Μάθηση βάσει Πληροφορίας - Μέρος 3

Θορυβώδη Δεδομένα, Υπερπροσαρμογή και Κλάδεμα Δέντρων (Pruning)

- Στην περίπτωση ενός δέντρου απόφασης, η υπερπροσαρμογή περιλαμβάνει τη διαίρεση των δεδομένων με βάση ένα άσχετο χαρακτηριστικό.

Η πιθανότητα εμφάνισης υπερπροσαρμογής αυξάνεται όσο το δέντρο βαθαίνει, επειδή οι ταξινομήσεις που προκύπτουν βασίζονται σε όλο και μικρότερα υποσύνολα καθώς το σύνολο δεδομένων διαμερίζεται μετά από κάθε έλεγχο χαρακτηριστικού στη διαδρομή.

- **Προ-κλάδεμα (Pre-Pruning)**: πρόωρη διακοπή της αναδρομικής διαμέρισης. Το προ-κλάδεμα είναι επίσης γνωστό ως **εμπρόσθιο κλάδεμα** (forward pruning).

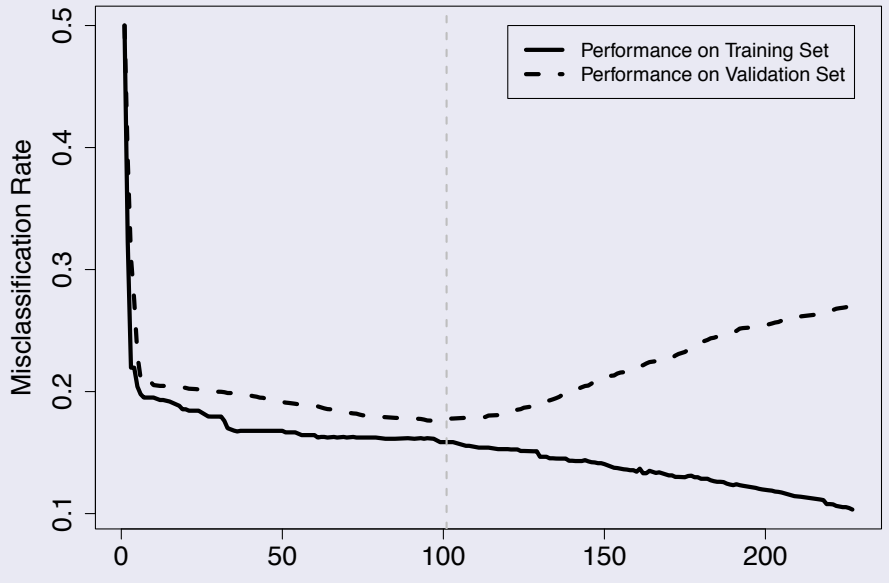
Συνήθεις προσεγγίσεις προ-κλαδέματος

1. **πρόωρη διακοπή**
2. **κλάδεμα** χ^2

- **Μετα-κλάδεμα (post pruning)**: επιτρέπουμε στον αλγόριθμο να αναπτύξει το δέντρο όσο θέλει και στη συνέχεια κλαδεύουμε τα κλαδιά που προκαλούν υπερπροσαρμογή.

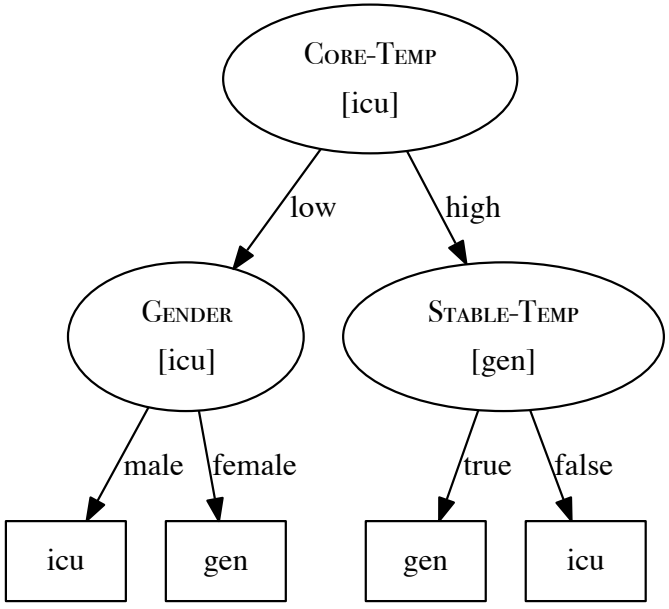
Συνήθης προσέγγιση μετα-κλαδέματος

- Χρησιμοποιώντας το σύνολο επικύρωσης, αξιολογούμε την ακρίβεια πρόβλεψης τόσο του πλήρως αναπτυγμένου δέντρου όσο και του κλαδεμένου αντιγράφου του. Αν το κλαδεμένο αντίγραφο δεν αποδίδει χειρότερα από το πλήρως αναπτυγμένο δέντρο, ο κόμβος είναι υποψήφιος για κλάδεμα.

