Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Λογικής και Επιστήμης Υπολογισμών

Προηγμένα Θέματα Αλγορίθμων

PAC Learning

Μία εισαγωγή βασισμένη στα κεφάλαια 2-3 του βιβλίου
Understanding Machine Learning: *From Theory to Algorithms*

Αργύρης Μουζάκης

*amouzakis*17@*hotmail.com*

7 Νοεμβρίου 2018

Περιεχόμενα

Μέρος I

Εισαγωγή στο *PAC Learning*

Κεφάλαιο 1

Εισαγωγή

Το *Machine Learning* (*ML*) είναι τομέας της τεχνητής νοημοσύνης που έχει αποκτήσει ιδιαίτερο ενδιαφέρον τα τελευταία χρόνια. Αφορά στη μελέτη των διαδικασιών όπου βασίζεται η αφομοίωση νέων εννοιών από ζωντανούς οργανισμούς, με σκοπό την υλοποίηση αντίστοιχων μηχανισμών σε υπολογιστές. Έτσι, καθίσταται δυνατή η δημιουργία υπολογιστικών διεργασιών με πολύ μεγαλύτερη δυνατότητα προσαρμογής.

Πρόκειται για πεδίο με διεπιστημονικό χαρακτήρα. Βρίσκεται στο μεταίχμιο των μαθηματικών και της πληροφορικής. Από τα πρώτα προέρχονται τα στατιστικά μοντέλα μέσω των οποίων περιγράφεται μαθηματικά η διαδικασία της μάθησης. Από την δεύτερη προέρχονται οι αλγόριθμοι και οι τεχνικές που επιτρέπουν την υλοποίηση αυτών των μοντέλων σε υπολογιστικό περιβάλλον.

Στο πλαίσιο αυτού του κειμένου παρουσιάζονται κάποιες εισαγωγικές έννοιες που βοηθούν στη μελέτη των αλγοριθμικών πτυχών του αντικειμένου. Στόχος αποτελεί η παρουσίαση της έννοιας του *PAC learning* και η απόδειξη βασικών ιδιοτήτων που απορρέουν από αυτό.

Κεφάλαιο 2

Μια ομαλή αρχή

2.1 Θεμελιώδεις Έννοιες

Αρχικά, δίνουμε κάποιους ορισμούς που επιτρέπουν τη φορμαλιστική διατύπωση του προβλήματος της μάθησης.

Το κεντρικό πρόβλημα του ML αφορά στην εκπαίδευση ενός μαθητευόμενου, ώστε να μπορεί να χωρίσει σε κατηγορίες τα στοιχεία ενός συνόλου με βάση τα χαρακτηριστικά τους. Συγκεκριμένα, ο μαθητευόμενος έχει πρόσβαση:

- Στο *domain set* $\mathcal{X}$, που αναπαριστά το σύνολο όλων των αντικειμένων που θέλουμε να κατηγοριοποιηθούν. Κάθε στοιχείο του συνόλου αναπαρίσταται με βάση κάποιες μεταβλητές που δηλώνουν τα χαρακτηριστικά του.
- Στο *label set* $\mathcal{Y}$. Για αρχή, θεωρούμε πως αποτελείται από τις τιμές 0 και 1. Σε κάθε στοιχείο του $\mathcal{X}$ αντιστοιχίζεται ακριβώς μία από αυτές τις τιμές, γεγονός που διαχωρίζει τα στοιχεία του.
- Στο δείγμα εκπαίδευσης S. Μια ακολουθία $S = ((x_1, y_1), ..., (x_m, y_m))$ που αποτελείται από στοιχεία του $\mathcal{X}$ που γνωρίζουμε τις τιμές $y_i \in \mathcal{Y}$ που τους αντιστοιχούν.

Επιπλέον, το αποτέλεσμα της εκπαίδευσης είναι μία συνάρτηση $h_S : \mathcal{X} \to \mathcal{Y}$ (υπόθεση/ταξινομητής) και δηλώνει την αντίληψη που διαμορφώθηκε για το πώς πρέπει να κατηγοριοποιηθούν τα στοιχεία του *domain set* με βάση το δείγμα. Θεωρούμε πως υπάρχει σωστός ταξινομητής f, που όμως δεν είναι γνωστός και που θέλουμε να προσεγγίσουμε όσο γίνεται καλύτερα. Τέλος, θεωρούμε πως υπάρχει μία (επίσης άγνωστη) κατανομή $\mathcal{D}$ με βάση την οποία παράγονται τα στοιχεία του $\mathcal{X}$.

Θα ορίσουμε κάποια μεγέθη που μας δίνουν τη δυνατότητα να αξιολογήσουμε τους διάφορους ταξινομητές. Ένα τέτοιο μέγεθος είναι η πιθανότητα, επιλέγοντας τυχαία κάποιο $x \in \mathcal{X}$, η τιμή που του αποδίδει ο ταξινομητής να διαφέρει από τη σωστή. Το μέγεθος αυτό λέγεται πραγματικό ρίσκο και δίνεται από τον τύπο $L_{(\mathcal{D},f)}(h) \stackrel{def}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{def}{=} \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq f(x)\})$. Το μέγεθος αυτό δεν μπορεί να υπολογιστεί ακριβώς, διότι δεν γνωρίζουμε ούτε την f ούτε τη $\mathcal{D}$. Για αυτό το λόγο ορίζουμε το εμπειρικό ρίσκο (*empirical risk*), το

οποίο αφορά στην απόκλιση των τιμών του ταξινομητή από τις πραγματικές, εφόσον περιοριζόμαστε στα στοιχεία του δείγματος. Ο τύπος που το εκφράζει είναι $L_S(h) \stackrel{def}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$, όπου $[m] = \{1, ..., m\}$.

