ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ενότητα 7 – Supervised Learning
Decision Trees (Δέντρα αποφάσεων)
Part 2

Μέθοδοι Μηχανικής Μάθησης στα Χρηματοοικονομικά

Αθανάσιος Σάκκας, Επ. Καθηγητής, ΟΠΑ

- Στην ενότητα αυτή θα επικεντρωθούμε στην πρόβλεψη όταν η target μεταβλητή είναι συνεχής μεταβλητή (continuous variable).

- Στην περίπτωση αυτή κατασκευάζουμε ένα δέντρο όπου αντί να μεγιστοποιήσουμε το αναμενόμενο κέρδος πληροφοριών (expected information gain), μεγιστοποιούμε την αναμενόμενη μείωση στο μέσο τετράγωνο σφάλμα (δηλαδή διακύμανση (the expected decrease in mean squared error (i.e. variance). Μαθηματικά ελαχιστοποιούμε το

$$Prob(X \geq Z) \times (mse\ if\ (X \geq Z)) + Prob(X < Z) \times (mse\ if\ (X < Z))$$

- Θα εξετάσουμε το παράδειγμα τιμής σπιτιού στην Αϊόβα (δεκαετίες '000) **Εφαρμογή Iowa House Price** που είχαμε δει στην ενότητα 5 και λαμβάνουμε υπόψη μόνο τη συνολική ποιότητα (Overall Quality) και την περιοχή διαβίωσης (Living Area).

- Μετά από data cleaning καταλήγουμε σε 2908 σπίτια και θα χρησιμοποιήσουμε 47 χαρακτηριστικά.
- Χρησιμοποιούμε 1800 παρατηρήσεις στο training set, 600 στο validation set και 508 στο test set.
- Πρέπει να βρούμε το χαρακτηριστικό που θα μπει στη ρίζα του δέντρου και επειδή και τα 2 χαρακτηριστικά είναι συνεχείς μεταβλητές, πρέπει να υπολογίσουμε και το threshold.
- Και για τα δύο χαρακτηριστικά υπολογίζουμε το βέλτιστο threshold.
- Το αναμενόμενο mse είναι μικρότερο για το χαρακτηριστικό συνολική ποιότητα (Overall Quality) και έχει ένα βέλτιστο threshold = 7.5.
- Άρα το χαρακτηριστικό συνολική ποιότητα (Overall Quality) τοποθετείται στη ρίζα του δέντρου. Αναλυτικά δείτε τον παρακάτω πίνακα.

Για τη ρίζα του δέντρου έχουμε τον παρακάτω πίνακα

Feature	Threshold, Q	No. Obs ≤Q	mse of Obs ≤Q	No. obs > Q	mse obs >Q	Exp mse
Overall Quality	7.5	1,512	2,376	288	7,312	3,166
Living Area	1,482.5	949	1,451	851	6,824	3,991

