



Διαδικασία εκμάθησης

Βήμα 1

Υπόθεση 1 (πραγματοποιησιμότητα - realisability)

Υπάρχει $h^* \in$ H τέτοια ώστε $L_{(\mathcal{D},f)}(h^*) = 0$

Η Υπόθεση 1 πρακτικά διασφαλίζει ότι το σύνολο H περιέχει τουλάχιστον έναν καλό ταξινομητή, τον h^*, δεδομένης της κατανομής $\mathbb{D}$ και του ιδανικού ταξινομητή f

Δηλαδή, η Υπόθεση 1 συνεπάγεται ότι με πιθανότητα 1 θα έχουμε $L_{\mathbb{D}}(h^*) = 0$ (αφού $\mathbf{x}_i \sim \mathcal{D}(X)$)

Αυτό σημαίνει ότι με πιθανότητα 1 θα έχουμε $L_{\mathbb{D}}(h_{\mathbb{D}}) = 0$

Υπόθεση 2 (ανεξάρτητα, πανοποιότυπα κατανεμημένα παραδείγματα - independently, identically distributed examples - iid assumption)

Κάθε $\mathbf{x}_i \in \mathbb{D}$ επιλέγεται με ανεξάρτητη δειγματοληψία με βάση την $\mathcal{D}$ και συσχετίζεται με την ετικέτα y_i με βάση την f

Η Υπόθεση 2 πρακτικά διασφαλίζει ότι το σύνολο δεδομένων εκπαίδευσης $\mathbb{D}$ θα είναι ένα παράθυρο παρατήρησης των $\mathcal{D}$ και f

Συμβολικά γράφουμε για την υπόθεση iid ότι $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} \sim \mathcal{D}^n$

Βήμα 1

Το σφάλμα $L_{(\mathcal{D},f)}(h)$ εξαρτάται από τη δειγματοληψία για την κατασκευή του $\mathbb{D}$

Δηλαδή, το $L_{(\mathcal{D},f)}(h)$ είναι μία τυχαία μεταβλητή, στην οποία εμπλέκεται η iid δειγματοληψία του $\mathcal{D}$

Έστω ότι το $\mathbb{D}$ είναι ένα παράθυρο παρατήρησης του $\mathcal{D}$, με κάποια μικρή πιθανότητα δ να είναι μη αντιπροσωπευτικό

Αυτό σημαίνει ότι το $L_{(\mathcal{D},f)}(h_{\mathbb{D}})$ δεν μπορεί να έχει καλύτερο σφάλμα από το δ

Παρατήρηση 1 (παράμετρος εμπιστοσύνης πρόβλεψης - prediction confidence parameter)

Αν δ η πιθανότητα να επιλέξουμε ένα μη αντιπροσωπευτικό δείγμα, τότε $(1-\delta)$ είναι η μέγιστη εμπιστοσύνη πρόβλεψης

Όσον αφορά το f, δεν είναι ρεαλιστικό να θεωρήσουμε ότι το προσεγγίζουμε χωρίς σφάλμα κατά την επισήμανση (labeling)

Θεωρούμε ότι υπάρχει μία μικρή πιθανότητα ϵ να ταξινομήσουμε λάθος ένα δείγμα, κατά την κατασκευή του $\mathbb{D}$

Η ίδια αυτή μικρή πιθανότητα ϵ μπορεί να εισάγει ένα επιπλέον σφάλμα κατά τον υπολογισμό του $L_{(\mathcal{D},f)}(h_{\mathbb{D}})$

Παρατήρηση 2 (προσεγγιστικά ορθή πρόβλεψη - approximately correct prediction)

Αν ϵ η πιθανότητα να ταξινομήσουμε λάθος ένα δείγμα, τότε αν ισχύει ότι $L_{(\mathcal{D},f)}(h_{\mathbb{D}}) \leq \epsilon$ θεωρούμε ότι η $h_{\mathbb{D}}$ ταξινομεί σωστά