2.2 *Empirical Risk Minimization* (*ERM*)

Με βάση τους ορισμούς που δόθηκαν πριν, κάνουμε μία πρώτη απόπειρα επίλυσης του προβλήματος της εύρεσης βέλτιστου ταξινομητή. Συγκεκριμένα, απαιτούμε το εμπειρικό ρίσκο να έχει την ελάχιστη δυνατή τιμή. Αυτή η σκέψη, όμως, δεν αρκεί από μόνη της. Ο λόγος είναι πως η ελάχιστη τιμή που μπορεί να πάρει το εμπειρικό ρίσκο είναι το 0. Ένας προφανής τρόπος να επιτευχθεί αυτό είναι θεωρώντας τον ταξινομητή:

$$h(x) = \begin{cases} y_i & if\ \exists i \in [m], s.t.\ x_i = x \\ 0 & otherwise \end{cases}$$

Ο ταξινομητής αυτός προβλέπει σωστά τις τιμές των αντικειμένων που ανήκουν στο δείγμα, καθώς και των αντικειμένων εκτός δείγματος όπου αντιστοιχεί η τιμή 0. Έστω ότι το δείγμα είναι μικρό και υπάρχουν πολλά αντικείμενα που δεν ανήκουν σε αυτό στα οποία αντιστοιχεί η τιμή 1. Τότε, τα αποτελέσματα που προκύπτουν μπορεί να είναι αυθαίρετα άσχημα, παρά το γεγονός ότι $L_S(h) = 0$. Το φαινόμενο αυτό καλείται $overfitting$. Αν και ο εν λόγω ταξινομητής φαίνεται "τεχνητός", υπό την έννοια ότι έχει σχεδιαστεί για να αποτύχει, υπάρχει αναπαράστασή του με πολυώνυμα (βλ. παράρτημα).

Προκειμένου να αποφευχθεί το προηγούμενο φαινόμενο, θεωρούμε πως υπάρχει μία συγκεκριμένη κλάση υποθέσεων $\mathcal{H}$ που μεταξύ των στοιχείων της προσπαθούμε να βρούμε ποιο προσεγγίζει καλύτερα τον πραγματικό ταξινομητή f. Έτσι, αντί να επιτρέπουμε στους αλγορίθμους μάθησης να επιλέγουν ταξινομητή με μοναδικό κριτήριο την ελαχιστοποίηση του εμπειρικού ρίσκου, αντιμετωπίζουμε πλέον το πρόβλημα ως ένα πρόβλημα αναζήτησης όπου ο χώρος λύσεων είναι η $\mathcal{H}$. Αυτή η παραλλαγή της τεχνικής λέγεται *empirical risk minimization with inductive bias*. Έτσι, στόχος μας πλέον είναι να βρούμε την υπόθεση $h_S \in \mathcal{H}$ που ελαχιστοποιεί το εμπειρικό ρίσκο (συμβολίζουμε $ERM_{\mathcal{H}}(S) = \underset{h \in \mathcal{H}}{argmin}\ L_S(h)$- αν υπάρχουν περισσότερες από μία θέσεις ελαχίστου, επιλέγεται μία αυθαίρετα και στην προηγούμενη έκφραση πρέπει να γράψουμε $\in$ αντί για $=$). Υπό την προϋπόθεση ότι έχει επιλεγεί σωστά η κλάση υποθέσεων $\mathcal{H}$ (δηλαδή, αποτελείται από υποθέσεις που δεν οδηγούν σε $overfitting$), αυτός ο περιορισμός μπορεί να οδηγήσει σε αισθητά καλύτερα αποτελέσματα. Στα επόμενα, θα θεωρήσουμε πως η κλάση υποθέσεων που έχει δοθεί έχει επιλεγεί σωστά.

2.3 Πεπερασμένες κλάσεις υποθέσεων

Σε αυτό το μέρος θεωρούμε πως αντιμετωπίζουμε προβλήματα μάθησης όπου η κλάση υποθέσεων είναι πεπερασμένη. Στο πλαίσιο μίας πρώτης προσέγγισης αυτής της κατηγορίας προβλημάτων, κάνουμε μία παραδοχή που, αν και αρκετά περιοριστική, θα μας βοηθήσει σημαντικά σε αυτό το στάδιο. Πρόκειται για την παραδοχή ικανοποιησιμότητας (*realizability assumption*):

Παραδοχή 1. *Υποθέτουμε πως υπάρχει υπόθεση $h^* \in \mathcal{H}$ με $L_{(\mathcal{D},f)}(h^*) = 0$. Έτσι, για οποιοδήποτε δείγμα S που έχει προκύψει από το $\mathcal{X}$ σύμφωνα με την κατανομή $\mathcal{D}$ και τα στοιχεία του δείγματος κατηγοριοποιούνται βάσει της f, ισχύει $L_S(h^*) = 0$.*

Εφόσον ισχύει το παραπάνω, στόχος μας είναι να εντοπίσουμε μία υπόθεση h^* για την οποία ισχύει $L_{(\mathcal{D},f)}(h^*) = 0$. Δοθέντος, λοιπόν, ενός δείγματος S, εφόσον ισχύουν τα προηγούμενα, η εκτέλεση του ERM αφήνει στο τέλος μόνο όσες υποθέσεις έχουν μηδενικό εμπειρικό ρίσκο. Δεδομένου ότι το $L_{(\mathcal{D},f)}(h_S)$ δεν μπορεί να υπολογιστεί, το μόνο που μας μένει είναι να επιλέξουμε μία από τις εναπομείνασες υποθέσεις, ελπίζοντας ότι θα αποδειχθεί βέλτιστη. Θα εξετάσουμε υπό ποιες συνθήκες έχουμε βάσιμες πιθανότητες να ισχύει αυτό.

Επιπλέον, κάνουμε την παραδοχή ανεξαρτησίας και ισονομίας για το δείγμα (*i.i.d. assumption*):

Παραδοχή 2. *Τα στοιχεία του $\mathcal{X}$ που συγκροτούν το δείγμα έχουν προκύψει με δειγματοληψία του $\mathcal{X}$ σύμφωνα με την κατανομή $\mathcal{D}$ και είναι ανεξάρτητα. Έτσι, έχουμε $S|_x = (x_1, ..., x_m) \sim \mathcal{D}^m$.*

Δεδομένου ότι ο τρόπος που επιλέγεται το S είναι τυχαίος, αντίστοιχα τυχαία είναι και η επιλογή του ταξινομητή h_S που προκύπτει από αυτό. Αυτό ισχύει και για το $L_{(\mathcal{D},f)}(h_S)$. Ως εκ τούτου, δεν μπορούμε γενικά να προσδοκούμε πως θα καταλήξουμε σε αποτέλεσμα για το οποίο να ισχύει $L_{(\mathcal{D},f)}(h_S) = 0$.

Όμως, ισχύει ότι $\mathbb{P}[L_S(h_S) \xrightarrow[m\to\infty]{} L_{(\mathcal{D},f)}(h_S)] = 1$ (βλ. παράρτημα). Επομένως, αυξάνοντας το μέγεθος του δείγματος, έχουμε καλή πιθανότητα το δείγμα να είναι αντιπροσωπευτικό. Αυτό σημαίνει πως, για αρκούντως μεγάλο δείγμα, το αποτέλεσμα που θα δώσει ο αλγόριθμος ERM μπορεί με μεγάλη πιθανότητα να είναι σχεδόν βέλτιστο.

