ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Τεχνητή Νοημοσύνη

16η διάλεξη (2023-24)

Ίων Ανδρουτσόπουλος

http://www.aueb.gr/users/ion/

Οι διαφάνειες αυτής της διάλεξης βασίζονται σε ύλη των βιβλίων: (α) *Artificial Intelligence – A Modern Approach* των S. Russel και P. Norvig, 2η και 4η έκδοση, Prentice Hall, 2003 και 2020, (β) *Τεχνητή Νοημοσύνη* των Βλαχάβα κ.ά., 3η έκδοση, Β. Γκιούρδας Εκδοτική, 2006 και (γ) *Machine Learning* του T. Mitchell, McGraw-Hill, 1997. Τα περισσότερα σχήματα των διαφανειών βασίζονται σε αντίστοιχα σχήματα των διαφανειών που συνοδεύουν τα πρώτα δύο βιβλία.

Τι θα ακούσετε σήμερα

- Αφελείς ταξινομητές Bayes.
- Αλγόριθμος ID3.
- Θόρυβος και υπερ-προσαρμογή.
- Αλγόριθμος Τυχαίου Δάσους (Random Forest)

Αφελείς ταξινομητές Bayes (Naive Bayes)

- Παράδειγμα: εισερχόμενο μήνυμα, παριστάνεται ως:

$$\vec{X} = \langle X_1, X_2, \ldots, X_m \rangle = \langle 0, 1, \ldots, 1 \rangle$$

- Συνάρτηση ταξινόμησης:

$$h(\vec{X}) = 1 \text{ , αν} \quad \mathrm{P}(\mathrm{C} = 1 | \vec{\mathrm{X}}) > \mathrm{P}(\mathrm{C} = 0 | \vec{\mathrm{X}})$$

- Με το θεώρημα του Bayes (εδώ c = 0 ή c = 1):

$$P(C = c \mid \vec{X}) = \frac{P(C = c)\, P(\vec{X} \mid C = c)}{P(\vec{X})}$$

Πρέπει να εκτιμηθούν οι πιθανότητες όλων των συνδυασμών x_1, x_2, ..., $x_m \mid c$. Πάρα πολλοί και πολλοί είναι σπάνιοι στα δεδομένα μας.

Παραδοχή ανεξαρτησίας

- Οι αφελείς ταξινομητές Bayes κάνουν την παραδοχή ότι οι τιμές των $X_1, \ldots, X_m$ είναι **ανεξάρτητες δεδομένης της τιμής της *C***.
  - Συνήθως δεν ισχύει, αλλά στην πράξη καλά αποτελέσματα.

$$P(\overrightarrow{X} = \langle x_1, x_2, \ldots, x_m \rangle \mid C = c) =$$

$$P(X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_m = x_m \mid C = c) \approx$$

$$P(X_1 = x_1 \mid C = c) \cdot \ldots \cdot P(X_m = x_m \mid C = c) =$$

$$\prod_{i=1}^{m} P(X_i = x_i \mid C = c)$$

Αφελείς ταξινομητές Bayes – συνέχεια

- Τότε:

$$P(C=1|\vec{X}) = \frac{P(C=1)\cdot\prod_{i=1}^{m} P(X_i = x_i \mid C=1)}{P(\vec{X})}$$

$$P(C=0|\vec{X}) = \frac{P(C=0)\cdot\prod_{i=1}^{m} P(X_i = x_i \mid C=0)}{P(\vec{X})}$$

- Τώρα όλες οι πιθανότητες μπορούν να εκτιμηθούν εύκολα από τα παραδείγματα εκπαίδευσης.
- Οι παρονομαστές δεν μας χρειάζονται, γιατί είναι ίδιοι.

