Εκπαίδευση μοντέλου πρόβλεψης υπερπαραμέτρων αλγορίθμου knn

Νησιώτη Ελένη
9 Φεβρουαρίου 2017

1 Automl: Ρύθμιση Μοντέλου

Η ρύθμιση ενός μοντέλου μηχανικής μάθησης συνίσταται στην εύρεση των βέλτιστων υπερπαραμέτρων, διαδικασία που υπόκειται στην κρίση του αναλυτή και εμπειρικά αποδεδειγμένα επηρεάζει την απόδοση του τελικού μοντέλου. Η συνηθέστερη τεχνική, που έχει υιοθετήσει η κοινότητα της μηχανικής μάθησης είναι η αναζήτηση των υπερπαραμέτρων σε ένα προκαθορισμένο πλέγμα και η επιλογή των βέλτιστων με cross-validation. Οι Bergstra et al. [1] απέδειξαν ότι αυτή η τεχνική είναι υποδεέστερη της τυχαίας αναζήτησης, καθώς αν και οι υπερπαράμετροι κυμαίνονται σε μεγάλο εύρος, μόνο περιορισμένες τιμές επηρεάζουν τη συμπεριφορά του αλγορίθμου, με αποτέλεσμα να δημιουργούνται υποπεριοχές ενδιαφέροντος πρακτικά μη εντοπίσιμες μόνο με πλεγματική αναζήτηση.

Η επιστήμη του Automl έχει αναγνωρίσει το στάδιο της ρύθμισης μοντέλου ως επιδεχόμενο αυτοματοποίησης λόγω της σημασίας και της χρονικής και υπολογιστικής απαιτητικότητάς του. Ο κύριος άξονας της βιβλιογραφίας είναι η υιοθέτηση τεχνικών βελτιστοποίησης συναρτήσεων μαύρου κουτιού με προσαρμογή τους στις ιδιαιτερότητες των προβλημάτων μηχανικής μάθησης. Παραδείγματα αυτής της προσπάθειας αποτελούν οι βιβλιοθήκες Hyperopt [9], Spearmint [10], HPOlib [8], οι οποίες υλοποιούν συναρτήσεις SMBO, η χρήση γενετικών αλγορίθμων [11] και τεχνικών βασισμένων σε gradient descent [14].

Στόχος μας είναι η εύρεση των βέλτιστων υπερπαραμέτρων να γίνεται εξολοκλήρου με ένα μοντέλο πρόβλεψης και όχι με αναζήτηση. Προς αυτό το σκοπό θα εκπαιδευτεί ένα μοντέλο πρόβλεψης της υπερπαραμέτρου k ενός knn με χρήση μετα-

χαρακτηριστικών,το οποίο ονομάζουμε HPP μοντέλο. Η πιθανή και διαπιστωμένη [6] αδυναμία εκπαίδευσης ενός ικανοποιητικά ακριβούς μοντέλου μπορεί να αντιμετωπιστεί με την πρόβλεψη, αντί για μία τιμής, ενός διαστήματος βέλτιστων υπερπαραμέτρων, προερχόμενου από τα διαστήματα πρόβλεψης (*prediction intervals*) του μοντέλου HPP. Στη συνέχεια και με βάση αυτό το διάστημα θα εκπαιδεύεται ένα ensemble μοντέλων με την τεχνική του model selection [3]. Κίνητρα για αυτή την προσπαθεια αποτελούν τόσο η επιτάχυνση της διαδικασίας ρύθμισης ενός μοντέλου όσο και η δημιουργία μετα-γνώσης για τη συνάρτηση μεταξύ βέλτιστων υπερπαραμέτρων και φύσης του προβλήματος (χαρακτηριστικά του σετ δεδομένων και του αλγορίθμου μηχανικής μάθησης).

2 Πειράματα

Στόχος των πειραμάτων είναι ο καθορισμός του HPP μοντέλου. Στην ενότητα 3 θα αξιολογήσουμε την απόδοση του συστήματος, που χρησιμοποιεί το μοντέλο για τη ρύθμιση ενός knn και εκπαιδεύει ένα ensemle knn μοντέλων.