Βήμα 2

Έστω $\mathbb{D}|_x = \{x_1, x_2, \ldots, x_n\}$ το σύνολο των παραδειγμάτων που περιέχονται στο $\mathbb{D}$

Όσο μεγαλύτερο το n τόσο μικρότερη η πιθανότητα να μην είναι το $\mathbb{D}|_x$ αντιπροσωπευτικό για το $\mathcal{D}$

Πρακτικά, πρέπει να βρούμε ένα άνω φράγμα για την πιθανότητα να μην είναι το $\mathbb{D}|_x$ αντιπροσωπευτικό για το $\mathcal{D}$

Με τον τρόπο αυτό θα έχουμε βρει ένα άνω φράγμα για την πιθανότητα $\mathcal{D}^n\left(\{\mathbb{D}|_x : L_{(\mathcal{D},f)}(h_{\mathbb{D}}) > \epsilon\}\right)$

Θεωρούμε ότι το $h \in \mathrm{H}$ είναι μία κακή υπόθεση, αν $L_{(\mathcal{D},f)}(h) > \epsilon$

Έστω $\mathrm{H}_B = \{h \in \mathrm{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$ το σύνολο των κακών υποθέσεων

Έστω $\mathrm{M} = \{\mathbb{D}|_x : \exists h \in \mathrm{H}_B : L_{\mathbb{D}}(h) = 0\} = \bigcup_{h \in \mathrm{H}_B} \{\mathbb{D}|_x : L_{\mathbb{D}}(h) = 0\}$ το σύνολο των κακών συνόλων δεδομένων εκπαίδευσης

Λόγω της πραγματοποιησιμότητας (Υπόθεση 1), ισχύει ότι $\{\mathbb{D}|_x : L_{(\mathcal{D},f)}(h_{\mathbb{D}}) > \epsilon\} \subseteq \mathrm{M}$

Συνεπώς, $\mathcal{D}^n\left(\{\mathbb{D}|_x : L_{(\mathcal{D},f)}(h_{\mathbb{D}}) > \epsilon\}\right) \leq \mathcal{D}^n(\mathrm{M}) = \mathcal{D}^n\left(\bigcup_{h \in \mathrm{H}_B} \{\mathbb{D}|_x : L_{\mathbb{D}}(h) = 0\}\right) \leq \sum_{h \in \mathrm{H}_B} \mathcal{D}^n\left(\{\mathbb{D}|_x : L_{\mathbb{D}}(h) = 0\}\right)$

Βήμα 2

Λόγω της υπόθεσης iid ισχύει ότι $\mathcal{D}^n\left(\{\mathbb{D}|_x : L_{\mathbb{D}}(h)=0\}\right) = \mathcal{D}^n\left(\{\mathbb{D}|_x : \forall i \in \mathbb{N}_n\ h(x_i)=f(x_i)\}\right) = \prod_{i=1}^{n} \mathcal{D}\left(\{x_i : h(x_i)=f(x_i)\}\right)$

Όμως, για κάθε δειγματοληψία ενός στοιχείου του $\mathcal{D}\left(\{x_i : h(x_i)=f(x_i)\}\right) = 1 - L_{(\mathcal{D},f)}(h) \leq 1-\epsilon$ αφού $h \in \mathsf{H}_B$

Χρησιμοποιώντας την ανίσωση $1-\epsilon \leq e^{-\epsilon}$ έχουμε $\mathcal{D}^n\left(\{\mathbb{D}|_x : L_{\mathbb{D}}(h)=0\}\right) \leq (1-\epsilon)^n \leq e^{-\epsilon n}$ και τελικά ισχύει ότι:

$$\boxed{\mathcal{D}^n\left(\{\mathbb{D}|_x : L_{(\mathcal{D},f)}(h_{\mathbb{D}}) > \epsilon\}\right) \leq |\mathsf{H}_B| e^{-\epsilon n} \leq |\mathsf{H}| e^{-\epsilon n}}$$