Διατυπώνοντας πιο αυστηρά τον προηγούμενο συλλογισμό, θεωρούμε $\delta, \epsilon \in (0,1)$, απαιτώντας $L_{(\mathcal{D},f)}(h_S) \leqslant \epsilon$ με πιθανότητα (παράμετρο εμπιστοσύνης) τουλάχιστον $1-\delta$. Αναζητούμε κάτω φράγμα για το μέγεθος του δείγματος ώστε να ισχύει το προηγούμενο. Έτσι, οδηγούμαστε στο ακόλουθο θεώρημα:

Θεώρημα 2.3.1. *Έστω $\mathcal{H}$ πεπερασμένη κλάση υποθέσεων. Ακόμη, έστω $\delta, \epsilon \in (0,1)$ και ακέραιος $m \geqslant \frac{\ln\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon}$. Τότε, για οποιαδήποτε συνάρτηση f και οποιαδήποτε κατανομή $\mathcal{D}$ που είναι τέτοιες ώστε να ικανοποιείται η παραδοχή ικανοποιησιμότητας, για κάθε i.i.d. δείγμα μεγέθους m, με πιθανότητα τουλάχιστον $1-\delta$, θα έχουμε για την υπόθεση h_S που επιστρέφει ο ERM ότι $L_{(\mathcal{D},f)}(h_S) \leqslant \epsilon$.*

Απόδειξη. Θέλουμε να βρούμε ένα κάτω φράγμα για το m, ώστε να εξασφαλίζεται πως η πιθανότητα να καταλήξουμε σε μη αντιπροσωπευτικό δείγμα (και ενδεχομένως σε κακή υπόθεση) δεν θα υπερβαίνει το δ. Αυτό ισοδυναμεί με το να ισχύει $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) < \delta$.

Θεωρούμε το σύνολο των κακών υποθέσεων $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$. Παρατηρούμε πως, για $h \in \mathcal{H}_B$, θα πρέπει να υπάρχει τουλάχιστον ένα δείγμα S για το οποίο να ισχύει $L_S(h) = 0$. Πράγματι, αν για κάποιο h δεν ίσχυε το προηγούμενο, δεν θα υπήρχε λόγος να διατηρηθεί στο $\mathcal{H}$ αφού, μετά την επιλογή του S, θα απορριπτόταν με πιθανότητα 1 για οποιοδήποτε S. Για το λόγο αυτό, ορίζουμε το σύνολο των μη αντιπροσωπευτικών δειγμάτων:

$$M = \{S|_x : \exists h \in \mathcal{H}_B, s.t.\ L_S(h) = 0\} = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

Παρατηρούμε πως, προκειμένου η h_S που θα προκύπτει από το δείγμα να είναι κακή υπόθεση, είναι απαραίτητο να ισχύει $L_S(h_S) = 0$. Άρα, έχουμε $\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M$.

Από τα προηγούμενα έχουμε:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leqslant \mathcal{D}^m(M) \leqslant \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) =$$

$$= \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : \forall i \in [m], h(x_i) = f(x_i)\})$$

Όμως, έχουμε υποθέσει πως η επιλογή των x_i είναι ανεξάρτητη, οπότε:

$$\mathcal{D}^m(\{S|_x : \forall i \in [m], h(x_i) = f(x_i)\}) = \prod_{i \in [m]} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$$

Όμως, $\mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) = 1 - \mathcal{D}(\{x_i : h(x_i) \neq f(x_i)\}) = 1 - L_{(\mathcal{D},f)}(h)$.

Στα προηγούμενα, όμως, έχουμε θεωρήσει πως $h \in \mathcal{H}_B$, οπότε $L_{(\mathcal{D},f)}(h) > \epsilon$, που δίνει:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) < \sum_{h \in \mathcal{H}_B} (1-\epsilon)^m = |\mathcal{H}_B|(1-\epsilon)^m \leqslant |\mathcal{H}|e^{-\epsilon m}$$

(αφού $|\mathcal{H}_B| \leqslant |\mathcal{H}|$ και $1 - \epsilon \leqslant e^{-\epsilon}$)

Απαιτούμε, λοιπόν, να ισχύει:

$$\boxed{|\mathcal{H}|e^{-\epsilon m} \leqslant \delta \Leftrightarrow m \geqslant \frac{\ln\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon}}$$

Η απόδειξη έχει ολοκληρωθεί. ■

Κεφάλαιο 3

Ένα τυπικό μοντέλο μάθησης

3.1 *PAC Learning*

Πλέον, είμαστε έτοιμοι να δώσουμε τον ορισμό του *PAC learnability* (*PAC* : *Probably Approximately Correct*).

Ορισμός 3.1.1. Μία κλάση $\mathcal{H}$ λέμε πως είναι *PAC learnable* αν υπάρχουν συνάρτηση $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ και αλγόριθμος μάθησης ώστε: Για κάθε $\epsilon, \delta \in (0,1)$ και για κάθε *labeling function* $f : \mathcal{X} \to \{0,1\}$ και κατανομή $\mathcal{D}$ πάνω στον $\mathcal{X}$ που είναι τέτοια ώστε η παραδοχή ικανοποιησιμότητας να ισχύει και θεωρώντας *i.i.d.* δείγμα μεγέθους $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ που έχει προκύψει από τη $\mathcal{D}$, εκτελώντας τον αλγόριθμο πάνω σε αυτό να προκύπτει μία υπόθεση $h \in \mathcal{H}$ που με πιθανότητα τουλάχιστον $1 - \delta$ να δίνει $L_{(\mathcal{D},f)}(h) \leqslant \epsilon$.

Ο παραπάνω ορισμός περιλαμβάνει δύο παραμέτρους. Αυτές είναι το δ και το ϵ. Δηλώνουν την μέγιστη πιθανότητα επιλογής μη αντιπροσωπευτικού δείγματος (παράμετρος αβεβαιότητας) και την επιτρεπτή απόκλιση από το σωστό αποτέλεσμα (παράμετρος σφάλματος), αντίστοιχα.

Ερμηνεύοντας το προηγούμενο θεώρημα υπό το πρίσμα του παραπάνω ορισμού, αντιλαμβανόμαστε πως αποδείξαμε ότι οι πεπερασμένες κλάσεις είναι *PAC learnable*.