2.1 Προετοιμασία

Για τη διεξαγωγή των πειραμάτων συνέλεξα 124 σετ δεδομένων από διάφορες πηγές: UCI [17], kaggle [12], openml [13], data.world [4], Relational Dataset Repository [16]. Πρόκειται για σετ δεδομένων δυαδικής ταξινόμησης με πολύ ετερογενή χαρακτηριστικά (πλήθος και είδος χαρακτηριστικών, πλήθος παραδειγμάτων, τύπος αρχείου, παρουσία και αναπαράσταση άγνωστων τιμών, ισορροπία κλάσης). Η διαδικασία που χρησιμοποιήθηκε για τον καθαρισμό τους, ώστε να υπάρχει μια κοινή διεπαφή για το πείραμα ήταν:

- μετατροπή αρχείων σε comma-delimited .csv. Τα πηγαία αρχεία βρίσκονταν σε μορφές .csv, .txt, .xlsx, .arff, .mysql και είχαν κωδικοποίηση utf-8 ή utf-16.
- καθορισμός κλάσης. Στη πλειοψηφία των περιπτώσεων η κλάση αναγνωριζόταν χειροκίνητα από την περιγραφή του σετ δεδομένων. Συλλέχθηκαν και σετ δεδομένων που ήταν πολλαπλής ταξινόμησης και παλινδρόμησης. Στην πρώτη περίπτωση έγινε αντιστοίχηση σε δύο ουσιώδεις κλάσεις, ενώ στη δεύτερη βρέθηκε η μέση τιμή της μεταβλητής κλάσης και χρησιμοποιήθηκε ως κατώφλι για το διαχωρισμό των παραδειγμάτων σε δύο κλάσεις.
- αναγνώριση άγνωστων τιμών. Στα αρχεία που περιείχαν άγνωστες τιμές χρησιμοποιούνταν διάφοροι συμβολισμοί (”?”, ”*”, ””) οι οποίοι αντικαταστάθηκαν από κενά, ώστε να αναγνωρίζονται από την R ως NAs.
- αναγνώριση ημερομηνιών. Συγκεντρώθηκαν τα ονόματα των χαρακτηριστικών που αναφέρονταν σε ημερομηνίες, ώστε να διαβαστούν σωστά από την R.

Στη συνέχεια παράχθηκαν τα απαραίτητα μετα-δεδομένα για τα πειράματα. Για κάθε σετ δεδομένων τα μετα-χαρακτηριστικά υπολογίστηκαν με χρήση του πακέτου mfExtractor (που έγραψα εγώ), ενώ η βέλτιστη παράμετρος k βρέθηκε με χρήση του λογισμικού HPOlib [8], με τον αλγορίθμο βελτιστοποίησης Tree Parzen Estimator για ένα μοντέλο knn της βιβλιοθήκης caret και πλέγμα αναζήτησης $K = \{k \mid 1 \leq k \leq 10\}$.

Για τα πειράματα επιλογής αλγορίθμου το σετ δεδομένων χωρίστηκε σε 58 σετ δεδομένων για εκπαίδευση του HPP μοντέλου και 17 για έλεγχό του, καθώς και τον έλεγχο του τελικού συστήματος.

2.2 Ανάλυση μεταχαρακτηριστικών

Η βιβλιογραφία προσφέρει μια εκτεταμένη λίστα μετα-χαρακτηριστικών που μπορούν να χρησιμοποιηθούν για το χαρακτηρισμό ενός σετ δεδομένων. Οι δημιουργοί του autosklearn αναφέρουν τόσο το σύνολο των μετα-χαρακτηριστικών που δοκίμασαν [6] όσο και την τελική λίστα αυτών που συμπεριλήφθηκαν στο τελικό μοντέλο [7]. Οι Reif, Shafait, and Dengel [15] προτείνουν την τεχνική των μετα-μετα-χαρακτηριστικών. Στα πειράματά μας δοκιμάσαμε και τα δύο σετ και καταλήξαμε σε αυτό της ομάδας του autosklearn χωρίς τη χρήση landmarking μετα-χαρακτηριστικών . Καθώς το συγκεκριμένο σύνολο δεν αντιπροσωπεύει τα σετ δεδομένων που αποτελούνται μόνο από κατηγορικά χαρακτηριστικά, στόχος μας είναι αυτό το σύνολο να εμπλουτιστεί ώστε να αντικατοπτρίζει όλα τα πιθανά σετ δεδομένων.