Βήμα 3

Έστω H ένα πεπερασμένο σύνολο από υποθέσεις

Έστω $\delta \in (0,1)$ και $\epsilon > 0$

Έστω $n \in \mathbb{N}$ τέτοιος ώστε $n \geq \dfrac{\log(|\mathsf{H}|/\delta)}{\epsilon}$

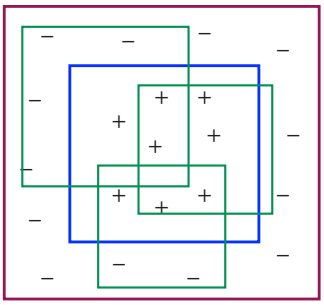
Τότε, για κάθε ταξινομητή f και κάθε κατανομή $\mathcal{D}$, αν ισχύει η υπόθεση της πραγματοποιησιμότητας

και κατασκευάσουμε με iid δειγματοληψία με βάση την $\mathcal{D}$ ένα σύνολο δεδομένων $\mathbb{D}$ μεγέθους n

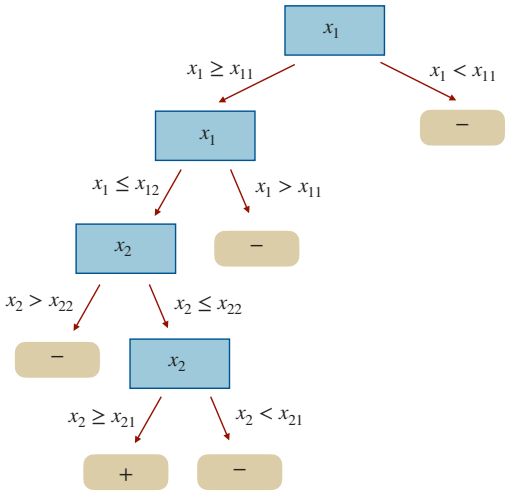
με πιθανότητα τουλάχιστον $1-\delta$, για κάθε ERM υπόθεση $h_{\mathbb{D}}$ θα ισχύει $L_{(\mathcal{D},f)}(h_{\mathbb{D}}) \leq \epsilon$

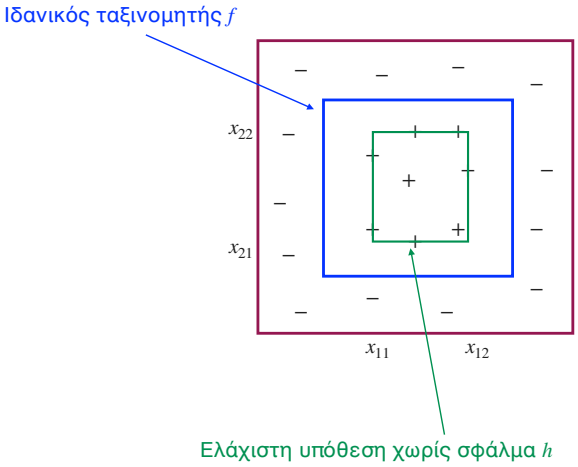
Παράδειγμα - Τετραγωνικοί ταξινομητές

- Επιλέγουμε τετραγωνικούς ταξινομητές, παράλληλους με τους άξονες, στο όριο του πεδίου τιμών του χώρου υποθέσεων
  - Μπορούν να χρησιμοποιηθούν δέντρα αποφάσεων
- *Η επιλογή αυτή διασφαλίζει ότι θα αποφύγουμε την υπερπροσαργογή;*

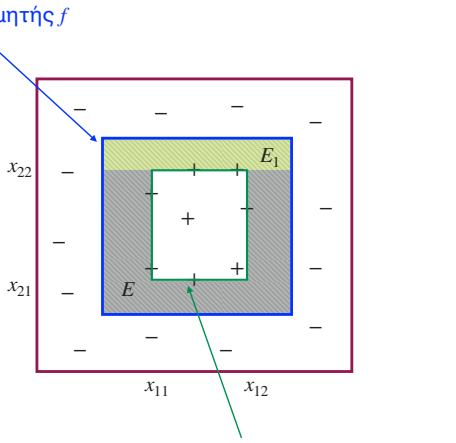


Παράδειγμα - Τετραγωνικοί ταξινομητές





Παράδειγμα - Τετραγωνικοί ταξινομητές



Η υπόθεση h μπορεί να υπολογιστεί εύκολα (σε πολυωνυμικό χρόνο)