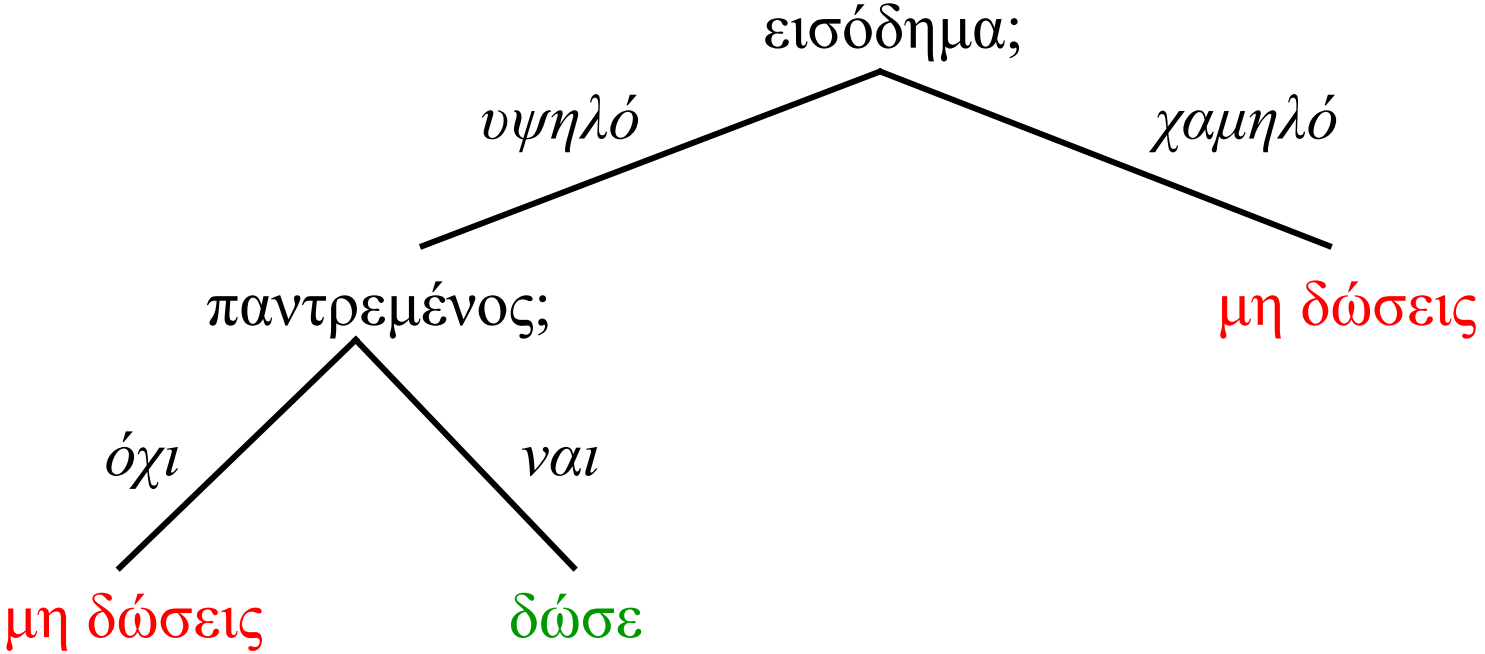
Εξομάλυνση πιθανοτήτων

- Προσπαθούμε να εκτιμήσουμε την $P(X_i = x_i \mid C = c)$.
  - 1η προσέγγιση: Σε πόσα μηνύματα εκπαίδευσης της κατηγορίας c εμφανίζεται η λέξη που αντιστοιχεί στη X_i;
  - Τι γίνεται όμως αν η λέξη της X_i δεν εμφανίζεται σε κανένα μήνυμα εκπαίδευσης της c; **Μηδενική εκτίμηση**.
  - Μηδενίζεται και ολόκληρο το $\prod_{i=1}^{m} P(X_i = x_i \mid C = c)$
  - Μηδενική $P(C = c \mid \vec{X})$ λόγω μόνο μιας ιδιότητας.
- Ένας τρόπος εξομάλυνσης: **εκτιμήτρια Laplace**.
  - Θεωρούμε κατά την εκτίμηση της $P(X_i = x_i \mid C = c)$ ότι υπάρχουν δύο ακόμη ψευτο-μηνύματα εκπαίδευσης κατηγορίας c: ένα που περιέχει τη λέξη της X_i και ένα που δεν την περιέχει.
  - Επομένως +1 στον αριθμητή της εκτίμησης, +2 στον παρονομαστή.
  - Γενικότερα, κατά την εκτίμηση τυχαίας μεταβλητής με k δυνατές τιμές, **+1** στον **αριθμητή** και **+*k*** στον **παρονομαστή**.

Χαρακτηριστικά Naive Bayes

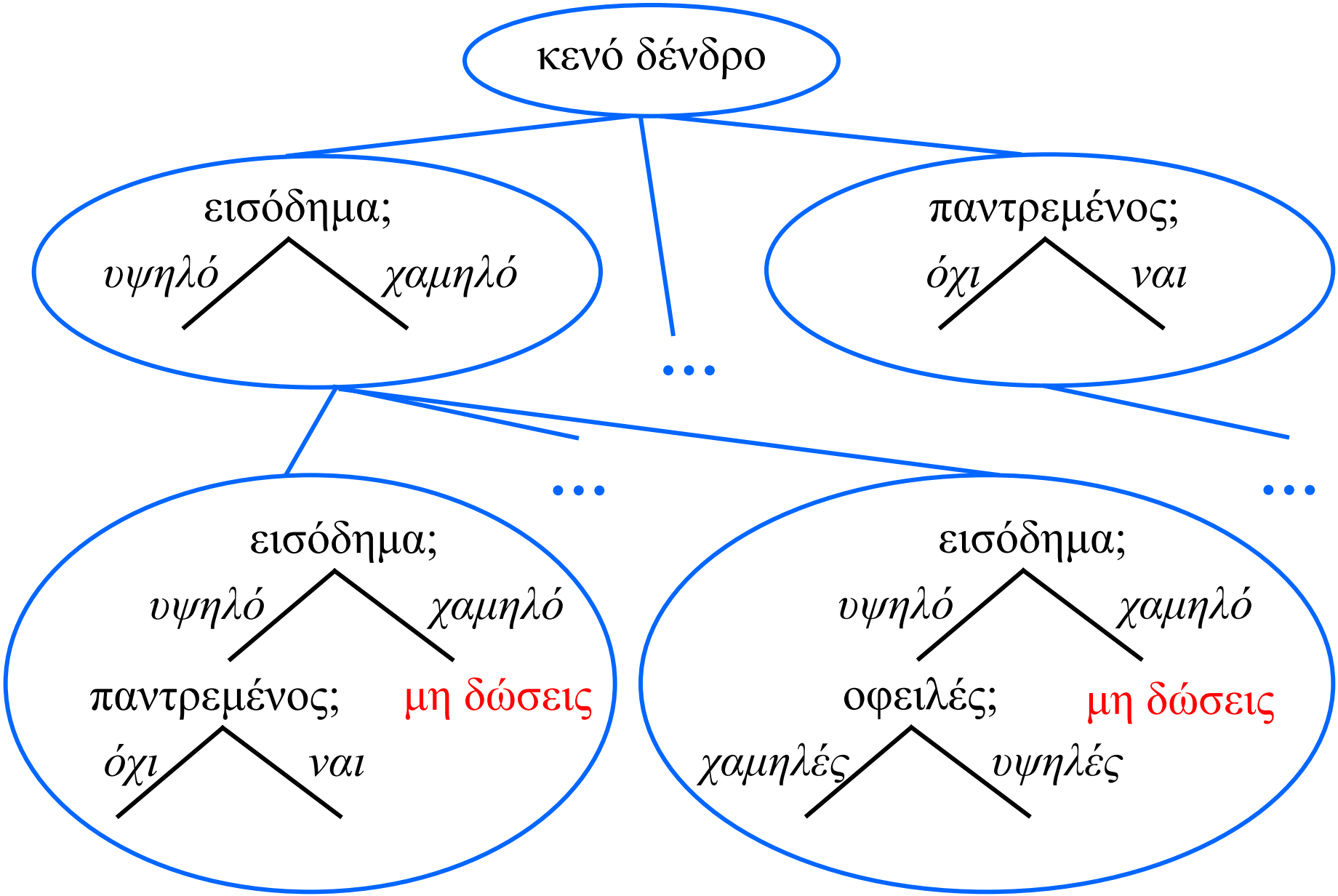
- Πολύ **μικρό υπολογιστικό κόστος**:
  - ***O*(*mN*)** κατά την **εκπαίδευση** για την εκτίμηση των πιθανοτήτων $P(X_i|C)$,
  - ***O*(*m*)** κατά την **κατάταξη** για τον υπολογισμό του γινομένου των $P(X_i|C)$,
  - όπου ***N*** το πλήθος των παραδειγμάτων εκπαίδευσης και ***m*** το πλήθος των ιδιοτήτων.
- Πολύ **μικρές απαιτήσεις μνήμης**:
  - ***O*(*m*)** για την αποθήκευση των εκτιμήσεων των $P(X_i|C)$.
- **Δεν μπορούμε** να παραστήσουμε άμεσα τη γνώση που απέκτησε με μορφή **λογικών κανόνων**.
  - Όπως και στην περίπτωση του *k*-NN.

