

---

# Εκπαίδευση μοντέλου πρόβλεψης υπερπαραμέτρων αλγορίθμου knn

---

Νησιώτη Ελένη  
6 Φεβρουαρίου 2017

## 1 Η βελτιστοποίηση στο AUTOML

Η βελτιστοποίηση των υπερπαραμέτρων ενός μοντέλου μηχανικής μάθησης αποτελεί υπολογιστικά και χρονικά απαιτητική διαδικασία. Η επιστήμη του Automl έχει επικεντρώσει την προσπάθειά της σε αυτοματοποιημένες τεχνικές ελαχιστοποίησης συναρτήσεων κόστους (Hyperopt, Spearmint). Η θεωρία της μεταμάθησης έχει χρησιμοποιηθεί με διαφορετικές τεχνικές όπως [2], [6].

Στόχος μας είναι η πλήρης αντικατάσταση της διαδικασίας αναζήτησης με πρόβλεψη βάσει των μετα-χαρακτηριστικών. Η αδυναμία ακριβούς πρόβλεψης μπορεί να εξουδετερωθεί εκπαιδεύοντας μοντέλα με παραμέτρους που προκύπτουν από το μοντέλο βάσει των διαστημάτων πρόβλεψης (prediction intervals) και χρησιμοποιώντας την τεχνική του ensemble selection [1].

## 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 [2] όσο και την τελική λίστα αυτών που συμπεριλήφθηκαν στο τελικό μοντέλο [3]. Οι Reif, Shafait, and Dengel [5] 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

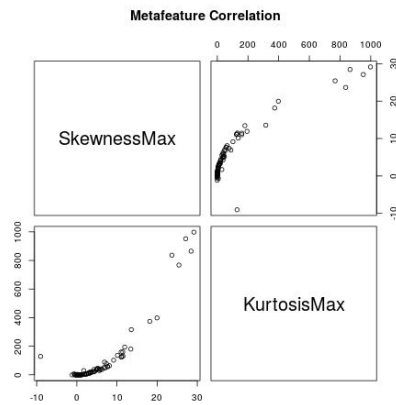
. Καθώς το συγκεκριμένο σύνολο δεν αντιπροσωπεύει τα σετ δεδομένων που αποτελούνται μόνο από κατηγορικά χαρακτηριστικά, στόχος μας είναι αυτό το σύνολο να εμπλουτιστεί ώστε να αντικατοπτρίζει όλα τα πιθανά σετ δεδομένων.

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

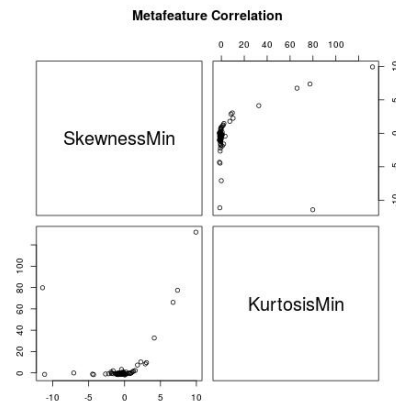
Πίνακας 2.1: Τελική λίστα μετα-χαρακτηριστικών

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

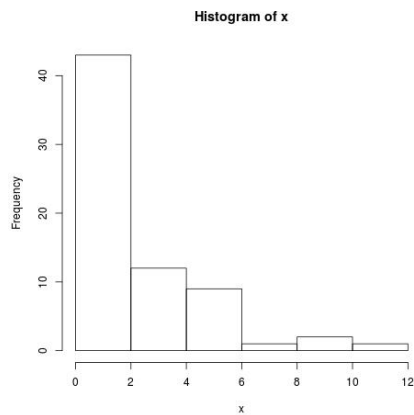
- υπάρχει μια μη-γραμμική συσχέτιση μεταξύ των μετα-χαρακτηριστικών που περιγράφουν όμοιες ιδιότητες της ασυμμετρίας και κυρτότητας (Σχήματα 2.1i, 2.1ii )
- όλα τα χαρακτηριστικά εκτός από τη Μέση τιμή ασυμμετρίας έχουν ξεκάθαρα μη-συμμετρική κατανομή (Σχήματα 2.1iii, 2.1iv )
- υπάρχουν πολλές εξωκείμενες τιμές ( Σχήματα 2.1v, 2.1vi)



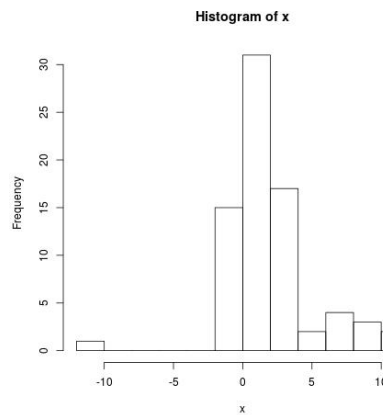
(i) Λογαριθμική συσχέτιση μεταξύ μέγιστων τιμών κυρτότητας και ασυμμετρίας