Υπάρχει αναλογία με τα προσεγγιστικά σχήματα *PTAS* και *FPTAS*. Εδώ, όμως, δεν μας μας απασχολεί τόσο η χρονική πολυπλοκότητα, όσο η πολυπλοκότητα δείγματος (*sample complexity*), που εκφράζεται μέσω της $m_{\mathcal{H}}$. Η συνάρτηση αυτή δηλώνει ποιο είναι το ελάχιστο μέγεθος δείγματος που πρέπει να έχουμε προκειμένου να εξασφαλίσουμε ότι θα έχουμε το επιθυμητό αποτέλεσμα. Αποδεικνύεται πως η $m_{\mathcal{H}}$ είναι φθίνουσα ως προς καθεμία από τις μεταβλητές της, εφόσον η άλλη μεταβλητή παραμένει σταθερή (βλ. παράρτημα).

Με βάση τα προηγούμενα, έχουμε δείξει ουσιαστικά πως:

Πόρισμα 1. *Κάθε πεπερασμένη κλάση υποθέσεων $\mathcal{H}$ είναι PAC learnable με πολυπλοκότητα δείγματος:*

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \left\lceil \frac{\ln\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon} \right\rceil$$

3.2 Γενικεύσεις του *PAC Learning*

Θα εξετάσουμε τι δυνατότητες υπάρχουν για γενικότερα μοντέλα μηχανικής μάθησης, εφόσον αγνοηθούν κάποιες από τις παραδοχές που κάναμε πριν.

3.2.1 Πέρα από την παραδοχή ικανοποιησιμότητας

Η πρώτη παραδοχή που μπορεί να μην ισχύει είναι η παραδοχή ικανοποιησιμότητας. Ο λόγος είναι ότι, οι μεταβλητές με τις οποίες περιγράφονται τα στοιχεία του πεδίου $\mathcal{X}$ δεν είναι απαραίτητο να καθορίζουν πλήρως την τιμή που του αντιστοιχεί η f. Προκειμένου να ληφθεί υπόψη και αυτό το ενδεχόμενο, εισάγουμε έναν ακόμη μη ντετερμινιστικό παράγοντα στην περιγραφή μας.

Στο γενικότερο μοντέλο το οποίο θεωρούμε η κατανομή $\mathcal{D}$ παύει να είναι ορισμένη πάνω στο πεδίο $\mathcal{X}$ και ορίζεται πλέον στο $\mathcal{X} \times \mathcal{Y}$. Έτσι, η $\mathcal{D}$ δηλώνει πλέον την από κοινού κατανομή των στοιχείων του $\mathcal{X}$ και των τιμών του $\mathcal{Y}$, εφόσον η αντίστοιχη των μεν στις δε είναι μη ντετερμινιστική. Επιπλέον, ορίζουμε τις κατανομές $\mathcal{D}_x$ και $\mathcal{D}((x, y)|x)$. Η πρώτη είναι περιθώρια της $\mathcal{D}$ και εκφράζει την πιθανότητα να προκύψει ένα συγκεκριμένο στοιχείο του $\mathcal{X}$, ανεξαρτήτως του y που του αντιστοιχεί. Η δεύτερη εκφράζει τις πιθανότητες των διαφόρων $y \in \mathcal{Y}$ για δεδομένο $x \in \mathcal{X}$. Ισχύει $\mathcal{D} = \mathcal{D}_x \cdot \mathcal{D}((x, y)|x)$.

Σύμφωνα με τα προηγούμενα, η συνάρτηση f παύει πλέον να έχει νόημα και αντικαθίσταται ουσιαστικά από την $\mathcal{D}((x, y)|x)$. Γι' αυτό και χρειάζεται να αναδιατυπώσουμε κάποιους ορισμούς που δόθηκαν πριν, κυρίως αναφορικά με την αξιολόγηση της επιτυχίας των ταξινομητών (οι ταξινομητές εξακολουθούν να είναι συναρτήσεις $h : \mathcal{X} \to \{0, 1\}$).

Ορισμός 3.2.1. Ορίζουμε ως πραγματικό ρίσκο ενός ταξινομητή h την ποσότητα $L_{\mathcal{D}}(h) \stackrel{def}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{def}{=} \mathcal{D}(\{(x, y) : h(x) \neq y\})$ και ως εμπειρικό ρίσκο το $L_S(h) \stackrel{def}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$ (όπως πριν).

Εφόσον σε κάθε x μπορεί να αντιστοιχούν περισσότερες της μίας τιμές με διαφορετική πιθανότητα και, ταυτοχρόνως, το αποτέλεσμα της μάθησης παραμένει συνάρτηση $f : \mathcal{X} \to \{0, 1\}$ είναι αδύνατο να ισχύει πλέον $L_{\mathcal{D}}(f) = 0$. Ως εκ τούτου, βέλτιστος ταξινομητής f πλέον είναι αυτός για τον οποίον $L_{\mathcal{D}}(f) \leqslant L_{\mathcal{D}}(f'), \forall f' : \mathcal{X} \to \mathcal{Y}$.

Ορίζουμε τον βέλτιστο κατά *Bayes* ταξινομητή, για τον οποίο αποδεικνύεται ότι, με βάση τις συνθήκες που διαμορφώνονται σύμφωνα με τους προηγούμενους ορισμούς, είναι ο καλύτερος που μπορούμε να έχουμε (βλ. παράρτημα):

Ορισμός 3.2.2. Για οποιαδήποτε κατανομή $\mathcal{D}$ ορισμένη στο $\mathcal{X} \times \{0,1\}$, ο βέλτιστος ταξινομητής $f_{\mathcal{D}} : \mathcal{X} \to \{0,1\}$ είναι:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & if\ \mathbb{P}[y=1|x] \geqslant \frac{1}{2} \\ 0 & otherwise \end{cases}$$

Δεδομένου ότι δεν γνωρίζουμε την κατανομή $\mathcal{D}$, η εύρεση του προηγούμενου ταξινομητή είναι αδύνατη. Επιπλέον, δεν μας εγγυάται κανείς ότι $f_{\mathcal{D}} \in \mathcal{H}$. Γι' αυτό και επιθυμούμε να βρούμε έναν ταξινομητή που μάλλον περίπου (*probably approximately*) να ελαχιστοποιεί το πραγματικό ρίσκο. Αυτό μας οδηγεί στον ορισμό του *Agnostic PAC learnability*, που γενικεύει τον αντίστοιχο για το *PAC learnability*:

Ορισμός 3.2.3. Μία κλάση υποθέσεων $\mathcal{H}$ καλείται *Agnostic PAC learnable* αν υπάρχει συνάρτηση $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ και αλγόριθμος μάθησης ώστε: για κάθε $\epsilon, \delta \in (0,1)$ και για κάθε κατανομή $\mathcal{D}$ πάνω στο $\mathcal{X} \times \mathcal{Y}$, θεωρώντας ένα *i.i.d.* δείγμα μεγέθους $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ που έχει προκύψει από τη $\mathcal{D}$ και εφαρμόζοντας πάνω σε αυτό τον αλγόριθμο, να προκύπτει μία υπόθεση $h \in \mathcal{H}$, για την οποία με πιθανότητα τουλάχιστον $1-\delta$ να έχουμε $L_{\mathcal{D}}(h) \leqslant \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$.

3.2.2 Πέρα από δυαδικά προβλήματα ταξινόμησης

Μέχρι τώρα θεωρούσαμε πως $\mathcal{Y} = \{0,1\}$ και ορίζαμε ως πραγματικό ρίσκο το $L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. Αυτές οι επιλογές έχουν ιδιαίτερα περιοριστικό χαρακτήρα, αφού δεν μας επιτρέπουν να αντιμετωπίσουμε προβλήματα ταξινόμησης με περισσότερες των δύο κατηγοριών (*multiclass classification*). Επιπλέον, θεωρούσαμε δεδομένο πως μας ενδιαφέρει μόνο η πιθανότητα να αποδίδει ο ταξινομητής διαφορετική τιμή από την πραγματική και όχι ο βαθμός στον οποίο αποκλίνουν οι δύο τιμές. Αυτό καθιστούσε αδύνατη την αντιμετώπιση αρκετών προβλημάτων στατιστικής (πχ προβλήματα παλινδρόμησης).

Για τους προηγούμενους λόγους, επιλέγουμε αφενός να επιτρέψουμε το $\mathcal{Y}$ να αποτελείται από περισσότερες από δύο τιμές, αφετέρου να επιτρέψουμε γενικότερες μορφές συναρτήσεων ρίσκου.

Ορισμός 3.2.4. Έστω μία κλάση υποθέσεων $\mathcal{H}$ και ένα πεδίο Z. Οποιαδήποτε συνάρτηση $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ λέγεται συνάρτηση σφάλματος.

Για τα προβλήματα που αντιμετωπίζουμε (προβλήματα πρόβλεψης) έχουμε $Z = \mathcal{X} \times \mathcal{Y}$. Ακόμη:

Ορισμός 3.2.5. Για κατανομή $\mathcal{D}$ πάνω στον Z, ορίζουμε ως συνάρτηση ρίσκου ενός ταξινομητή h το αναμενόμενο σφάλμα $L_{\mathcal{D}}(h) \stackrel{def}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h,z)]$. Ακόμη, για δείγμα $S = (z_1, ..., z_m)$, ορίζουμε ως εμπειρικό ρίσκο την ποσότητα:

$$L_S(h) \stackrel{def}{=} \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$$

Διακρίνουμε τους ακόλουθους δύο τύπους προβλημάτων και ορίζουμε τις κατάλληλες συναρτήσεις σφάλματος για καθέναν από αυτούς:

- **Σφάλμα 0-1:** Ορίζουμε:

$$\ell_{0-1}(h,(x,y)) \stackrel{def}{=} \begin{cases} 0 & if\ h(x) = y \\ 1 & if\ h(x) \neq y \end{cases}$$

  Αυτή είναι η συνάρτηση σφάλματος που χρησιμοποιείται για προβλήματα ταξινόμησης. Παρατηρούμε πως, αν τα x, y είναι τυχαία, πρόκειται για τυχαία μεταβλητή *Bernoulli*. Έτσι, έχουμε:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell_{0-1}(h,(x,y))] = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[\ell_{0-1}(h,(x,y)) = 1] = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

  που είναι η συνάρτηση ρίσκου που χρησιμοποιούσαμε ως τώρα.

- **Τετραγωνικό Σφάλμα:** Ορίζουμε $\ell_{sq}(h,(x,y)) \stackrel{def}{=} (h(x) - y)^2$. Χρησιμοποιείται σε προβλήματα παλινδρόμησης.

Μένει να γενικεύσουμε τον ορισμό του *PAC Learning*, προκειμένου να καλύπτει και αυτές τις περιπτώσεις συναρτήσεων σφάλματος. Έχουμε:

Ορισμός 3.2.6. Μία κλάση υποθέσεων $\mathcal{H}$ λέγεται *Agnostic PAC learnable* με σεβασμό σε ένα πεδίο Z και μία συνάρτηση σφάλματος $\ell : \mathcal{X} \times Z \to \mathbb{R}_+$, εάν υπάρχει συνάρτηση $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ και αλγόριθμος μάθησης, τέτοιοι ώστε: για κάθε $\epsilon, \delta \in (0,1)$ και για κάθε κατανομή $\mathcal{D}$ πάνω στο Z, εκτελώντας τον αλγόριθμο μάθησης σε έναν *i.i.d.* δείγμα μεγέθους $m \geqslant m_{\mathcal{H}}(\epsilon,\delta)$ να επιστρέφει ένα $h \in \mathcal{H}$, τέτοιο ώστε $L_{\mathcal{D}}(h) \leqslant \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$ με πιθανότητα τουλάχιστον $1-\delta$, όπου $L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h,z)]$.

Μέρος II

Παράρτημα

Σε αυτό το παράρτημα δίνονται κάποιες προτάσεις που στο βιβλίο όπου βασίστηκε αυτή η παρουσίαση δίνονται ως ασκήσεις προς λύση. Οι αποδείξεις αυτών των προτάσεων είναι αρκετά διδακτικές και οδηγούν στην καλύτερη κατανόηση του υλικού.

Κεφάλαιο 2 Άσκηση 1

Αποδεικνούμε πως, ο ταξινομητής:

$$h(x) = \begin{cases} y_i & if\ \exists i \in [m], s.t.\ x_i = x \\ 0 & otherwise \end{cases}$$

μπορεί να αναπαρασταθεί με πολυώνυμα. Συγκεκριμένα, αποδεικνύουμε πως υπάρχει πολυώνυμο p_S d μεταβλητών, που για κάθε $\boldsymbol{X} \in \mathbb{R}^d$ ισχύει $p_S(\boldsymbol{X}) \geqslant 0$ αν και μόνο αν $h_S(\boldsymbol{X}) = 1$.