Η υπόθεση h είναι υποσύνολο της f

Το πραγματικό σφάλμα της h περιορίζεται στην πιθανότητα ένα στιγμιότυπο να βρίσκεται στην περιοχή E της συμμετρικής διαφοράς των h και f

Έστω ότι το σφάλμα αυτό είναι το πολύ ϵ

Χωρίζουμε την E σε 4 υποπεριοχές-παραλληλόγραμμα και έστω ότι το σφάλμα της κάθε μίας από αυτές (έστω της E_1) είναι το πολύ $\epsilon/4$

Αφού $\mathbf{x}_i \sim \mathcal{D}(X)$ η πιθανότητα να μην επιλέξουμε κάποιο δείγμα στην E_1 κατά τη δειγματοληψία του $\mathbb{D}$ είναι το πολύ $1-\epsilon/4$

Συνεπώς, η πιθανότητα να μην επιλέξουμε συνολικά κανένα δείγμα στην E_1 κατά τη δειγματοληψία του $\mathbb{D}$ είναι το πολύ $(1-\epsilon/4)^n$

Έστω δ αυτή η πιθανότητα για τα 4 υποπαραλληλόγραμμα, άρα $4(1-\epsilon/4)^n \leq \delta$

Λύνοντας ως προς n, χρησιμοποιώντας την ταυτότητα $1-x \leq e^{-x}$ έχουμε $n \geq (4/\epsilon)\ln(4/\delta)$

Συνεπώς, αν $n \geq (4/\epsilon)\ln(4/\delta)$ τότε, με πιθανότητα τουλάχιστον $1-\delta$, η h θα έχει πραγματικό σφάλμα το πολύ ϵ

Ορισμός - Πιθανά προσεγγιστικά ορθώς εκπαιδεύσιμο σύνολο υποθέσεων H

Ένα σύνολο υποθέσεων H είναι πιθανά προσεγγιστικά ορθώς εκπαιδεύσιμο ή PAC εκπαιδεύσιμο
PAC learnable - Probably Approximately Correct learnable

αν υπάρχει συνάρτηση $n : (0,1)^2 \to \mathbb{N}$ και αλγόριθμος εκπαίδευσης learn ώστε:

Για κάθε $\epsilon, \delta \in (0,1)$

Για κάθε ταξινομητή $f : X \to \{0,1\}$ και κάθε κατανομή $\mathcal{D}$ πάνω στο X
για τα οποία ισχύει η υπόθεση της πραγματοποιησιμότητας του H

Για κάθε σύνολο δεδομένων εκμάθησης $\mathbb{D}$ με $|\mathbb{D}| \geq n_{(\delta,\epsilon)}$
με δείγματα του X που λαμβάνονται με βάση την $\mathcal{D}$ με iid δειγματοληψία και ταξινομούνται από τον f

Υπάρχει $h_{\mathbb{D}} = \text{learn}(\mathbb{D})$ για το οποίο με πιθανότητα τουλάχιστον $1-\delta$ ισχύει ότι $L_{(\mathcal{D},f)}(h_{\mathbb{D}}) \leq \epsilon$

S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms,
Cambridge University Press (2014)

Παρατηρήσεις

Η συνάρτηση $n : (0,1)^2 \to \mathbb{N}$ καθορίζει πόσα παραδείγματα χρειάζονται για να είναι εκπαιδεύσιμο το H

Είναι συνάρτηση των ϵ, δ αλλά η μορφή της εξαρτάται από τη δομή του H

Για παράδειγμα, όταν το H είναι πεπερασμένο, η n εξαρτάται λογαριθμικά από το $|H|$

$$n \geq \frac{\log(|H|/\delta)}{\epsilon}$$

Πρακτικά, έχει νόημα η χρήση της ελάχιστης n για την οποία ισχύει η PAC εκπαιδευσιμότητα
ώστε να παίρνουμε τον μικρότερο ακέραιο $n = |\mathbb{D}|$ για τον οποίο μπορούμε να βρούμε καλό ταξινομητή