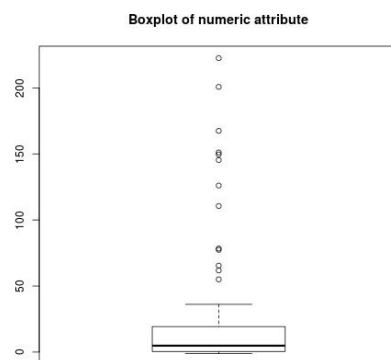
(ii) Λογαριθμική συσχέτιση μεταξύ ελάχιστων τιμών κυρτότητας και ασυμμετρίας



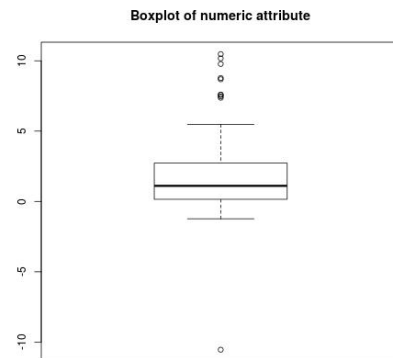
(iii) Θετική ασυμμετρία για την τυπική απόκλιση της ασυμμετρίας, όπως και για τα περισσότερα χαρακτηριστικά



(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 ασυμμετρίας παρουσιάζει επίσης εξωκείμενες τιμές

Σχήμα 2.1: Οπτικοποίηση μετα-χαρακτηριστικών

## 2.1 Επιλογή αλγορίθμου

Η αποθήκη σετ δεδομένων περιλαμβάνει 123 σετ δεδομένων, εκ των οποίων χρησιμοποιήθηκαν 75, καθώς τα υπόλοιπα είχαν μόνο κατηγορικά χαρακτηριστικά. Για αυτά έγινε εξαγωγή των μετα-χαρακτηριστικών και βελτιστοποίηση της υπερπαραμέτρου  $k$  του μοντέλου *knn* του πακέτου *caret* με χρήση του λογισμικού *HPOlib* [4] σε ένα πλέγμα αναζήτησης  $K = \{k \mid 1 \leq k \leq 10\}$  και τη τεχνική βελτιστοποίησης Tree Parzen Estimator.

Αρχικά εκπαιδεύτηκε ένα μοντέλο γραμμικής παλινδρόμησης με χρήση του μοντέλου *lm* για αξιολόγηση της γραμμικότητας του προβλήματος. Στη συνέχεια ενσωμάτωσα στο μοντέλο μετασχηματισμό *box-cox* λόγω των ασυμμετριών που είχαν παρατηρηθεί κατά την ανάλυση των μεταχαρακτηριστικών. Καθώς το αποτέλεσμα συνέχιζε να μην είναι ικανοποιητικό και παρατηρούνταν μοτίβα στο διάγραμμα διασποράς των *residuals* εξετάστηκε η χρησιμότητα ενός ισχυρότερου μοντέλου, του *RadialSvm*, το οποίο υλοποιεί έναν *svm* με ακτινική συνάρτηση πυρήνα.

Χρήση ENSEMBLE με βάση τα διαστήματα πρόβλεψης. Καθώς η απαίτηση ακριβούς πρόβλεψης της βέλτιστης τιμής μιας υπερπαραμέτρου κρίνεται, τουλάχιστον με τα τρέχοντα χαρακτηριστικά, υπερβολικά απαιτητική, όπως σχολίασαν και οι Feuer, Springenberg, and Hutter [2], 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 95% 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 υπερπαραμέτρο. Προϋπόθεση εγκυρότητας της παραπάνω μεθόδου είναι τα *residuals* να ακολουθούν κανονική κατανομή.

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  $e$  95% ποσοστημορίου μιας κανονικής κατανομής με μέση τιμή  $\mu$  τις προβλέψεις του μοντέλου και διακύμανση ίση με των *residuals* και δημιουργία του διαστήματος ως  $[\mu - e, \mu + e]$

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 παρήχθη 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  $e = 1 - accuracy$

Με βάση τα παραπάνω στοιχεία συμπεραίνω πως:

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

Πίνακας 2.2: Στατιστική σύγκριση μεθόδων

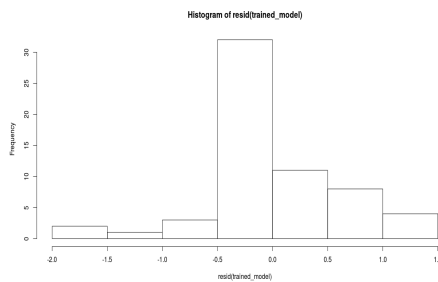
- Δεν μπορώ να απορρίψω την υπόθεση της μηδενικής διαφοράς μεταξύ οποιοδήποτε μεθόδων.
- Μπορώ οριακά να απορρίψω τη μονόπλευρη υπόθεση ότι το μοντέλο svm έχει μεγαλύτερο σφάλμα από το βέλτιστο.
- Επίσης οριακά απορρίπτω την υπόθεση ότι το μοντέλο svm έχει μεγαλύτερο σφάλμα από το μοντέλο lm.