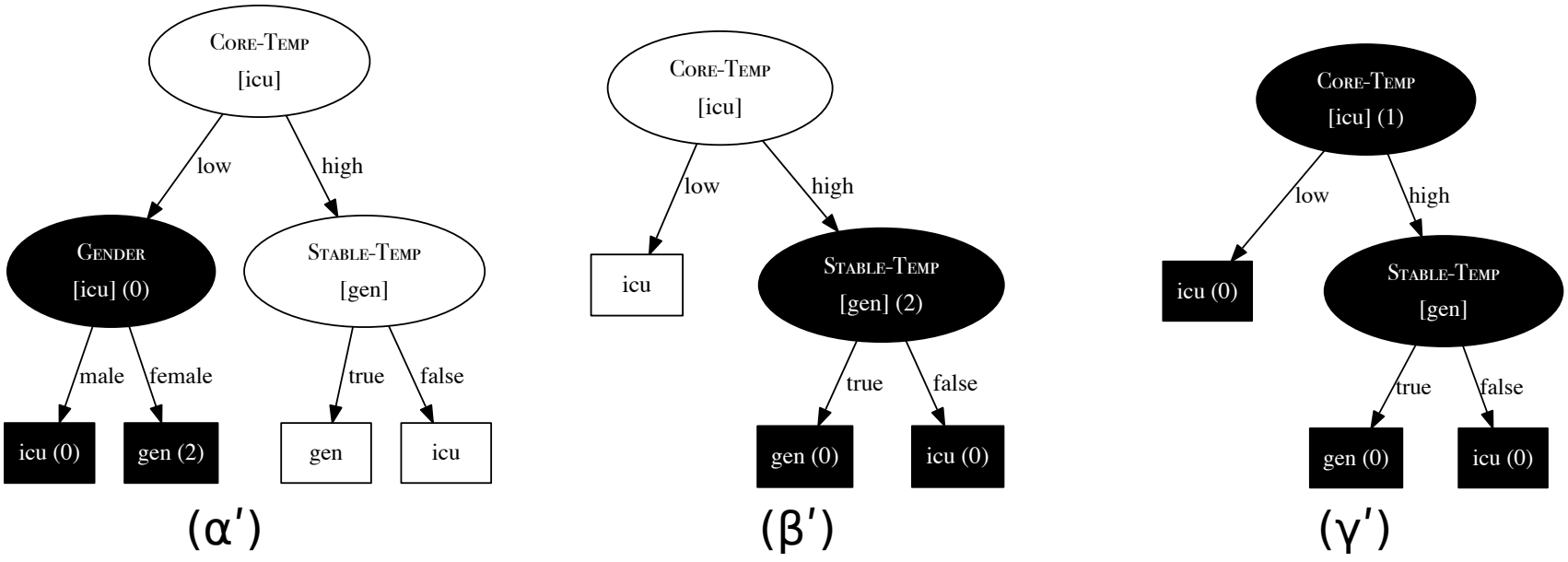


Πίνακας 1: Παράδειγμα συνόλου επικύρωσης για την εργασία δρομολόγησης μετεγχειρητικών ασθενών.

ID	Κεντρική Θερμ.	Σταθερή Θερμ.	Φύλο	Απόφαση
1	υψηλή	αληθές	άνδρας	gen
2	χαμηλή	αληθές	γυναίκα	icu
3	υψηλή	ψευδές	γυναίκα	icu
4	υψηλή	ψευδές	άνδρας	icu
5	χαμηλή	ψευδές	γυναίκα	icu
6	χαμηλή	αληθές	άνδρας	icu



Σχήμα 1: Το δέντρο απόφασης για την εργασία δρομολόγησης μετεγχειρητικών ασθενών.