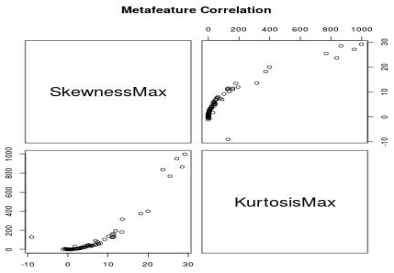
Μετα-χαρακτηριστικά
Κλάσμα χαρακτηριστικών για 95% διακύμανση των PCA
Κυρτότητα πρώτης PCA συνιστώσας
Ασσυμετρία πρώτης PCA συνιστώσας
Ελάχιστη ασυμμετρία
Μέγιστη ασυμμετρία
Μέση τιμή ασυμμετρίας
Τυπική απόκλιση ασυμμετρίας
Ελάχιστη κυρτότητα
Μέγιστη κυρτότητα
Μέση τιμή κυρτότητας
Τυπική απόκλιση κυρτότητας

Πίνακας 2.1: Τελική λίστα μετα-χαρακτηριστικών

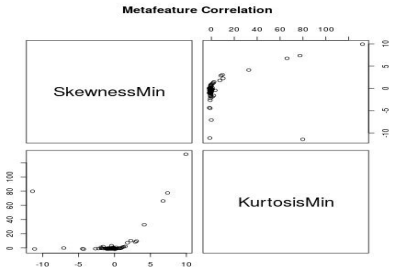
Η ανάλυση των μετα-χαρακτηριστικών του Σχήματος 2.2 οδήγησε στα ακόλουθα συμπεράσματα:

- υπάρχει μια μη-γραμμική συσχέτιση μεταξύ των μετα-χαρακτηριστικών που περιγράφουν όμοιες ιδιότητες της ασυμμετρίας και κυρτότητας (Σχήματα 2.1i, 2.1ii)

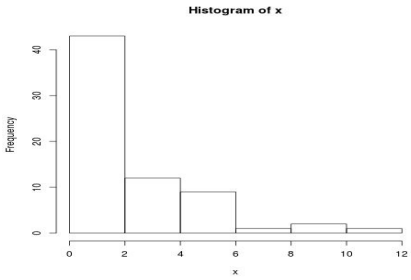
- όλα τα χαρακτηριστικά εκτός από τη Μέση τιμή ασυμμετρίας έχουν ξεκάθαρα μη-συμμετρική κατανομή (Σχήματα 2.1iii, 2.1iv)
- υπάρχουν πολλές εξωκείμενες τιμές (Σχήματα 2.1v, 2.1vi)



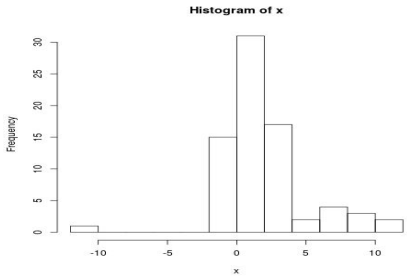
(i) Λογαριθμική συσχέτιση μεταξύ μέγιστων τιμών κυρτότητας και ασυμμετρίας



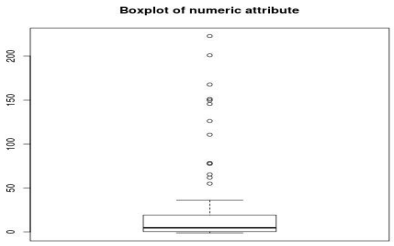
(ii) Λογαριθμική συσχέτιση μεταξύ ελάχιστων τιμών κυρτότητας και ασυμμετρίας