Με το παράδειγμα των παράλληλων στους άξονες παραλληλογράμμων (στην περίπτωση αυτή το H είναι άπειρο)
είδαμε ότι η PAC εκπαιδευσιμότητα εξαρτάται (εκτός από το πλήθος) και από τη μορφή των υποθέσεων του H

Στην περίπτωση των παράλληλων με τους άξονες παραλληλογράμων έχουμε:

$$n \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

Στη συνέχεια θα μελετήσουμε τον τρόπο με τον οποίο επηρεάζει η μορφή των υποθέσεων του H την εκπαιδευσιμότητά του

Το σύνολο δεδομένων εκπαίδευσης μπορεί να είναι παραπλανητικό και να οδηγήσει σε υπερπροσαρμογή

Ένας τρόπος να αντιμετωπιστεί αυτό είναι ο περιορισμός των υποθέσεων σε ένα συγκεκριμένο σύνολο H

Ο περιορισμός του H για να είναι αποδοτικός πρέπει να στηρίζεται σε πρότερη γνώση για το πρόβλημα
(για παράδειγμα μπορούμε να διακρίνουμε τους μεσαίου μεγέθους σκύλους, με βάση το ύψος και το βάρος τους,
με παράλληλα στους άξονες παραλληλόγραμμα)

Είναι η πρότερη γνώση απαραίτητη για την εκπαίδευση σε κάθε προβλημα, αποφεύγοντας την υπερπροσαρμογή;

Υπάρχει κάποιο H το οποίο μπορεί να εκπαιδευτεί σε κάθε πρόβλημα, χωρίς υπερπροσαρμογή;

Υπάρχει αλγόριθμος εκμάθησης learn και σύνολο δεδομένων εκμάθησης $\mathbb{D}$ μεγέθους n τέτοια ώστε
για κάθε κατανομή $\mathcal{D}$ στο $X \times Y$ μέσω της οποίας δέχεται ο learn με iid δειγματοληψία n δείγματα
να βρίσκει έναν ταξινομητή $h : X \to Y$ με χαμηλό πραγματικό σφάλμα $L_{\mathcal{D}}(h)$;

?

Θεώρημα No-Free-Lunch (NFL)

Έστω learn ένας αλγόριθμος εκμάθησης για το πρόβλημα της δυαδικής ταξινόμησης σε ένα σύνολο X

Έστω $n < \frac{|X|}{2}$ ένα άνω φράγμα στο μέγεθος του συνόλου δεδομένων εκπαίδευσης

Τότε, υπάρχει κατανομή $\mathscr{D}$ στο $X \times \{0,1\}$ τέτοια ώστε:

1. Υπάρχει ταξινομητής $f : X \to \{0,1\}$ με $L_{\mathscr{D}}(f) = 0$
2. Με πιθανότητα τουλάχιστον $1/7$, για ένα σύνολο δεδομένων εκπαίδευσης $\mathbb{D}$ με $\mathbb{D} \sim \mathscr{D}^n$, έχουμε $L_{\mathscr{D}}(\text{learn}(\mathbb{D})) \geq 1/8$

Υπάρχει αλγόριθμος εκμάθησης learn και σύνολο δεδομένων εκμάθησης $\mathbb{D}$ μεγέθους n τέτοια ώστε για κάθε κατανομή $\mathscr{D}$ στο $X \times Y$ μέσω της οποίας δέχεται ο learn με iid δειγματοληψία n δείγματα να βρίσκει έναν ταξινομητή $h : X \to Y$ με χαμηλό πραγματικό σφάλμα $L_{\mathscr{D}}(h)$;

ΟΧΙ

Sterkenburg, T.F., Grünwald, P.D. The no-free-lunch theorems of supervised learning
Synthese 199, 9979–10015 (2021)

ΘΕΩΡΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ - ΣΥΜΒΙΒΑΣΜΟΣ ΜΕΡΟΛΗΨΙΑΣ ΠΟΛΥΠΛΟΚΟΤΗΤΑΣ

Θεώρημα No-Free-Lunch (NFL)

Έστω learn ένας αλγόριθμος εκμάθησης για το πρόβλημα της δυαδικής ταξινόμησης σε ένα σύνολο X

Έστω $n < \frac{|X|}{2}$ ένα άνω φράγμα στο μέγεθος του συνόλου δεδομένων εκπαίδευσης

Τότε, υπάρχει κατανομή $\mathscr{D}$ στο $X \times \{0,1\}$ τέτοια ώστε:

1. Υπάρχει ταξινομητής $f : X \to \{0,1\}$ με $L_{\mathscr{D}}(f) = 0$
2. Με πιθανότητα τουλάχιστον $1/7$, για ένα σύνολο δεδομένων εκπαίδευσης $\mathbb{D}$ με $\mathbb{D} \sim \mathscr{D}^n$, έχουμε $L_{\mathscr{D}}(\text{learn}(\mathbb{D})) \geq 1/8$

Σκιαγράφηση απόδειξης (proof sketch)

Έστω $C \subset X$ με $|C| = 2n$

Η ιδέα είναι να δείξουμε ότι ένας αλγόριθμος εκπαίδευσης που παρατηρεί μόνο τα μισά δείγματα του C δεν μπορεί να γνωρίζει πως ταξινομούνται τα υπόλοιπα μισά δείγματα του C

Επομένως, θα αναζητήσουμε μία κατανομή $\mathscr{D}$ για να επιλέγουμε δείγματα από το C και έναν ταξινομητή f που επισημαίνει τα υπόλοιπα δείγματα του C με αντίθετο τρόπο από την $h = \text{learn}(\mathbb{D})$ αυξάνοντας έτσι το πραγματικό σφάλμα της h

Θεώρημα No-Free-Lunch (NFL)

Σκιαγράφηση απόδειξης (proof sketch)

Για τον "ανταγωνιστή" ταξινομητή f έχουμε τη δυνατότητα να επιλέξουμε μεταξύ όλων των $f_1, f_2, \ldots, f_T$ όπου $T = 2^{2n}$

δηλαδή μεταξύ όλων των διαφορετικών συναρτήσεων $C \to \{0,1\}$

Για οποιαδήποτε επιλογή f_i $(i \in \mathbb{N}_T)$, αν θεωρήσουμε την κατανομή $\mathcal{D}_i$ πάνω στο $C \times \{0,1\}$, με

$\mathcal{D}_i(\{x,y\}) = 1/|C|$, αν $y = f_i(x)$ αλλιώς $\mathcal{D}_i(\{x,y\}) = 0$

Δεν είναι δύσκολο να δούμε ότι $L_{\mathcal{D}_i}(f_i) = 0$

Επιπλέον, μπορεί να αποδειχθεί ότι $\max_{i \in \mathbb{N}_T} \mathbb{E}_{\mathbb{D} \sim \mathcal{D}_i^n}[L_{\mathcal{D}_i}(\text{learn}(\mathbb{D}))] \geq 1/4$

Αυτό σημαίνει ότι μπορεί να βρεθεί μία f_i που έχει μηδενικό σφάλμα για την κατανομή $\mathcal{D}_i$

ενώ ταυτόχρονα η learn βρίσκει ένα h με μικρό πραγματικό σφάλμα $L_{(\mathcal{D},f)}(h)$

Αποσύνθεση πραγματικού σφάλματος

Περιορίζοντας το σύνολο H αποφεύγουμε πιθανά την υπερπροσαρμογή

χρειαζόμαστε όμως ένα πλούσιο H για να μπορέσουμε να μειώσουμε το εμπειρικό σφάλμα

αλλά όχι τόσο πλούσιο (για παράδειγμα όλες τις συναρτήσεις) ώστε πρακτικά να μην εκπαιδεύεται