Μάθηση δέντρων απόφασης

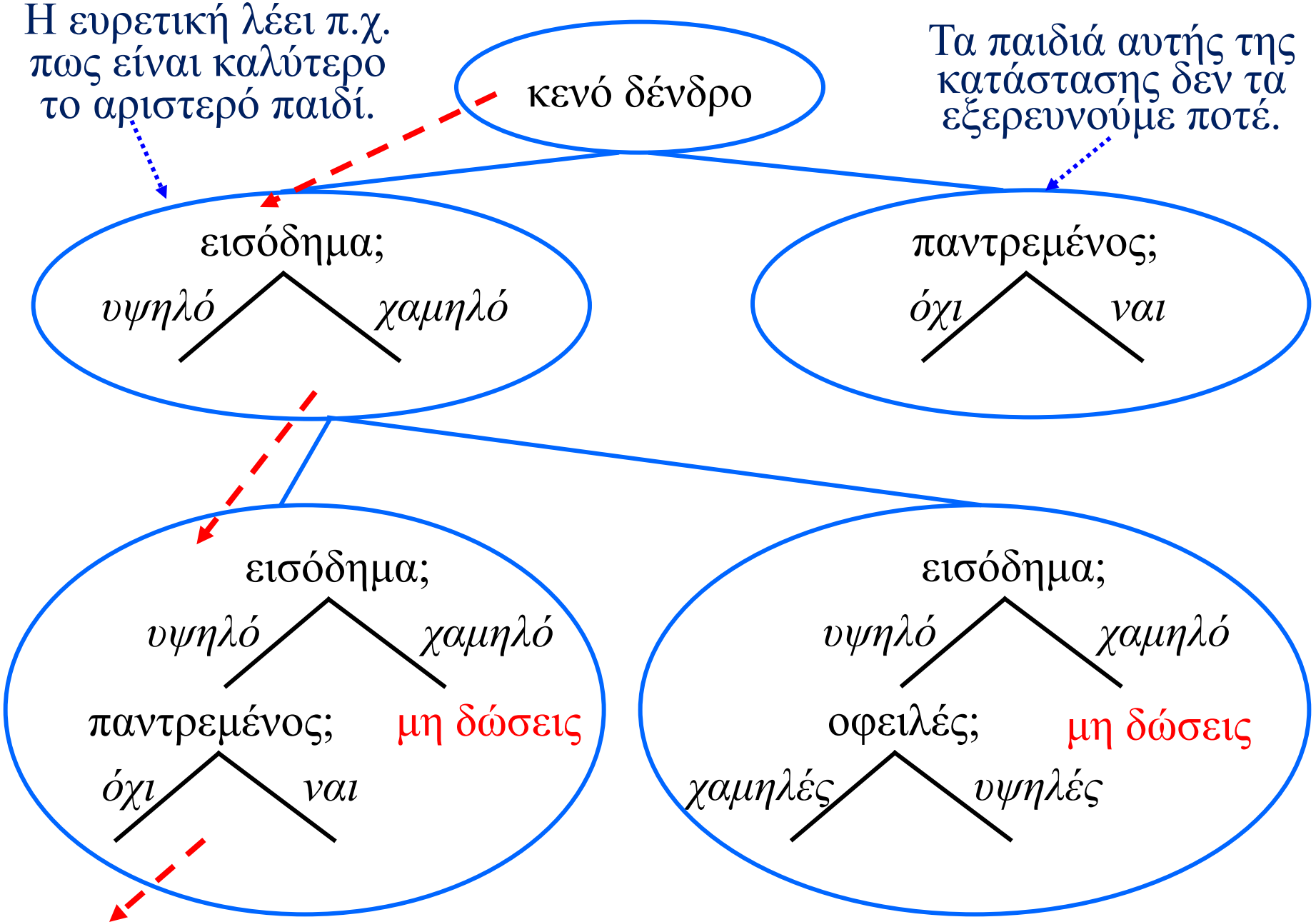


- Ο **αλγόριθμος ID3** κατασκευάζει δέντρα αυτής της μορφής από τα παραδείγματα εκπαίδευσης.
- Σε κάθε **εσωτερικό κόμβο** ελέγχουμε την **τιμή μιας ιδιότητας**.
- Τα **φύλλα** αντιστοιχούν σε **αποφάσεις**.