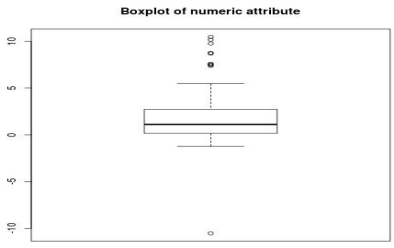
(iii) Θετική ασυμμετρία για την τυπική απόκλιση της ασυμμετρίας, όπως και για τα περισσότερα χαρακτηριστικά



(iv) Κατανομή κοντινή σε κανονική για τη μέση τιμή της ασυμμετρίας



(v) Εμφάνιση πολλών εξωκείμενων τιμών για τη μέση τιμή της κυρτότητας



(vi) Η κανονική κατανομή της μέσης τιμής της ασσυμετρίας παρουσιάζει επίσης εξωκείμενες τιμές

Σχήμα 2.1: Οπτικοποίηση μετα-χαρακτηριστικών

2.3 Επιλογή αλγορίθμου

Αρχικά εξετάστηκε η γραμμικότητα του προβλήματος με χρήση του μοντέλου lm της caret. Κατά την εκπαίδευση ενός μοντέλου παλινδρόμησης υπάρχει η απαίτηση κανονικής κατανομής και σταθερής διακύμανσης των residuals. Σε αντίθετη περίπτωση η χρήση των prediction intervals δεν είναι ορθή. Για να αντιμετωπιστεί αυτό χρησιμοποιούνται μετασχηματισμοί δύναμης, ένας εκ των οποίων είναι ο μετασχηματισμός Box-Cox. Εισήχθη από τους Box and Cox [2] και ορίζεται ως

$$y^{'} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{2.1}$$

Ο μετασχηματισμός αυτός μετατρέπει skewed κατανομές σε κανονικές και κατ' επέκταση εξασφαλίζει γραμμικότητα των χαρακτηριστικών με τη κλάση. Το βέλτιστο λ για ένα σετ δεδομένων υπολογίζεται με χρήση profile likelihood διαγραμμάτων, δηλαδή επιλέγεται το λ που μεγιστοποιεί την πιθανότητα τα residuals να έχουν κανονική κατανομή με την εφαρμογή του μετασχηματισμού.

Επίσης, εξετάστηκε η χρήση ενός svm, προκειμένου να διαπιστωθεί η αναγκαιότητα ενός πολυπλοκότερου μοντέλου από το γραμμικό. Χρησιμοποιήθηκε το μοντέλο RadialSvm της βιβλιοθήκης caret.

2.4 Περιγραφή ensemble

Καθώς η απαίτηση ακριβούς πρόβλεψης της βέλτιστης τιμής μιας υπερπαραμέτρου κρίνεται, τουλάχιστον με τα τρέχοντα χαρακτηριστικά, υπερβολικά απαιτητική, όπως σχολίασαν και οι Feurer, Springenberg, and Hutter [6], οι οποίοι αρκέστηκαν στη χρήση των προβλέψεων τους για warmstart αλγορίθμων βελτιστοποίησης, θα χρειαστεί περαιτέρω επεξεργασία του μοντέλου. Προς αυτό το σκοπό εκμεταλλεύτηκα τα διαστήματα πρόβλεψης που παράγονται από ένα γραμμικό μοντέλο και είναι προσβάσιμα μέσω της *caret*, ώστε να ορίσω ένα σύνολο βέλτιστων k για κάθε σετ δεδομένων και να δημιουργήσω έναν ensemble με αυτά. Το σύνολο αυτό ορίζεται ως τα ακέραια k που βρίσκονται στο 90% διάστημα εμπιστοσύνης της πρόβλεψης. Αν η βέλτιστη τιμή βρίσκεται μέσα σε αυτό το διάστημα τότε με χρήση του ensemble θεωρητικά θα εξασφαλιστεί αποτέλεσμα ισάξιο με ένα μοντέλο που θα προέβλεπε επακριβώς τη βέλτιστη υπερπαράμετρο.