Στη συνέχεια μπορούμε να μεταβούμε στο δεύτερο επίπεδο του δέντρου

Overall Quality ≤ 7.5

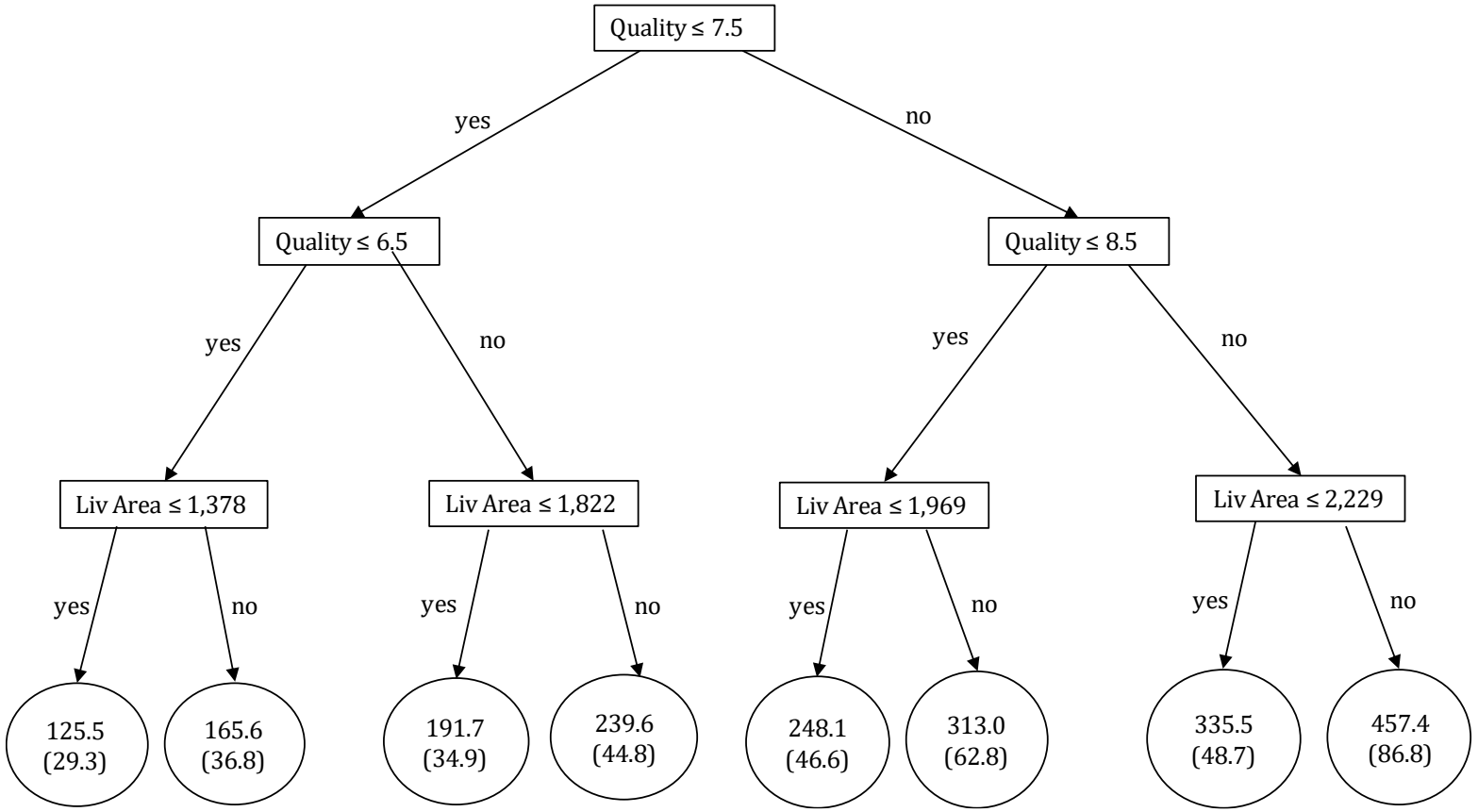
Feature	*Q*	*No. of obs ≤ Q*	*mse of obs ≤ Q*	*No. of obs > Q*	*mse of obs> Q*	*E(mse)*
Overall Quality	6.5	1,122	1,433	390	1,939	1,564
Living (sq. ft.)	1,412	814	1,109	698	2,198	1,612

Overall Quality > 7.5

Feature	*Q*	*No. of obs ≤ Q*	*mse of obs ≤ Q*	*No. of obs > Q*	*mse of obs> Q*	*E(mse)*
Overall Quality	8.5	214	3,857	74	8,043	4,933
Living (sq. ft.)	1,971	165	3,012	123	8,426	5, 324

- Η παραπάνω ανάλυση βρίσκεται στο *7. IOWA_Decision tree.xlsx* και *στο 7. IOWA_DecisionTree_Python*

The Tree



Ensemble Learning

- Γενικότερα, τα αποτελέσματα από πολλούς διαφορετικούς αλγόριθμους ML μπορούν να συνδυαστούν για να ληφθεί μια ενιαία εκτίμηση.
- Πολλοί αδύναμοι μαθητές (weak learners) μπορούν μερικές φορές να συνδυαστούν σε έναν δυνατό μαθητή (weak learner).
- Ο βαθμός στον οποίο αυτό είναι δυνατό εξαρτάται από τους συσχετισμούς μεταξύ των μαθητών (learners).

I. Bagging

- Δείγμα με αντικατάσταση για τη δημιουργία νέων datasets.
- Χρησιμοποιήστε μεθόδους ψηφοφορίας (voting) ή υπολογισμού του μέσου όρου (averaging) για την τελική εκτίμηση

II. Random Forest

- Αυτό περιλαμβάνει την κατασκευή πολλών δέντρων για παράδειγμα:
  - ❖Χρήση δειγμάτων bootstrapped από τα αρχικά δεδομένα.
  - ❖Χρησιμοποιώντας ένα τυχαίο υποσύνολο χαρακτηριστικών σε κάθε κόμβο.
  - ❖Τυχαιοποίηση ορίων thresholds κατά κάποιο τρόπο.

- Η τελική απόφαση μπορεί να είναι πλειοψηφία (majority vote) ή σταθμισμένη πλειοψηφία (weighted majority vote). Οι σταθμίσεις μπορούν να αντικατοπτρίζουν εκτιμήσεις πιθανοτήτων (όταν είναι διαθέσιμες) ή στοιχεία από ένα σύνολο test data set.

III. Boosting

- Οι προβλέψεις γίνονται διαδοχικά sequentially, κάθε μία προσπαθεί να διορθώσει το προηγούμενο σφάλμα.

- Μία προσέγγιση (AdaBoost) αυξάνει το βάρος που δίνεται σε εσφαλμένες ταξινομημένες παρατηρήσεις.

- Μια άλλη προσέγγιση (Gradient boosting) προσπαθεί να προσαρμόσει έναν νέο προγνωστικό παράγοντα στο σφάλμα που έκανε ο προηγούμενος προγνωστικός παράγοντας.