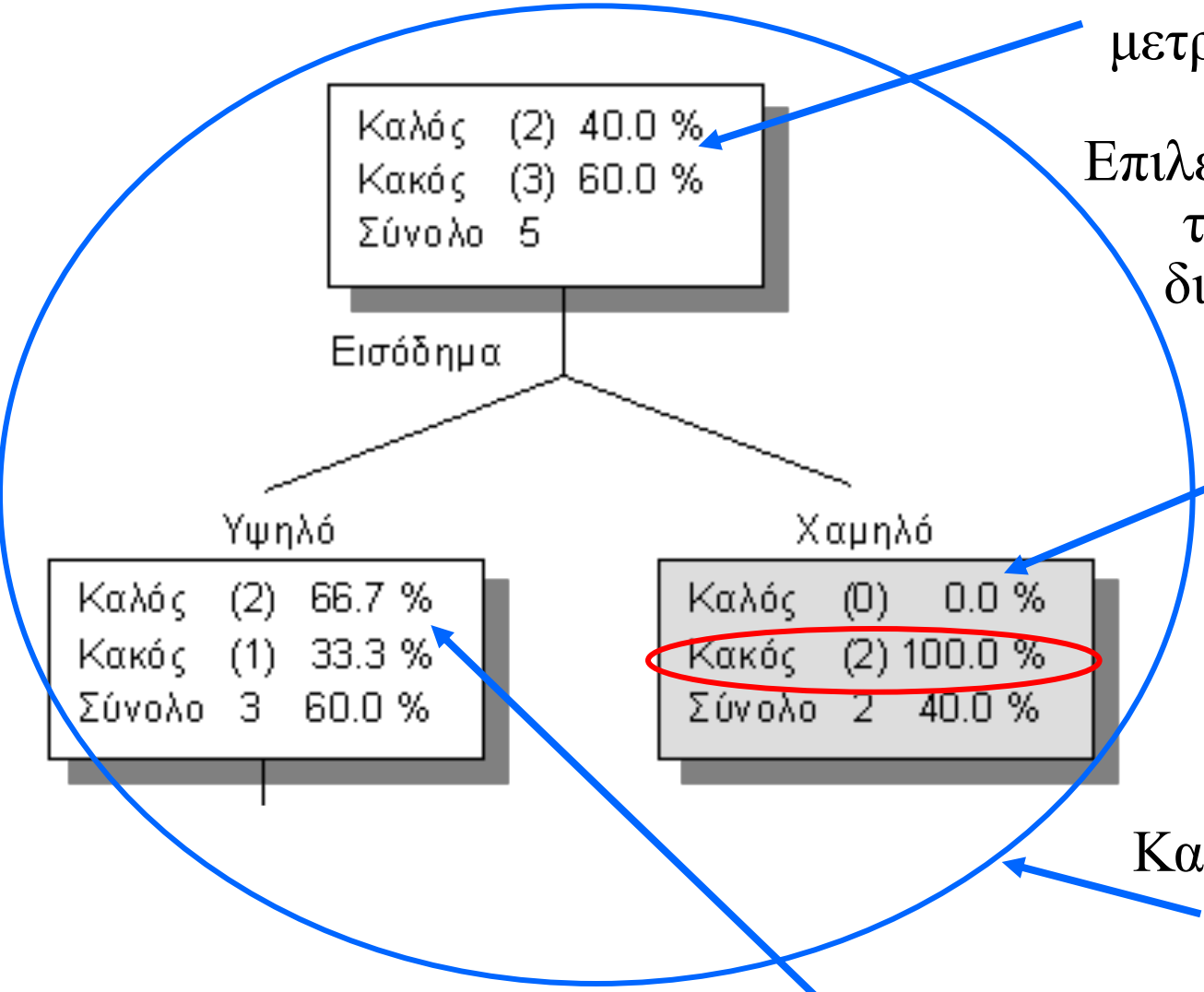
Χώρος αναζήτησης δέντρων απόφασης



Ο υποχώρος αναζήτησης του ID3



Καταστάσεις: περισσότερες λεπτομέρειες



Σε κάθε κόμβο του **δένδρου** μετράμε τα παραδείγματα κάθε κατηγορίας. Επιλέγουμε μέσω ευρετικής την ιδιότητα που τα διαχωρίζει καλύτερα.

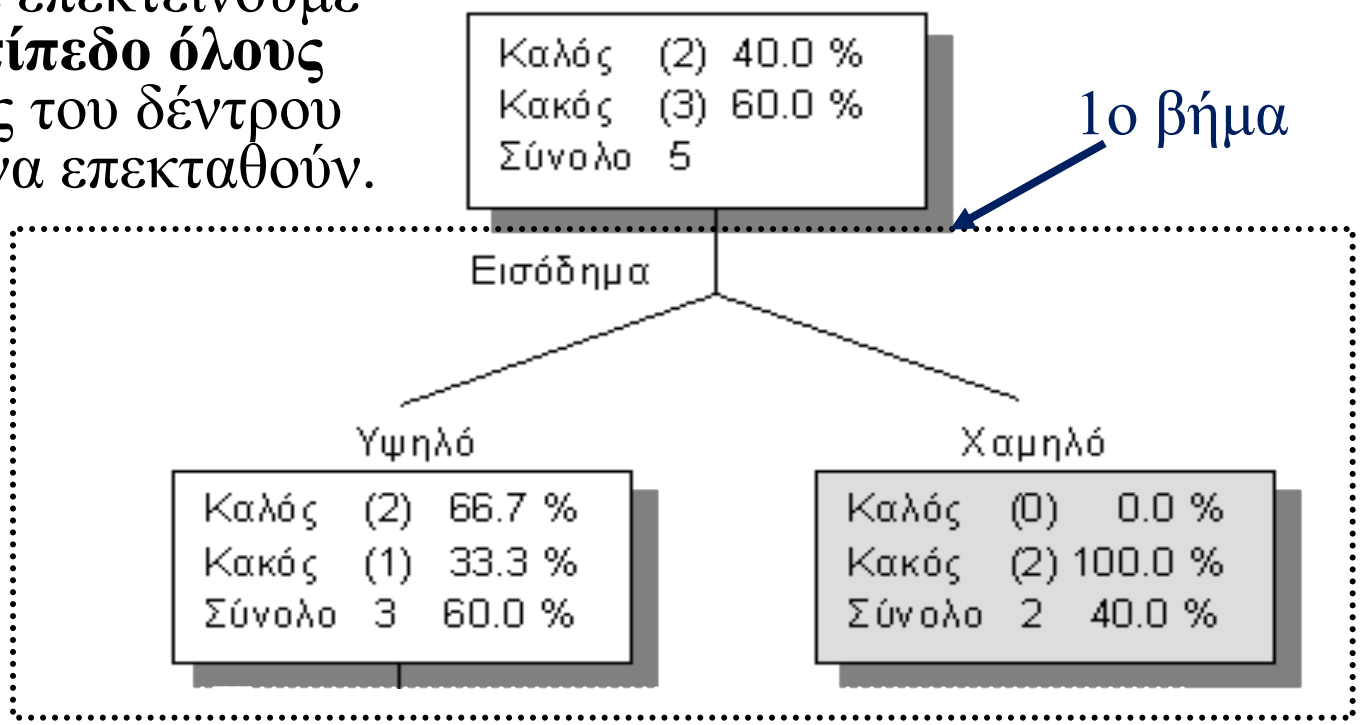
Όλα τα παραδείγματα εκπαίδευσης που έχουν μείνει ανήκουν σε **μία κατηγορία**. Απόφαση: **μη δώσεις**.