Για τον svm η εξαγωγή των prediction intervals δεν είναι αυτοματοποιημένη. Η τεχνική που ακολουθήθηκε ήταν η εξής: υπολογισμός της διακύμανσης των residuals, εύρεση του e 90% ποσοστημορίου μιας κανονικής κατανομής με μέση τιμή μ τις προβλέψεις του μοντέλου και διακύμανση ίση με των residuals και δημιουργία του διαστήματος ως $[\mu - e, \mu + e]$

2.5 Αποτελέσματα

2.5.1 Επιλογή βέλτιστου μοντέλου HPP

Για την εύρεση του καλύτερου αλγορίθμου χρησιμοποιήθηκαν στατιστικά τεστ. Συγκεκριμένα σύγκρινα τα σφάλματα μεταξύ των μοντέλων *lm* με μετασχηματισμό Box-Cox, *RadialSvm* και του βελτιστοποιημένου μοντέλου που παρήχθηκε από την *HPOlib* στο 20% των σετ δεδομένων με χρήση του paired Wilcoxon-rank-sum τεστ σε επίπεδο εμπιστοσύνης 95% για την παρατήρηση στατιστικά σημαντικών διαφορών. Ως σφάλμα ορίζω $e = 1 - accuracy$

Μέθοδοι		Υπόθεση		
Μέθοδος 1	Μέθοδος 2	Δίπλευρη	Αριστερή	Δεξία
svm	HPOlib	0.0673	0.971	0.0381
lm + BoxCox	HPOlib	0.1353	0.9406	0.0673
svm	lm + BoxCox	0.09058	0.9608	0.04529

Πίνακας 2.2: Στατιστική σύγκριση μεθόδων

Με βάση τα παραπάνω στοιχεία συμπεραίνω πως:

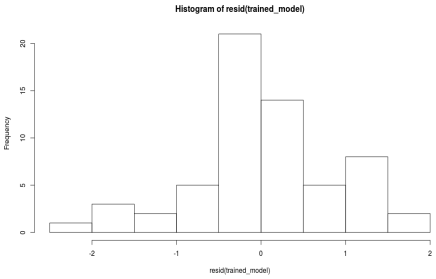
- Δεν μπορώ να απορρίψω την υπόθεση της μηδενικής διαφοράς μεταξύ οποιοδήποτε μεθόδων.
- Μπορώ οριακά να απορρίψω τη μονόπλευρη υπόθεση ότι το μοντέλο svm έχει μεγαλύτερο σφάλμα από το βέλτιστο.
- Επίσης οριακά απορρίπτω την υπόθεση ότι το μοντέλο svm έχει μεγαλύτερο σφάλμα από το μοντέλο lm.

Για τα τελικά πειράματα θα επιλέξω το μοντέλο με μετασχηματισμό Box-Cox, καθώς παρουσιάζει οριακά χειρότερη συμπεριφορά από το svm, αλλά είναι απλούστερο.

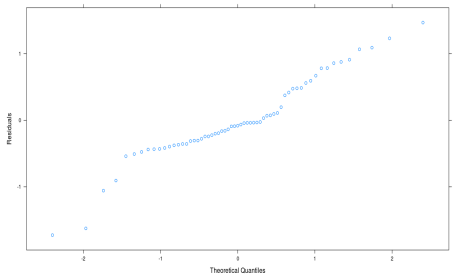
2.5.2 Αξιολόγηση HPP μοντέλου

Στη συνέχεια αξιολογούμε την απόδοση του HPP μοντέλου. Η ρίζα του μέσου τετραγωνικού σφάλματος, υπολογισμένη με loocv είναι 2.56. Στα σχήματα φαίνεται ότι τα residuals του τελικού μοντέλου ακολουθούν σχετικά κανονική κατανομή, επομένως η χρήση των prediction intervals είναι έγκυρη.