Σχήμα 2: Οι επαναλήψεις του κλαδέματος μειωμένου σφάλματος για το δέντρο απόφασης στο Σχήμα 1, χρησιμοποιώντας το σύνολο επικύρωσης του Πίνακα 1. Το υποδέντρο που εξετάζεται για κλάδεμα σε κάθε επανάληψη επισημαίνεται με μαύρο. Η πρόβλεψη που επιστρέφει κάθε μη τερματικός κόμβος αναγράφεται σε αγκύλες. Το ποσοστό σφάλματος για κάθε κόμβο δίνεται σε παρενθέσεις.

Πλεονεκτήματα του κλαδέματος:

- Τα μικρότερα δέντρα είναι ευκολότερα στην ερμηνεία.
- Αυξημένη ακρίβεια γενίκευσης όταν υπάρχει θόρυβος στα δεδομένα εκπαίδευσης (**απόσβεση θορύβου**).

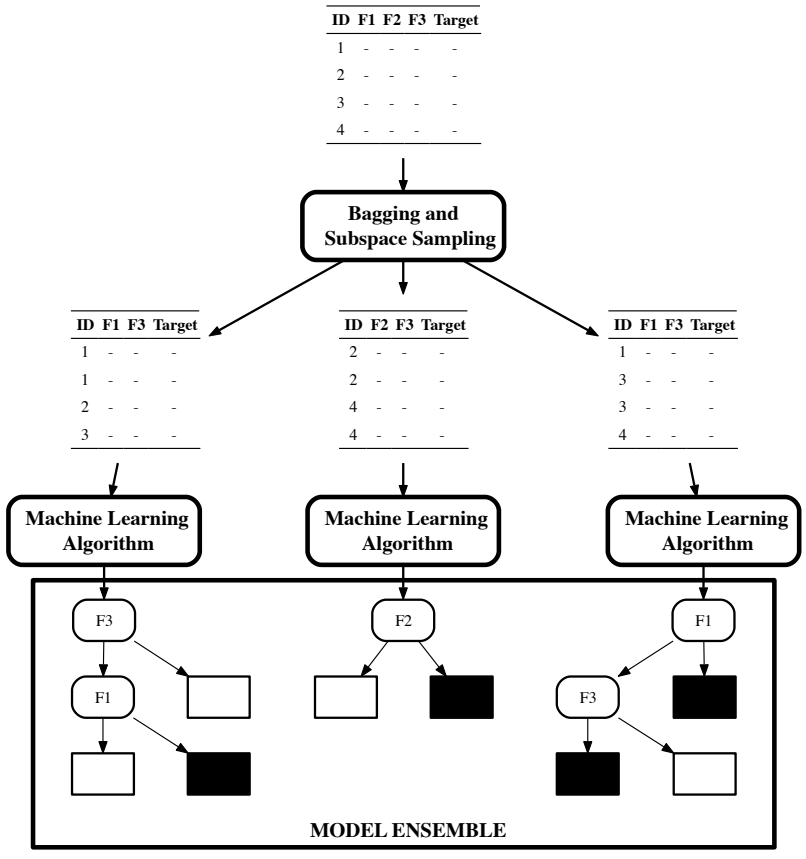
Σύνολα Μοντέλων

- Αντί να δημιουργούν ένα μόνο μοντέλο, δημιουργούν ένα σύνολο μοντέλων και στη συνέχεια κάνουν προβλέψεις συγκεντρώνοντας τις εξόδους αυτών των μοντέλων.
- Ένα μοντέλο πρόβλεψης που αποτελείται από ένα σύνολο μοντέλων ονομάζεται **σύνολο μοντέλων**.
- Για να λειτουργήσει αυτή η προσέγγιση, τα μοντέλα που βρίσκονται στο σύνολο πρέπει να διαφέρουν μεταξύ τους.

- Υπάρχουν δύο καθιερωμένες προσεγγίσεις για τη δημιουργία συνόλων:
  1. **bagging**.
  2. **boosting**.