Κατάσταση του χώρου αναζήτησης.

Τα παραδείγματα εκπαίδευσης **δεν** ανήκουν όλα σε μία κατηγορία. **Επέκτεινε** το δέντρο κάτω από αυτόν τον κόμβο.

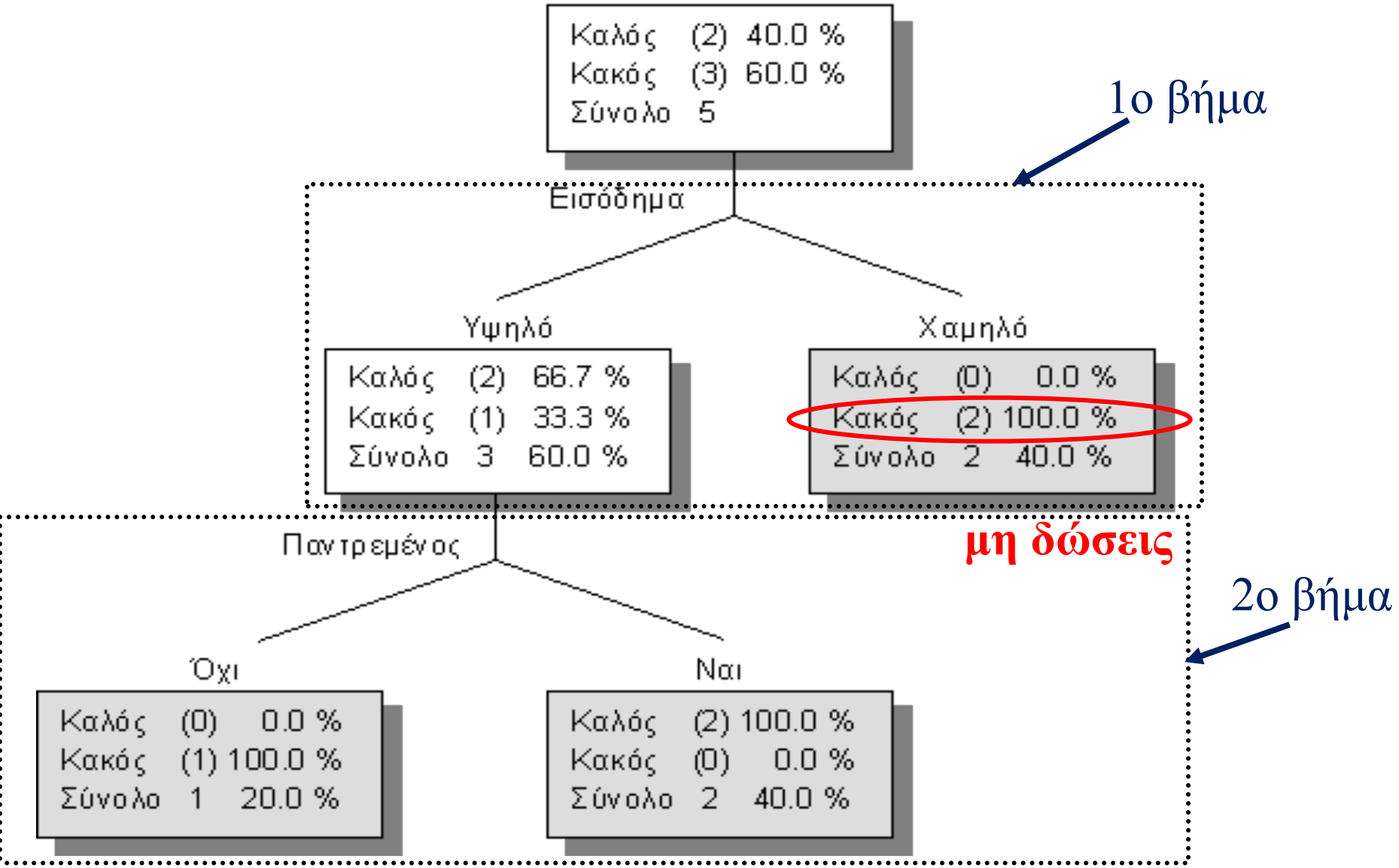
ID3: επέκταση δένδρου απόφασης

Σε κάθε βήμα επεκτείνουμε **κατά ένα επίπεδο όλους** τους κόμβους του δέντρου που μπορούν να επεκταθούν.