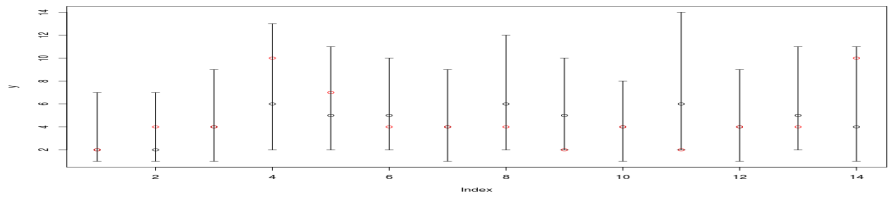
Εφαρμόζωντας την τεχνική με τα prediction intervals στο σετ ελέγχου παρατηρούμε ότι πάντα συμπεριλαμβάνουμε τη βέλτιστη λύση



Σχήμα 2.2: Ιστόγραμμα residuals



Σχήμα 2.3: Q-Q διάγραμμα



Σχήμα 2.4: Διάγραμμα διασποράς προβλέψεων με prediction intervals

3 Προτεινόμενο σύστημα

Έχοντας εκπαιδεύσει το HPP μοντέλο, θα αξιολογήσουμε το σύστημά μας για τα 17 σετ δεδομένων ελέγχου. Για το καθένα το σύστημα εξάγει τα μετα-χαρακτηριστικά, προβλέπει τα βέλτιστα k με χρήση του HPP μοντέλου και σχηματίζει τον ensemble με την τεχνική του model selection. Η απόδοση του ensemble αξιολογείται ως η ακρίβεια με την τεχνική 80% – 20% holdout.

Διαθέτοντας την επιτευχθείσα ακρίβεια και τις ακρίβειες που είχε πετύχει η HPOlib κατά τη βελτιστοποίηση των υπερπαραμέτρων μπορούμε να ερευνήσουμε αν οι δύο μέθοδοι έχουν σημαντική στατιστική διαφορά. Προς αυτό το σκοπό εφαρμόζουμε το paired Wilcoxon-rank-sum τεστ με επίπεδο εμπιστοσύνης 95%, σύμφωνα με το οποίο η υπόθεση ότι το προτεινόμενο σύστημα έχει χαμηλότερη ακρίβεια απορρίπτεται με p-value 0.02864.

Τα διαγράμματα προφίλ απόδοσης (performance profile plots) [5] αποτελούν ένα εργαλείο αξιολόγησης και σύγκρισης της απόδοσης εργαλείων βελτιστοποίησης. Χρησιμοποιούνται σε περιπτώσεις εφαρμογής διαφορετικών τεχνικών βελτιστοποίησης σε ένα σύνολο προβλημάτων ως εναλλακτική απεικόνιση εκτενών πινάκων, μιας συνηθισμένης και προβληματικής λύσης. Το προφίλ απόδοσης είναι η αθροιστική συνάρτηση κατανομής μιας τεχνικής για μία μετρική απόδοσης.

Ως μετρική απόδοσης ορίζουμε το λόγο της απόδοσης της τρέχουσας τεχνικής προς τη μεγαλύτερη απόδοση που επιτεύχθηκε από οποιαδήποτε τεχνική για ένα συγκε-

κριμένο σετ δεδομένων, δηλαδή

$$r_{p,s} = \frac{t_{p,s}}{\max\{t_{p,s} : s \in S\}} \tag{3.1}$$

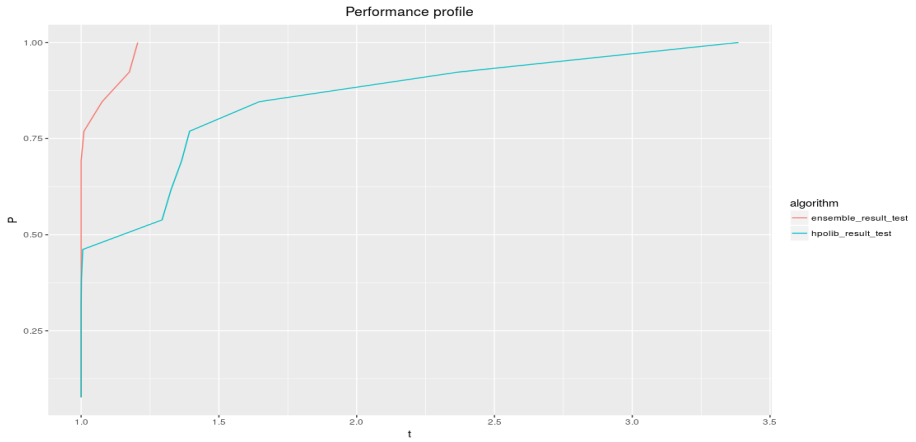
όπου r ο λόγος απόδοσης, t η ακρίβεια, p το σετ δεδομένων και s η τεχνική.

