



Ενότητα 9 – Supervised Learning *Decision Trees (Δέντρα αποφάσεων)* *Part 2*

Μέθοδοι Μηχανικής Μάθησης στα Χρηματοοικονομικά

Αθανάσιος Σάκκας, 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 τετράγωνικό σφάλμα (mean squared error), δηλαδή τη διακύμανση (variance). Μαθηματικά ελαχιστοποιούμε το

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

- Θα εξετάσουμε το παράδειγμα τιμής σπιτιού στην Αϊόβα (δεκαετίες '000) **Εφαρμογή Iowa House Price** 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 $\leq Q$	mse of Obs $\leq Q$	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

<i>Feature</i>	<i>Q</i>	<i>No. of obs $\leq Q$</i>	<i>mse of obs $\leq Q$</i>	<i>No. of obs $> Q$</i>	<i>mse of obs $> Q$</i>	<i>E(mse)</i>
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

<i>Feature</i>	<i>Q</i>	<i>No. of obs $\leq Q$</i>	<i>mse of obs $\leq Q$</i>	<i>No. of obs $> Q$</i>	<i>mse of obs $> Q$</i>	<i>E(mse)</i>
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 *9. IOWA_Decision tree.xlsx* και στο *9. IOWA_DecisionTree_Python*

The Tree