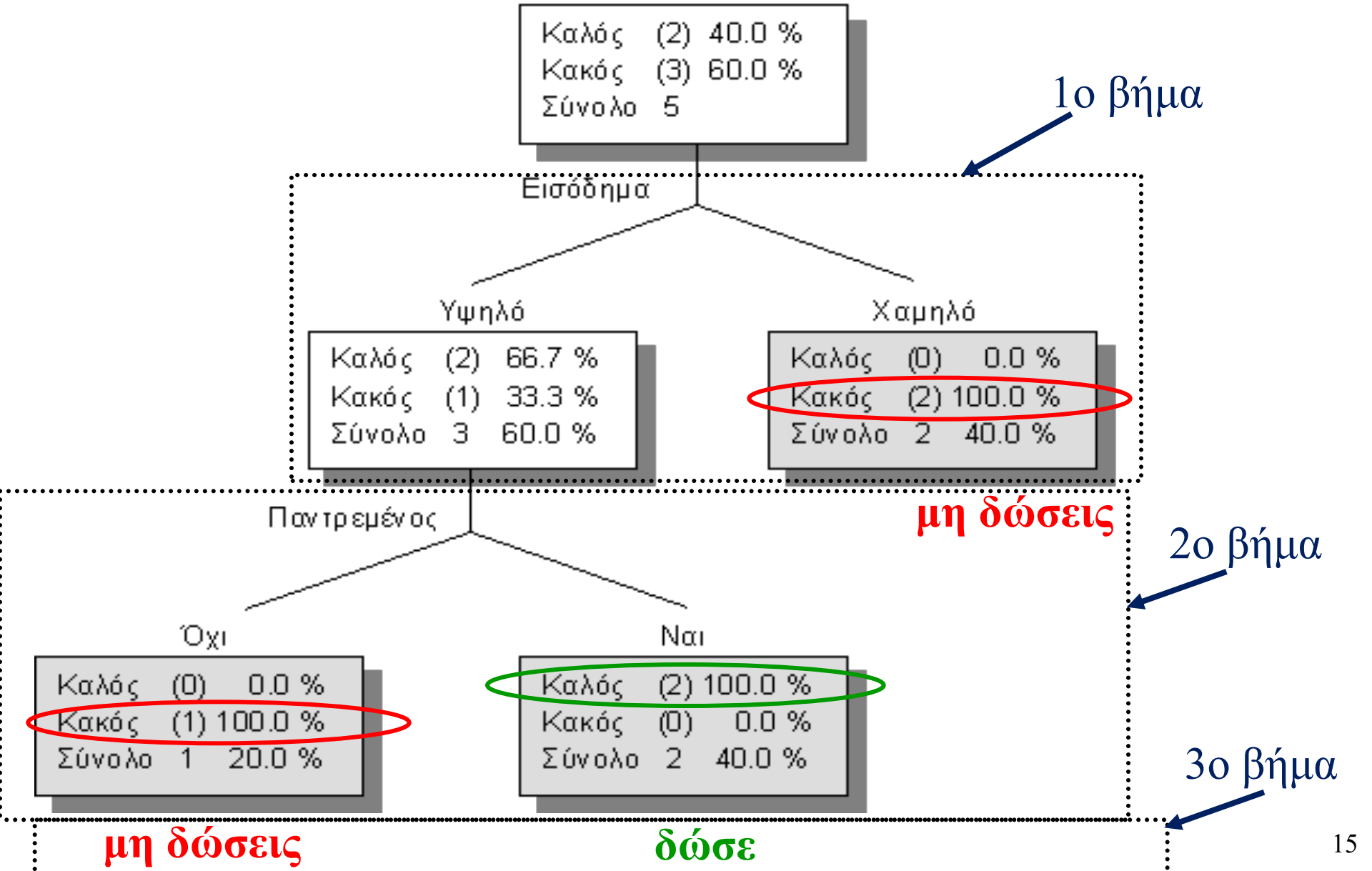


Θεωρούμε επέκταση και την προσθήκη φύλλου-απόφασης.

ID3: επέκταση δένδρου απόφασης



ID3: επέκταση δένδρου απόφασης



ID3: Αναρρίχηση λόφου

- **Σε κάθε κατάσταση** (ημιτελές δέντρο) του χώρου αναζήτησης, επιλέγει να **επεκτείνει** κάθε κόμβο του δέντρου απόφασης (που πρέπει να επεκταθεί) με την **ιδιότητα** που **εκτιμά** πως είναι η **χρησιμότερη**.
  - Χρησιμοποιεί ως ευρετική συνάρτηση αξιολόγησης των ιδιοτήτων το **κέρδος πληροφορίας** (IG).
- Προκύπτει έτσι μια **μοναδική κατάσταση-παιδί** (νέο δέντρο απόφασης) στην οποία μεταβαίνουμε.
  - Το **μέτωπο** περιέχει πάντα **μία μόνο κατάσταση**. Δεν υπάρχει δυνατότητα εξέτασης εναλλακτικών μονοπατιών.

ID3: επιλογή ιδιότητας

κενό δένδρο

- Ιδιότητες που μπορούν να χρησιμοποιηθούν:
  - Οφειλές (X_1), εισόδημα (X_2), παντρεμένος (X_3).
  - Απάντηση: ***C* = 1** (δώσε), ***C* = 0** (μη δώσεις).

$$H(C) = -\frac{2}{5}\cdot\log_2\frac{2}{5} - \frac{3}{5}\cdot\log_2\frac{3}{5} = 0.97$$

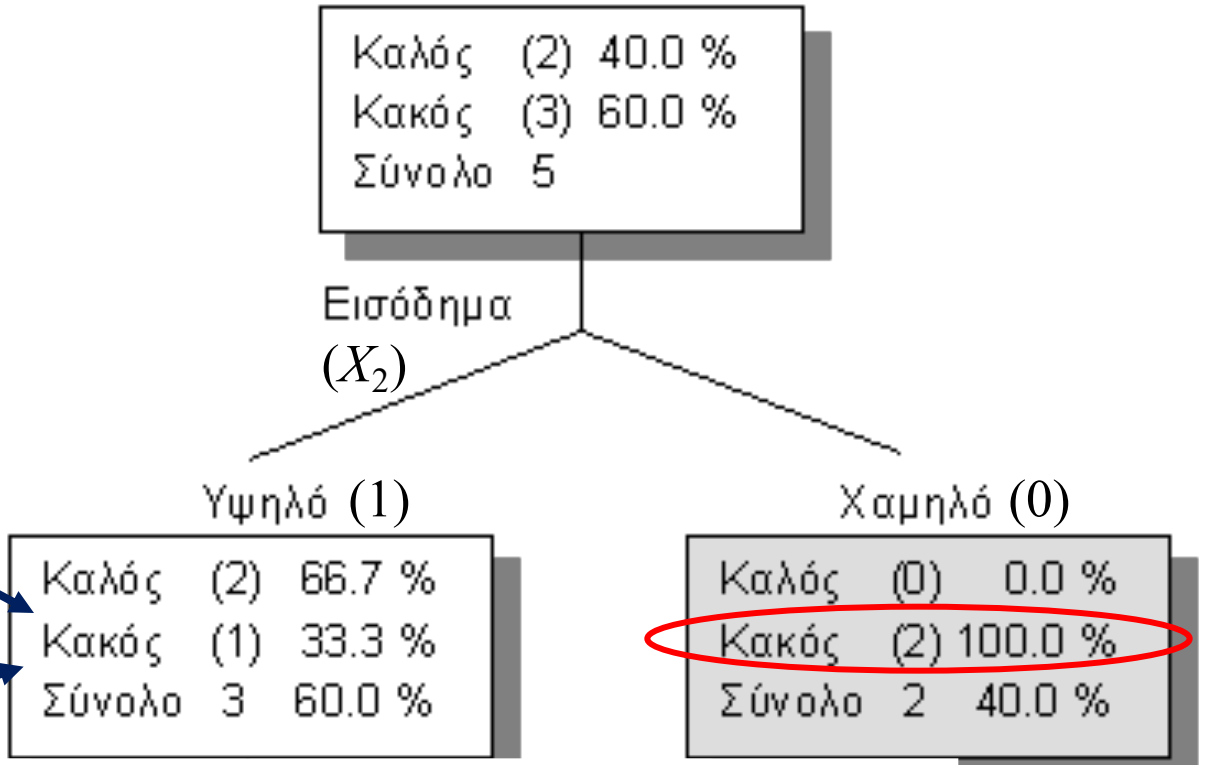
$$IG(C, X_1) = H(C) - \sum_{x\in\{0,1\}} P(X_1 = x)\cdot H(C \mid X_1 = x) = 0.02$$