Έστω ότι ένας ERM αλγόριθμος εκπαίδευσης του H με είσοδο $\mathbb{D}$ καταλήγει σε μία υπόθεση $h_{\mathbb{D}}$. Τότε:

$$L_{\mathcal{D}}(h_{\mathbb{D}}) = \epsilon_{app} + \epsilon_{est}$$

$$\epsilon_{app} = \min_{h \in H} L_{\mathcal{D}}(h)$$

$$\epsilon_{est} = L_{\mathcal{D}}(h_{\mathbb{D}}) - \min_{h \in H} L_{\mathcal{D}}(h)$$

Σφάλμα προσέγγισης (approximation error): ποιο είναι το ρίσκο από τον περιορισμό στο συγκεκριμένο σύνολο υποθέσεων, πόσο προβληματική είναι η επαγωγική μεροληψία

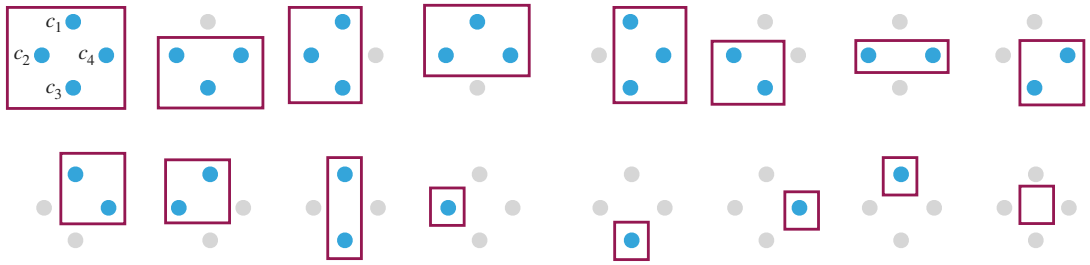
Σφάλμα εκτίμησης (estimation error): ποιο είναι το ρίσκο η υπόθεση που θα επιλέξει ο αλγόριθμος εκπαίδευσης να μην έχει το καλύτερο πραγματικό σφάλμα, παρότι έχει το καλύτερο εμπειρικό

Παράδειγμα (παράλληλα στους άξονες παραλληλόγραμμα)

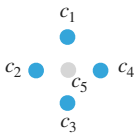
Η $= \{h(a_1, a_2, b_1, b_2) : a_1 \leq a_2$ και $b_1 \leq b_2\}$, όπου $h(a_1, a_2, b_1, b_2) = 1$ αν και μόνο αν $a_1 \leq x_1 \leq a_2$ και $b_1 \leq x_2 \leq b_2$

Αν θεωρήσουμε $C = \{c_1, c_2, \ldots, c_k\}$, ποιο το μέγιστο k για το οποίο για κάθε $f : C \to \{0,1\}$ υπάρχει ένα $h \in$ Η τέτοιο ώστε $h = f$;

$k = 4$



$k = 5$



Δεν υπάρχει $h \in$ Η τέτοιο ώστε $h = f^* = \{c_1 \to 1, c_2 \to 1, c_3 \to 1, c_4 \to 1, c_5 \to 0\}$

Παράδειγμα (παράλληλα στους άξονες παραλληλόγραμμα) - Παρατηρήσεις

Η υπόθεση της πραγματοποιησιμότητας διασφαλίζει ότι δεν θα παρατηρήσουμε συναρτήσεις όπως η f^*
αφού γνωρίζουμε ότι υπάρχει $h^* \in$ Η με $L_{(\mathcal{D},f)}(h^*) = 0$

Επομένως, όσα δείγματα και να πάρουμε για το $\mathbb{D}$ δεν θα ξεφύγουμε από τη μορφή των συναρτήσεων του $k = 4$

Επιπλέον, για όλα τα υπόλοιπα στοιχεία του X που δεν είναι στο $\mathbb{D}$ δεν θα υπάρχουν συναρτήσεις της μορφής της f^*
με $L_{\mathcal{D}}(f^*) = 0$