Το διάγραμμα απεικονίζει τη τιμή

$$\rho_\tau = \frac{size\{p \in P : r_{p,s} \leq \tau\}}{n_p} \tag{3.2}$$

η οποία εκφράζει την πιθανότητα μία τεχνική να βρίσκεται σε απόσταση τ από τον καλύτερο λόγο απόδοσης. Επομένως το σημείο $\tau = 1$ εκφράζει τη πιθανότητα μία τεχνική να είναι η βέλτιστη.

Από το διάγραμμα προφίλ απόδοσης που ακολουθεί μπορούμε να συμπεραίνουμε πως το δικό μας μοντέλο είναι πιθανότερα το βέλτιστο.



Σχήμα 3.1: Διάγραμμα απόδοσης για σύγκριση χρήσης προτεινόμενου μοντέλου με βελτιστοποιημένου HPOlib

Βιβλιογραφία

[1] James Bergstra et al. “Algorithms for Hyper-parameter Optimization”. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS’11. Granada, Spain: Curran Associates Inc., 2011, pp. 2546–2554. ISBN: 978-1-61839-599-3.

[2] G. E. P. Box and D. R. Cox. “An Analysis of Transformations”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2 (1964), pp. 211–252. ISSN: 00359246.

[3] Rich Caruana et al. “Ensemble Selection from Libraries of Models”. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: ACM, 2004, pp. 18–. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015432.

[4] *data.world*. 2016. URL: https://data.world/ (visited on 02/06/2017).

[5] Elizabeth D. Dolan and Jorge J. Moré. “Benchmarking optimization software with performance profiles”. In: *Mathematical Programming* 91.2 (2002), pp. 201–213. ISSN: 1436-4646. DOI: 10.1007/s101070100263.

[6] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. “Using Meta-learning to Initialize Bayesian Optimization of Hyperparameters”. In: *Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection - Volume 1201*. MLAS’14. Prague, Czech Republic: CEUR-WS.org, 2014, pp. 3–10. ISBN: 1613-0073.

[7] github. *autosklearn*. 2016. URL: https://github.com/automl/auto-sklearn (visited on 02/06/2017).

[8] github. *HPOlib*. 2016. URL: https://github.com/automl/HPOlib (visited on 02/06/2017).

[9] github. *hyperopt*. 2016. URL: https://github.com/hyperopt/hyperopt (visited on 02/06/2017).

[10] github. *Spearmint*. 2016. URL: https://github.com/HIPS/Spearmintt (visited on 02/06/2017).

[11] Frank Hutter et al. “An Experimental Investigation of Model-based Parameter Optimisation: SPO and Beyond”. In: *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*. GECCO ’09. Montreal, Québec, Canada: ACM, 2009, pp. 271–278. ISBN: 978-1-60558-325-9. DOI: 10.1145/1569901.1569940.

[12] *Kaggle*. 2016. URL: https://www.kaggle.com/datasets (visited on 02/06/2017).

[13] *OpenML*. 2016. URL: https://openml.org/ (visited on 02/06/2017).

[14] Fabian Pedregosa. *Hyperparameter optimization with approximate gradient*. Version 1.

[15] Matthias Reif, Faisal Shafait, and Andreas Dengel. *Meta 2-Features: Providing Meta-Learners More Information*.

[16] *Relational Dataset Repository*. 2016. URL: https://relational.fit.cvut.cz/ (visited on 02/06/2017).

[17] *UCI*. 2016. URL: https://archive.ics.uci.edu/ml/datasets.html (visited on 02/06/2017).

Γλωσσάρι

HPP	Hyperparameter Prediction. 2, 3, 6, 7
loocv	leave-one-out validation. 6
SMBO	Sequential model-based Optimization. 1