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

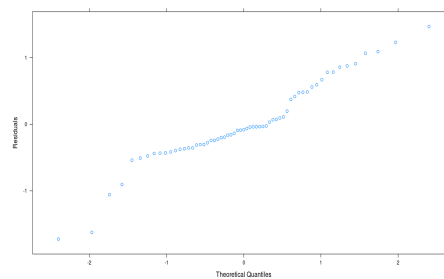
### 3 Περιγραφή τελικού μοντέλου

#### 3.1 Αξιολόγηση μοντέλου πρόβλεψης βελτιστοποιημένων κ

Στη συνέχεια αξιολογούμε την ικανότητα του μοντέλου lm με boxcox μετασχηματισμό. Στα σχήματα φαίνεται ότι τα residuals του τελικού μοντέλου ακολουθούν σχετικώς κανονική κατανομή, επομένως η χρήση των prediction intervals είναι έγκυρη.

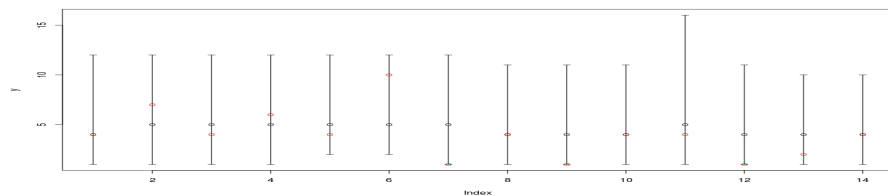


Σχήμα 3.1: Ιστόγραμμα residuals



Σχήμα 3.2: Q-Q διάγραμμα

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



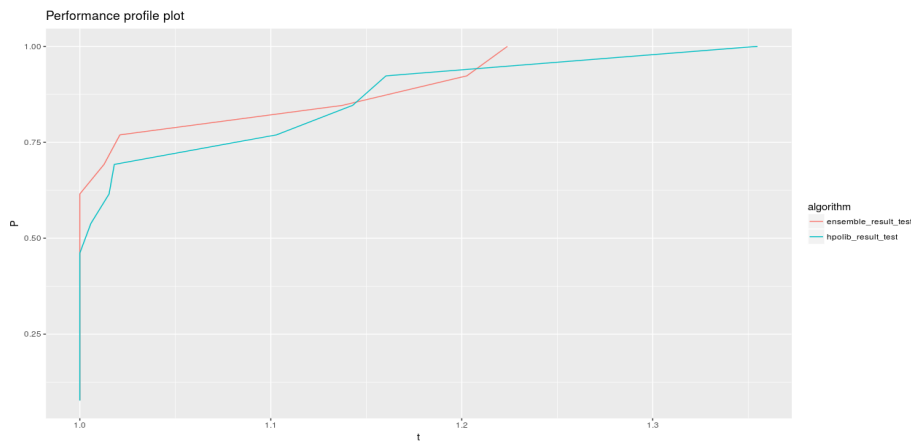
Σχήμα 3.3: Διάγραμμα διασποράς προβλέψεων με prediction intervals

### 3.2 Αξιολόγηση ENSEMBLE

Η αξιολόγηση του ensemble που παράγεται με βάση το βέλτιστο μοντέλο πρόβλεψης βελτιστοποιημένων  $k$  και τη τεχνική model selection. Εφαρμόζεται στο 20% του αρχικού σετ δεδομένων, καθώς η εκπαίδευση έγινε στο υπόλοιπο 80%. Μετρήθηκε η ακρίβεια (accuracy) και συγκρίθηκε με αυτή που παρήχθηκε από την HPOLib (η οποία αντιστοιχεί στην ακρίβεια με βελτιστοποιημένο  $k$ ).

Η μηδενική υπόθεση δεν απορρίφθηκε με paired Wilcoxon-rank-sum τεστ, καθώς το p-value ήταν 0.6949.

Ακολουθεί το διάγραμμα απόδοσης (*performance plot*) [Dolan2002] από το οποίο συμπεραίνουμε πως το δικό μας μοντέλο είναι πιθανότερα το βέλτιστο.



Σχήμα 3.4: Διάγραμμα απόδοσης για σύγκριση χρήσης προτεινόμενου μοντέλου με βελτιστοποιημένου HPOLib

## Βιβλιογραφία

- [1] 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.
- [2] 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.
- [3] github. *autosklearn*. 2016. URL: <https://github.com/automl/HPOLib> (visited on 02/06/2017).
- [4] github. *HPOLib*. 2016. URL: <https://github.com/automl/auto-sklearn> (visited on 02/06/2017).
- [5] Matthias Reif, Faisal Shafait, and Andreas Dengel. *Meta 2-Features: Providing Meta-Learners More Information*.
- [6] Carlos Soares, Pavel B. Brazdil, and Petr Kuba. “A Meta-Learning Method to Select the Kernel Width in Support Vector Regression”. In: *Machine Learning* 54.3 (2004), pp. 195–209. ISSN: 1573-0565. DOI: 10.1023/B:MACH.0000015879.28004.9b.