Άρα, δεν θα μπορούν να βρεθούν εύκολα συναρτήσεις-αντιπαραδείγματα
όπως αυτά που χρησιμοποιήθηκαν στην απόδειξη του θεωρήματος NFL

Με τον τρόπο αυτό κατανοούμε γιατί είναι PAC learnable τα παράλληλα στους άξονες παραλληλόγραμμα
και ποια είναι η μορφή των προβλημάτων ταξινόμησης που μπορούν να επιλύσουν

Συνεπώς, ο αριθμός k είναι ενδεικτικός για την εκπαιδευσιμότητα και την επαγωγική μεροληψία του Η

Ορισμός - Περιορισμός του H στο C

Έστω H ένα σύνολο υποθέσεων από το X στο $\{0,1\}$

Έστω $C = \{c_1, c_2, \ldots, c_n\} \subset X$ ένα πεπερασμένο υποσύνολο του X

Περιορισμός (restriction) του H στο C είναι το σύνολο των υποθέσεων $h \in$ H με $h : C \to \{0,1\}$ δηλαδή το σύνολο H_C των διανυσμάτων $\{0,1\}^n$ με $H_C = \{(h(c_1), h(c_2), \ldots, h(c_n)) : h \in H\}$

Ορισμός - Κατακερματισμός του C από το H

Ένα σύνολο υποθέσεων H κατακερματίζει (shatters) ένα πεπερασμένο σύνολο $C \subset X$ αν ο περιορισμός του H στο C είναι το σύνολο όλων των συναρτήσεων από $C \to \{0,1\}$

Παράδειγμα (παράλληλα στους άξονες παραλληλόγραμμα)

Έστω $H = \{h(a_1, a_2, b_1, b_2) : a_1 \le a_2$ και $b_1 \le b_2\}$, όπου $h(a_1, a_2, b_1, b_2) = 1$ αν και μόνο αν $a_1 \le x_1 \le a_2$ και $b_1 \le x_2 \le b_2$

Το H κατακερματίζει τα σύνολα $C^{(1)} = \{c_1\}, C^{(2)} = \{c_1, c_2\}, C^{(3)} = \{c_1, c_2, c_3\}, C^{(4)} = \{c_1, c_2, c_3, c_4\}$

ενώ δεν κατακερματίζει το σύνολο $C^{(5)} = \{c_1, c_2, c_3, c_4, c_5\}$ όπως και κάθε σύνολο $C^{(n)} = \{c_1, c_2, \ldots, c_n\}$ με $n > 5$

Ορισμός - Διάσταση VC του H

Έστω H ένα σύνολο υποθέσεων από το X στο $\{0,1\}$

Η VC διάσταση (VC dimension) του H είναι η μέγιστη πληθικότητα n των συνόλων $C \subset X$ που κατακερματίζονται από το H

Αν το H κατακερματίζει σύνολα $C \subset X$ οποιασδήποτε πληθικότητας, τότε έχει άπειρη VC διάσταση

Θεώρημα (εναλλακτική διατύπωση του Θεωρήματος NFL)

Έστω H ένα σύνολο υποθέσεων από το X στο $\{0,1\}$

Έστω ένα σύνολο δειγμάτων του X μεγέθους n

Αν υπάρχει $C \subset X$ μεγέθους $2n$ που κατακερματίζεται από το H, τότε:

για κάθε αλγόριθμο εκμάθησης learn για το πρόβλημα της δυαδικής ταξινόμησης στο X

υπάρχει κατανομή $\mathcal{D}$ στο $X \times \{0,1\}$ και υπόθεση $h \in$ H με $L_{\mathcal{D}}(h) = 0$

αλλά με πιθανότητα τουλάχιστον $1/7$ στην επιλογή $\mathbb{D} \sim \mathcal{D}^n$ έχουμε $L_{\mathcal{D}}(\text{learn}(\mathbb{D})) \ge 1/8$

Θεώρημα

Κάθε σύνολο υποθέσεων H άπειρης VC διάστασης δεν είναι PAC εκπαιδεύσιμο