Απόδειξη. Καταρχάς, παρατηρούμε πως, αφού τα πολυώνυμα είναι συνεχείς συναρτήσεις, αν για κάποιο $\boldsymbol{X} \in \mathbb{R}^d$ ισχύει $p_S(\boldsymbol{X}) > 0$, θα υπάρχουν άπειρα τέτοια $\boldsymbol{X}$. Όμως, τα σημεία στα οποία θέλουμε $p_S(\boldsymbol{X}) \geqslant 0$ είναι πεπερασμένα (είναι μόνο τα στοιχεία του $S|_x$ που τους αντιστοιχεί το 1). Ως εκ τούτου, θέλουμε ένα μη θετικό πολυώνυμο, το οποίο να μηδενίζεται μόνο για τα $\boldsymbol{X} \in \mathbb{R}^d$ για τα οποία $h_S(\boldsymbol{X}) = 1$. Έστω $\boldsymbol{r_1, ..., r_k} \in \mathbb{R}^d, k \leqslant m$ τα στοιχεία του S για τα οποία έχουμε $h_S(\boldsymbol{r_i}) = 1$. Αυτά γράφονται στη μορφή $\boldsymbol{r_i} = (r_{1,i}, ..., r_{d,i})$. Αν $\boldsymbol{X} = (X_1, ..., X_d) \in \mathbb{R}^d$, θεωρούμε το πολυώνυμο:

$$p_S(\boldsymbol{X}) = -\prod_{i=1}^{k} \left[\sum_{j=1}^{d} (X_j - r_{j,i})^2 \right]$$

Γραφόμενο αναλυτικά, το προηγούμενο πολυώνυμο δίνει:

$$p_S(\boldsymbol{X}) = -[(X_1 - r_{1,1})^2 + ... + (X_d - r_{d,1})^2]...[(X_1 - r_{1,k})^2 + ... + (X_d - r_{d,k})^2]$$

Κάθε παράγοντας του γινομένου είναι μη αρνητικός, αφού όλοι οι όροι είναι τετράγωνα. Προκειμένου δε να μηδενιστεί το πολυώνυμο, πρέπει και αρκεί $\boldsymbol{X} = (r_{1,i}, ..., r_{d,i}) = \boldsymbol{r_i}$, για κάποιο $i \in [k]$. Τα προηγούμενα, σε συνδυασμό με το αρνητικό πρόσημο, μας εξασφαλίζουν πως $p_S(\boldsymbol{X}) \leqslant 0, \forall \boldsymbol{X} \in \mathbb{R}^d$, με την ισότητα να ισχύει μόνο για τα στοιχεία του δείγματος όπου αντιστοιχεί η τιμή 1. Άρα:

$$\boxed{h_S(\boldsymbol{X}) = \mathbb{1}\{p_S(\boldsymbol{X}) \geqslant 0\}, \forall \boldsymbol{X} \in \mathbb{R}^d}$$

όπου με $\mathbb{1}\{A\}$ συμβολίζουμε τη δείκτρια συνάρτηση του συνόλου Α.

■

Κεφάλαιο 2 Άσκηση 2

Αποδεικνούμε πως $\mathbb{P}[L_S(h_S) \xrightarrow[m\to\infty]{} L_{(\mathcal{D},f)}(h_S)] = 1$.

Απόδειξη. Από τον ορισμό του εμπειρικού ρίσκου, είναι:

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}$$

Παρατηρούμε πως:

$$\mathop{\mathbb{P}}_{x_i \sim \mathcal{D}}[\mathbb{1}\{h(x_i) \neq y_i\} = 1] = \mathop{\mathbb{P}}_{x_i \sim \mathcal{D}}[h(x_i) \neq f(x_i)] = L_{(\mathcal{D},f)}(h)$$

Άρα, οι $\mathbb{1}\{h(x_i) \neq y_i\}$ είναι *i.i.d.* μεταβλητές *Bernoulli* με πιθανότητα επιτυχίας $L_{(\mathcal{D},f)}(h)$ και το εμπειρικό ρίσκο αποτελεί τον εμπειρικό τους μέσο. Αυτό μας εξασφαλίζει ότι $\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_S(h_S)] = L_{(\mathcal{D},f)}(h_S)$. Εφαρμόζοντας δε τον ισχυρό νόμο των μεγάλων αριθμών έχουμε:

$$\boxed{\mathbb{P}[L_S(h_S) \xrightarrow[m\to\infty]{} L_{(\mathcal{D},f)}(h_S)] = 1}$$

γεγονός που ολοκληρώνει την απόδειξη.

■

Κεφάλαιο 3 Άσκηση 1

Αποδεικνύουμε πως η συνάρτηση που εκφράζει την πολυπλοκότητα είναι αύξουσα και ως προς τις δύο μεταβλητές της.

Απόδειξη. Υποθέτουμε, για δοσμένο $\delta \in (0,1)$, πως υπάρχουν $\epsilon_1, \epsilon_2 \in (0,1)$ με $\epsilon_1 < \epsilon_2$, τέτοια ώστε $m_{\mathcal{H}}(\epsilon_1, \delta) < m_{\mathcal{H}}(\epsilon_2, \delta)$. Υποθέτουμε πως εκτελούμε τον αλγόριθμο μάθησης που έχουμε για το συγκεκριμένο πρόβλημα. Σε αυτή την περίπτωση, εκτελώντας τον αλγόριθμο μάθησης που έχουμε συσχετίσει με το συγκεκριμένο πρόβλημα, προκύπτουν δύο λύσεις. Στην πρώτη, για *i.i.d.* δείγμα μεγέθους $m_1 = m_{\mathcal{H}}(\epsilon_1, \delta)$ έχουμε με πιθανότητα τουλάχιστον $1-\delta$ ότι $L_{(\mathcal{D},f)}(h_S) \leqslant \epsilon_1$. Στην δεύτερη, για *i.i.d.* δείγμα μεγέθους $m_2 = m_{\mathcal{H}}(\epsilon_2, \delta) > m_{\mathcal{H}}(\epsilon_1, \delta)$ έχουμε με πιθανότητα τουλάχιστον $1-\delta$ ότι $L_{(\mathcal{D},f)}(h_S) \leqslant \epsilon_2$. Όμως, παρατηρούμε πως $L_{(\mathcal{D},f)}(h_{\mathcal{S}}) \leqslant \epsilon_1 < \epsilon_2 \Rightarrow L_{(\mathcal{D},f)}(h_S) < \epsilon_2$. Συνεπώς, στην πρώτη λύση, η τιμή του πραγματικού σφάλματος φράσσεται άνω από την ίδια τιμή που έχουμε στην δεύτερη λύση και το μέγεθος του δείγματος είναι μικρότερο. Όμως, οι τιμές της συνάρτησης $m_{\mathcal{H}}$ εκφράζουν την πολυπλοκότητα δείγματος του προβλήματος, γεγονός που αποκλείει να υπάρχει λύση στο δεύτερο πρόβλημα με μέγεθος δείγματος μικρότερο από m_2. Έτσι, καταλήγουμε σε άτοπο. Άρα, $m_{\mathcal{H}}(\epsilon_2, \delta) \leqslant m_{\mathcal{H}}(\epsilon_1, \delta)$.