- Όταν χρησιμοποιούμε **bagging** (ή **bootstrap aggregating**), κάθε μοντέλο στο σύνολο εκπαιδεύεται σε ένα τυχαίο δείγμα του συνόλου δεδομένων, γνωστό ως **δείγμα bootstrap**.
- Κάθε τυχαίο δείγμα έχει το ίδιο μέγεθος με το σύνολο δεδομένων και χρησιμοποιείται **δειγματοληψία με επανατοποθέτηση**.
- Κατά συνέπεια, από κάθε δείγμα bootstrap θα λείπουν ορισμένα από τα στιγμιότυπα του συνόλου δεδομένων. Έτσι, κάθε δείγμα bootstrap θα είναι διαφορετικό και αυτό σημαίνει ότι τα μοντέλα που εκπαιδεύονται σε διαφορετικά δείγματα bootstrap θα είναι επίσης διαφορετικά.

- Όταν το bagging χρησιμοποιείται με δέντρα απόφασης, κάθε δείγμα bootstrap χρησιμοποιεί μόνο ένα τυχαία επιλεγμένο υποσύνολο των περιγραφικών χαρακτηριστικών του συνόλου δεδομένων. Αυτό είναι γνωστό ως **δειγματοληψία υποχώρου**.
- Ο συνδυασμός bagging, δειγματοληψίας υποχώρου και δέντρων απόφασης είναι γνωστός ως μοντέλο **τυχαίου δάσους**.



Σχήμα 3: Η διαδικασία δημιουργίας ενός συνόλου μοντέλων με χρήση bagging και δειγματοληψίας υποχώρου.

- Το boosting λειτουργεί δημιουργώντας επαναληπτικά μοντέλα και προσθέτοντάς τα στο σύνολο.
- Η επανάληψη σταματά όταν προστεθεί ένας προκαθορισμένος αριθμός μοντέλων.
- Όταν χρησιμοποιούμε **boosting**, κάθε νέο μοντέλο που προστίθεται στο σύνολο είναι μεροληπτικά προσανατολισμένο στο να δίνει μεγαλύτερη προσοχή στα στιγμιότυπα που τα προηγούμενα μοντέλα ταξινόμησαν λανθασμένα.
- Αυτό γίνεται προσαρμόζοντας σταδιακά το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση των μοντέλων. Για να γίνει αυτό χρησιμοποιούμε ένα **σταθμισμένο σύνολο δεδομένων**.

Σταθμισμένο σύνολο δεδομένων

- Κάθε στιγμιότυπο έχει ένα συσχετισμένο βάρος $\mathbf{w}_i \geq 0$.
- Αρχικά ορίζεται ίσο με $\frac{1}{n}$, όπου n είναι ο αριθμός στιγμιοτύπων στο σύνολο δεδομένων.
- Αφού κάθε μοντέλο προστεθεί στο σύνολο, ελέγχεται στα δεδομένα εκπαίδευσης και τα βάρη των στιγμιοτύπων που ταξινομεί σωστά μειώνονται, ενώ τα βάρη των στιγμιοτύπων που ταξινομεί λανθασμένα αυξάνονται.
- Αυτά τα βάρη χρησιμοποιούνται ως κατανομή από την οποία γίνεται δειγματοληψία του συνόλου δεδομένων για τη δημιουργία ενός αναπαραγόμενου συνόλου εκπαίδευσης, όπου η επανάληψη ενός στιγμιοτύπου είναι ανάλογη του βάρους του.

Κατά τη διάρκεια κάθε επανάληψης εκπαίδευσης, ο αλγόριθμος:

1. Υπολογίζει το συνολικό σφάλμα ϵ
2. Αυξάνει τα βάρη των λανθασμένων:

$$\mathbf{w}[i] \leftarrow \mathbf{w}[i] \times \left(\frac{1}{2\epsilon}\right) \tag{1}$$

3. Μειώνει τα βάρη των σωστών:

$$\mathbf{w}[i] \leftarrow \mathbf{w}[i] \times \left(\frac{1}{2(1-\epsilon)}\right) \tag{2}$$

Υπολογισμός συντελεστή εμπιστοσύνης:

$$\alpha = \frac{1}{2} \ln \left(\frac{1 - \epsilon}{\epsilon} \right) \tag{3}$$

- Όσο μικρότερο το ϵ, τόσο μεγαλύτερο το α
- Το μοντέλο αποκτά μεγαλύτερη βαρύτητα

- Αφού δημιουργηθεί το σύνολο μοντέλων, το σύνολο κάνει προβλέψεις χρησιμοποιώντας σταθμισμένη συγκέντρωση των προβλέψεων που παράγονται από τα επιμέρους μοντέλα.
- Τα βάρη που χρησιμοποιούνται σε αυτή τη συγκέντρωση είναι απλώς οι συντελεστές εμπιστοσύνης που σχετίζονται με κάθε μοντέλο.