$$IG(C, X_2) = H(C) - \sum_{x\in\{0,1\}} P(X_2 = x)\cdot H(C \mid X_2 = x) = 0.42$$

$$IG(C, X_3) = H(C) - \sum_{x\in\{0,1\}} P(X_3 = x)\cdot H(C \mid X_3 = x) = 0.17$$

ID3: επιλογή ιδιότητας

Προσοχή: οι πιθανότητες εκτιμούνται τώρα από τα παραδείγματα **αυτού** του κόμβου.
π.χ. $P(C=1) = 2/3$



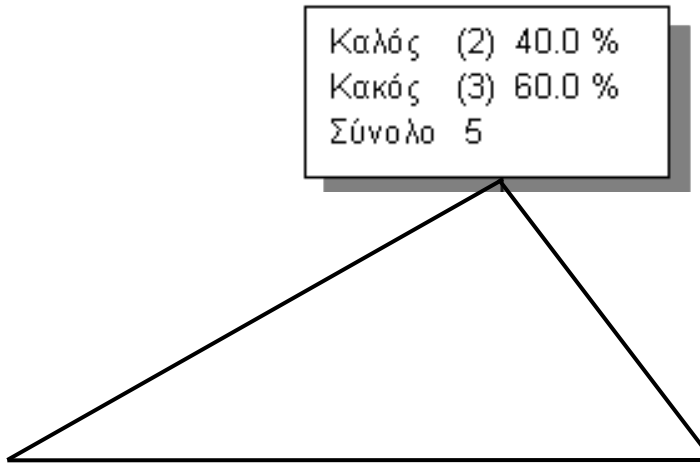
Ιδιότητες που μπορούν να χρησιμοποιηθούν:

Οφειλές (X_1), παντρεμένος (X_3).

Υπολογίζουμε **εκ νέου**: $IG(C, X_1) = \ldots$, $IG(C, X_3) = \ldots$

$IG(C, X_1) < IG(C, X_3)$. Άρα **επιλέγουμε το X_3** (παντρεμένος).

Αρχική κλήση του ID3



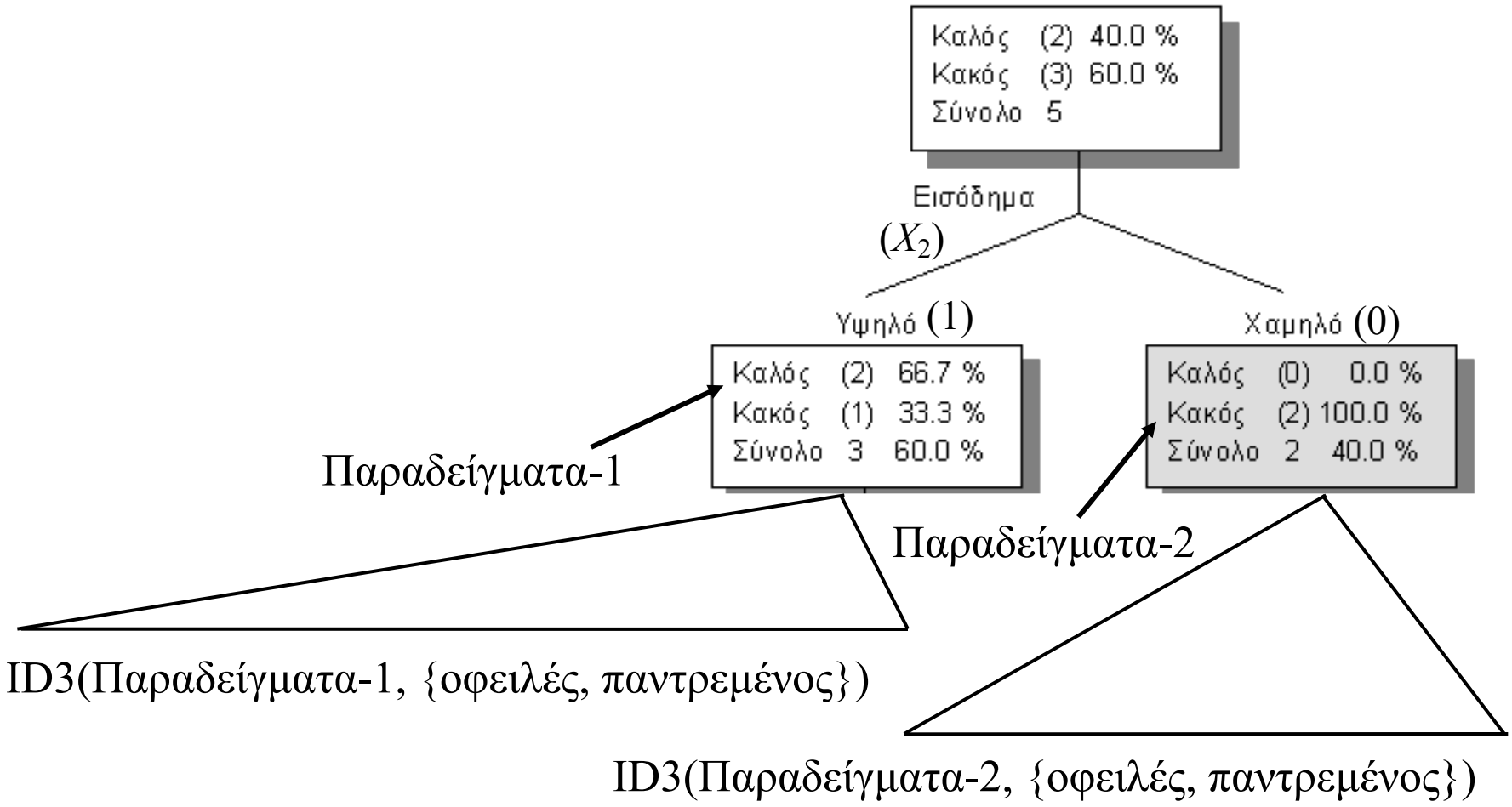
ID3(Παραδείγματα, {εισόδημα, οφειλές, παντρεμένος})

Παραδείγματα: 2 καλοί και 3 κακοί.