Αντίστοιχα προκύπτει και η μονοτονία ως προς τη δεύτερη μεταβλητή.

■

Κεφάλαιο 3 Άσκηση 2

Έστω $\mathcal{X}$ διακριτό πεδίο και $\mathcal{H}_{singleton} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, όπου $h_z(x) \equiv \mathbb{1}\{x = z\}$ και $h^- \equiv 0$. Θεωρώντας πως έχουμε κατανομή $\mathcal{D}$ και σωστή *labelling function* f που μαζί με την $\mathcal{H}_{singleton}$ να ικανοποιούν την παραδοχή ικανοποιησιμότητας, αυτό σημαίνει πως το πολύ σε ένα στοιχείο του $\mathcal{X}$ μπορεί να αντιστοιχεί η τιμή 1. Διατυπώνουμε αλγόριθμο βασιζόμενο στην αρχή ERM και δείχνουμε πως η παραπάνω κλάση υποθέσεων είναι *PAC learnable*.

Απόδειξη. Ο αλγόριθμος που διατυπώνουμε είναι ο ακόλουθος: αν υπάρχει στοιχείο z στο δείγμα όπου να αντιστοιχεί η τιμή 1, τότε επίστρεψε τον ταξινομητή h_z. Στην αντίθετη περίπτωση, επίστρεψε τον h^-. Σύμφωνα με όσα αναφέρθηκαν παραπάνω σχετικά με την παραδοχή ικανοποιησιμότητας, οι λύσεις που επιστρέφονται από τον αλγόριθμο αυτό έχουν μηδενικό εμπειρικό ρίσκο. Επομένως, υλοποιείται η αρχή ERM.

Προκειμένου, τώρα, να δείξουμε πως η κλάση αυτή είναι $PAC\ learnable$, πρέπει να βρούμε ένα κάτω φράγμα για το μέγεθος του δείγματος ώστε για $\epsilon, \delta \in (0,1)$, να έχουμε:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) < \delta$$

Αντίστοιχα με την απόδειξη για τις πεπερασμένες κλάσεις, ορίζουμε:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$$

$$M = \{S|_x : \exists h \in \mathcal{H}_B, s.t.\ L_S(h) = 0\}$$

Είναι:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leqslant \mathcal{D}^m(M)$$

Ένα δείγμα είναι μη αντιπροσωπευτικό όταν υπάρχει z στο οποίο αντιστοιχεί η τιμή 1, αλλά δεν υπάρχει στοιχείο του δείγματος ίσο με αυτό. Επομένως, έχουμε:

$$\mathcal{D}^m(M) = \mathcal{D}^m(\{S|_x : \forall i \in [m], x_i \neq z\}) = \prod_{i\in[m]} \mathcal{D}(\{x_i : x_i \neq z\})$$

Όμως, για το πραγματικό ρίσκο ισχύει:

$$L_{(\mathcal{D},f)}(h_S) = \mathcal{D}(\{x : h_S(x) \neq f(x)\}) = \mathcal{D}(\{x : x = z\})$$

(σημειώνεται πως, με βάση την προηγούμενη παρατήρηση, για να έχει νόημα η απόδειξη, πρέπει $\mathcal{D}(\{x : x = z\}) > \epsilon$, αλλιώς είναι $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) = 0$ και το ζητούμενο ισχύει για όλα τα $\delta > 0$ ανεξαρτήτως δείγματος)

Άρα, έχουμε:

$$\mathcal{D}^m(M) = \prod_{i\in[m]} (1 - L_{(\mathcal{D},f)}(h_S)) < (1-\epsilon)^m \leqslant e^{-\epsilon m}$$

Απαιτούμε, λοιπόν:

$$\boxed{e^{-\epsilon m} \leqslant \delta \Leftrightarrow m \geqslant \frac{\ln(\frac{1}{\delta})}{\epsilon}}$$

Ισοδύναμα, λέμε πως η πολυπλοκότητα δείγματος του προβλήματος είναι:

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \left\lceil \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon} \right\rceil$$

Η απόδειξη έχει ολοκληρωθεί.

■

Κεφάλαιο 3 Άσκηση 5

Θεωρούμε πεδίο $\mathcal{X}$ και $\mathcal{D}_1, ..., \mathcal{D}_m$ ακολουθία κατανομών πάνω σε αυτό. Ορίζουμε την κατανομή $\bar{\mathcal{D}}_m = \frac{\mathcal{D}_1 + ... + \mathcal{D}_m}{m}$. Θεωρούμε πεπερασμένη κλάση υποθέσεων $\mathcal{H}$ και σωστή $labelling\ function\ f : \mathcal{X} \to \{0,1\}$. Ακόμη, θεωρούμε δείγμα

μεγέθους m. Η επιλογή των στοιχείων του πραγματοποιείται ανεξάρτητα, αλλά όχι με βάση την ίδια κατανομή. Συγκεκριμένα, θεωρούμε πως $x_i \sim \mathcal{D}_i$. Έστω παράμετρος σφάλματος $\epsilon \in (0,1)$. Αποδεικνύουμε πως ισχύει:

$$\mathbb{P}[\exists h \in \mathcal{H} : L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \wedge L_S(h) = 0] \leqslant |\mathcal{H}| e^{-\epsilon m}$$