Απλό Παράδειγμα Boosting

Έστω 5 παρατηρήσεις για ταξινόμηση:

$+1$ ή -1

ID	Κλάση	M_1	Σωστό
1	+1	+1	N
2	+1	-1	O
3	-1	-1	N
4	+1	+1	N
5	-1	+1	O

Ιδέα

Το M_1 κάνει λάθη στα ID 2 και 5 $\Rightarrow$ αυτά παίρνουν μεγαλύτερο βάρος στο επόμενο μοντέλο.

Boosting: Αλλαγή Βαρών

Αρχικά:

$$w_i = \frac{1}{5} = 0.20$$

Μετά το M_1:

ID	M_1 σωστό;	Βάρος
1	N	↓
2	O	↑
3	N	↓
4	N	↓
5	O	↑

Ιδέα

Λάθη $\Rightarrow$ μεγαλύτερο βάρος $\Rightarrow$ έμφαση στο M_2

Boosting: Τελική Πρόβλεψη

Μετά από μερικές επαναλήψεις έχουμε ένα σύνολο μοντέλων:

$$M_1, M_2, M_3$$

Κάθε μοντέλο έχει έναν συντελεστή εμπιστοσύνης:

$$\alpha_1, \alpha_2, \alpha_3$$

Μοντέλο	Πρόβλεψη	Βάρος μοντέλου
M_1	$+1$	0.7
M_2	-1	0.4
M_3	$+1$	0.9

Άρα:

$$+1 : 0.7 + 0.9 = 1.6$$

$$-1 : 0.4$$

Τελική πρόβλεψη: $+1$

- Ποια προσέγγιση πρέπει να χρησιμοποιήσουμε; Το bagging είναι απλούστερο στην υλοποίηση και στην παραλληλοποίηση από το boosting και, επομένως, μπορεί να είναι καλύτερο ως προς την ευκολία χρήσης και τον χρόνο εκπαίδευσης.
- Τα εμπειρικά αποτελέσματα δείχνουν ότι:
  - τα σύνολα ενισχυμένων δέντρων απόφασης ήταν το μοντέλο με την καλύτερη απόδοση από όσα δοκιμάστηκαν για σύνολα δεδομένων με έως 4.000 περιγραφικά χαρακτηριστικά.
  - τα σύνολα τυχαίων δασών (βασισμένα στο bagging) απέδωσαν καλύτερα για σύνολα δεδομένων που περιείχαν περισσότερα από 4.000 χαρακτηριστικά.

Σύνοψη

- Το μοντέλο **δέντρου απόφασης** κάνει προβλέψεις με βάση ακολουθίες ελέγχων στις τιμές των περιγραφικών χαρακτηριστικών ενός ερωτώμενου στιγμιοτύπου.
- Ο αλγόριθμος είναι ένας τυπικός αλγόριθμος για την επαγωγή δέντρων απόφασης από ένα σύνολο δεδομένων.

Δέντρα Απόφασης: Πλεονεκτήματα

- Είναι ερμηνεύσιμα.
- Χειρίζονται τόσο κατηγορικά όσο και συνεχή περιγραφικά χαρακτηριστικά.
- Έχουν την ικανότητα να μοντελοποιούν τις αλληλεπιδράσεις μεταξύ περιγραφικών χαρακτηριστικών, αν και αυτή η ικανότητα περιορίζεται αν χρησιμοποιηθεί **προ-κλάδεμα**.
- Είναι σχετικά ανθεκτικά στην **κατάρα της διαστατικότητας**.
- Είναι σχετικά ανθεκτικά στον θόρυβο του συνόλου δεδομένων αν χρησιμοποιηθεί **κλάδεμα**.

Δέντρα Απόφασης: Πιθανά μειονεκτήματα

- Τα δέντρα γίνονται μεγάλα όταν χειρίζονται συνεχή χαρακτηριστικά.
- Τα δέντρα απόφασης είναι πολύ εκφραστικά και ευαίσθητα στο σύνολο δεδομένων· ως αποτέλεσμα, μπορούν να υπερπροσαρμοστούν στα δεδομένα αν υπάρχουν πολλά χαρακτηριστικά (κατάρα της διαστατικότητας).
- Είναι eager learners και μπορεί να επηρεαστούν από **concept drift**.