Απόδειξη. Αρχικά, αναφορικά με τις συναρτήσεις ρίσκου, ισχύει ότι:

$$L_{(\bar{\mathcal{D}}_m, f)}(h) = \bar{\mathcal{D}}_m(\{x : h(x) \neq f(x)\}) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{D}_i(\{x : h(x) \neq f(x)\}) =$$

$$= \frac{1}{m}\sum_{i=1}^{m} L_{(\mathcal{D}_i, f)}(h)$$

Το ενδεχόμενο την πιθανότητα του οποίου θέλουμε να φράξουμε γράφεται:

$$\{\exists h \in \mathcal{H} : L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \wedge L_S(h) = 0\} = \bigcup_{h \in \mathcal{H}} \{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \wedge L_S(h) = 0\}$$

Συνεπώς, έχουμε:

$$\mathbb{P}[\exists h \in \mathcal{H} : L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \wedge L_S(h) = 0] \leqslant$$

$$\leqslant \sum_{h \in \mathcal{H}} \mathbb{P}[L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \wedge L_S(h) = 0] =$$

$$= \sum_{h \in \mathcal{H}} \mathbb{P}[L_S(h) = 0]\mathbb{P}[L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon | L_S(h) = 0]$$

Παρατηρούμε πως, για κάθε μεμονωμένο στοιχείο h της κλάσης υποθέσεων, ή ισχύει ότι $L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon$, ή ισχύει το αντίθετο. Τα προηγούμενα δε ανεξαρτήτως του ενδεχομένου $\{L_S(h) = 0\}$, με το οποίο δεσμεύουμε το $\{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon\}$ στην παραπάνω έκφραση. Συνεπώς, έχουμε

$$\mathbb{P}[L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon | L_S(h) = 0] = \mathbb{1}\{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon\}$$

Επομένως, το ζητούμενο γράφεται:

$$\sum_{h \in \mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon\}\mathbb{P}[L_S(h) = 0] =$$

$$= \sum_{h \in \mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon\}\left[\prod_{i=1}^{m} \mathcal{D}_i(\{x : h(x) = f(x)\})\right] =$$

$$= \sum_{h \in \mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon\}\left[\prod_{i=1}^{m} \left(1 - L_{(\mathcal{D}_i, f)}(h)\right)\right]$$

Με εφαρμογή της ανισότητας αριθμητικού-γεωμετρικού μέσου έχουμε:

$$\prod_{i=1}^{m} \left(1 - L_{(\mathcal{D}_i, f)}(h)\right) \leqslant \left[\frac{1}{m}\sum_{i=1}^{m} \left(1 - L_{(\mathcal{D}_i, f)}(h)\right)\right]^m = [1 - L_{(\bar{\mathcal{D}}_m, f)}(h)]^m$$

(η τελευταία ισότητα ισχύει λόγω της σχέσης για τις συναρτήσεις ρίσκου που αποδείξαμε στην αρχή).

Επομένως, έχουμε δείξει προς το παρόν ότι:

$$\mathbb{P}[\exists h \in \mathcal{H} : L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon \wedge L_S(h) = 0] \leqslant$$

$$\leqslant \sum_{h\in\mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon\}[1 - L_{(\bar{\mathcal{D}}_m,f)}(h)]^m$$

Παρατηρούμε πως, οι όροι για τους οποίους δεν ισχύει ότι $L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon$ δεν συνεισφέρουν στο άθροισμα, λόγω της δείκτριας. Ως εκ τούτου, προκύπτει το άνω φράγμα:

$$\sum_{h\in\mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon\}[1 - L_{(\bar{\mathcal{D}}_m,f)}(h)]^m < \sum_{h\in\mathcal{H}} \mathbb{1}\{L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon\}(1-\epsilon)^m \leqslant$$

$$\leqslant \sum_{h\in\mathcal{H}} (1-\epsilon)^m = |\mathcal{H}|(1-\epsilon)^m$$

Άρα, λοιπόν, έχουμε, τελικά:

$$\boxed{\mathbb{P}[\exists h \in \mathcal{H} : L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon \wedge L_S(h) = 0] < |\mathcal{H}|(1-\epsilon)^m \leqslant |\mathcal{H}|e^{-\epsilon m}}$$

Το ζητούμενο έχει αποδειχθεί.

■

Κεφάλαιο 3 Άσκηση 7

Εδώ αποδεικνύουμε πως ο ταξινομητής $Bayes$ είναι ο καλύτερος που μπορούμε να έχουμε για δυαδικά προβλήματα ταξινόμησης με συναρτήσεις σφάλματος 0-1 (η απόδειξη που δίνεται είναι για διακριτό χώρο $\mathcal{X}$).

Απόδειξη. Θεωρούμε έναν αυθαίρετο ταξινομητή $h : \mathcal{X} \to \{0, 1\}$. Από τον ορισμό του πραγματικού σφάλματος και τον τύπο ολικής πιθανότητας έχουμε:

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] = \sum_{x^*\in\mathcal{X}} \mathbb{P}[x = x^*] \mathop{\mathbb{P}}_{y\sim\mathcal{D}((x^*,y)|x^*)}[h(x^*) \neq y]$$

Επιπλέον, έχουμε:

$$\mathop{\mathbb{P}}_{y\sim\mathcal{D}((x,y)|x^*)}[h(x^*) \neq y] =$$
$$= \mathbb{P}[y = 0|x = x^*]\mathbb{P}[h(x^*) \neq 0] + \mathbb{P}[y = 1|x = x^*]\mathbb{P}[h(x^*) \neq 1] =$$
$$= \mathbb{P}[y = 0|x = x^*]\mathbb{P}[h(x^*) = 1] + \mathbb{P}[y = 1|x = x^*]\mathbb{P}[h(x^*) = 0]$$

Παρατηρούμε πως, η τιμή του $h(x^*)$ καθορίζεται ντετερμινιστικά. Ως εκ τούτου, ακριβώς ένα εκ των $\mathbb{P}[h(x^*) = 0], \mathbb{P}[h(x^*) = 1]$ είναι 0 και το άλλο είναι 1. Προφανώς, συμφέρει να είναι 1 αυτό που πολλαπλασιάζεται με το μικρότερο από τα $\mathbb{P}[y = 1|x = x^*], \mathbb{P}[y = 0|x = x^*]$. Έτσι, οδηγούμαστε στον ταξινομητή:

$$\boxed{f_{\mathcal{D}}(x) = \begin{cases} 1 & if\ \mathbb{P}[y = 1|x] \geqslant \frac{1}{2} \\ 0 & otherwise \end{cases}}$$

Η απόδειξη έχει ολοκληρωθεί.

■