Ιδιότητες = {εισόδημα, οφειλές, παντρεμένος}.

Καλύτερη ιδιότητα: εισόδημα.

Αναδρομική κλήση του ID3



Αλγόριθμος ID3

συνάρτηση ID3(*παραδείγματα*, *ιδιότητες*, *προεπιλεγμένη*)

είσοδοι: *παραδείγματα*: σύνολο παραδειγμάτων εκπαίδευσης

ιδιότητες: σύνολο διαθέσιμων ιδιοτήτων

προεπιλεγμένη: προεπιλεγμένη κατηγορία

αν *παραδείγματα* = {} **τότε επίστρεψε** *προεπιλεγμένη κατηγορία*

διαφορετικά αν όλα τα *παραδείγματα* ανήκουν στην ίδια κατηγορία **τότε επίστρεψε** αυτή την κατηγορία

διαφορετικά αν *ιδιότητες* = {} **τότε επίστρεψε** την κατηγορία που είναι συχνότερη στα *παραδείγματα*

διαφορετικά ...

Αλγόριθμος ID3 (συνέχεια)

Επιλέγουμε την ιδιότητα που παρέχει το μεγαλύτερο κέρδος πληροφορίας.

διαφορετικά

καλύτερη ← επιλογή-ιδιότητας(*ιδιότητες*, *παραδείγματα*)

δέντρο ← νέο δέντρο που στη ρίζα του ελέγχει την *καλύτερη*

έστω *m* η συχνότερη κατηγορία μεταξύ των *παραδειγμάτων*

για κάθε δυνατή τιμή v_i της *καλύτερης*

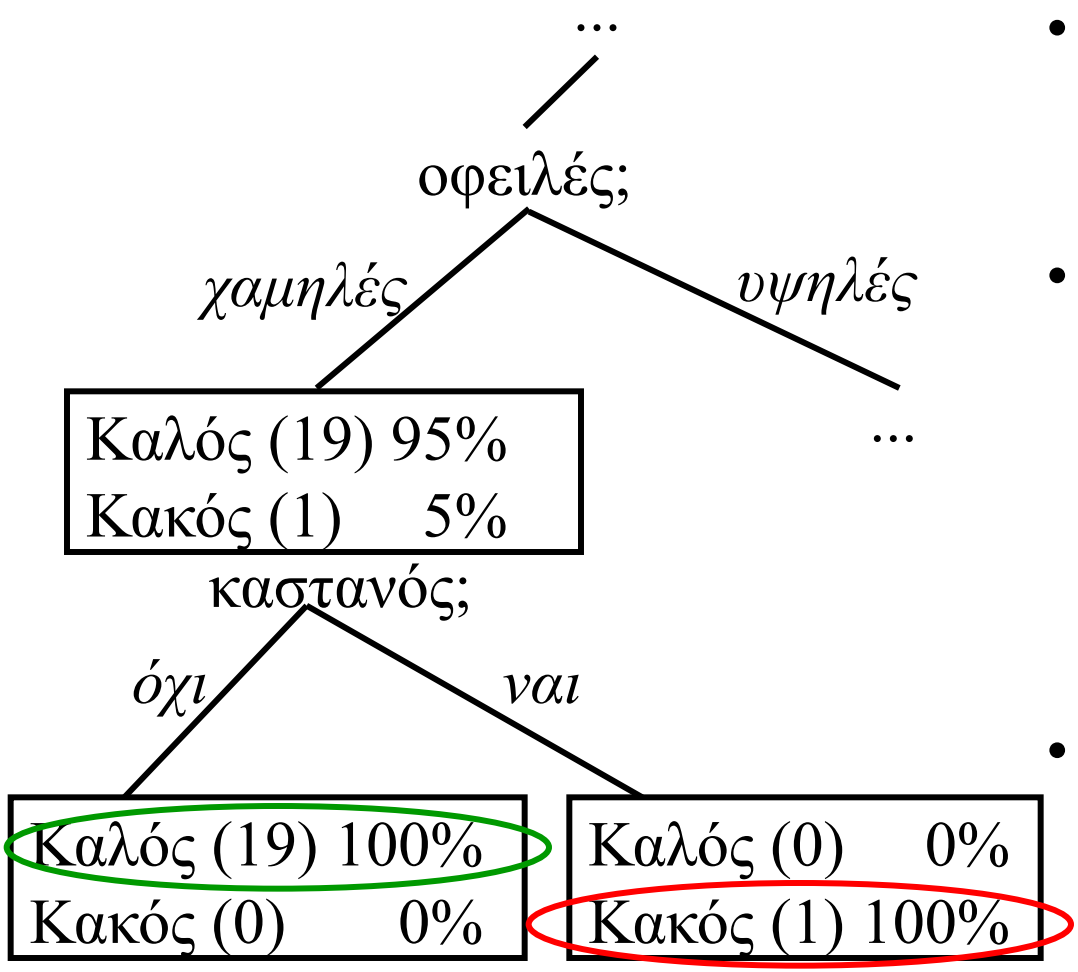
$παραδείγματα_i \leftarrow \{\pi \in παραδείγματα \mid \pi.καλύτερη = v_i\}$

υποδέντρο ← ID3($παραδείγματα_i$, *ιδιότητες* – *καλύτερη*, *m*)

πρόσθεσε κλαδί με ετικέτα v_i στο *δέντρο* που να οδηγεί από τη ρίζα στο *υποδέντρο*

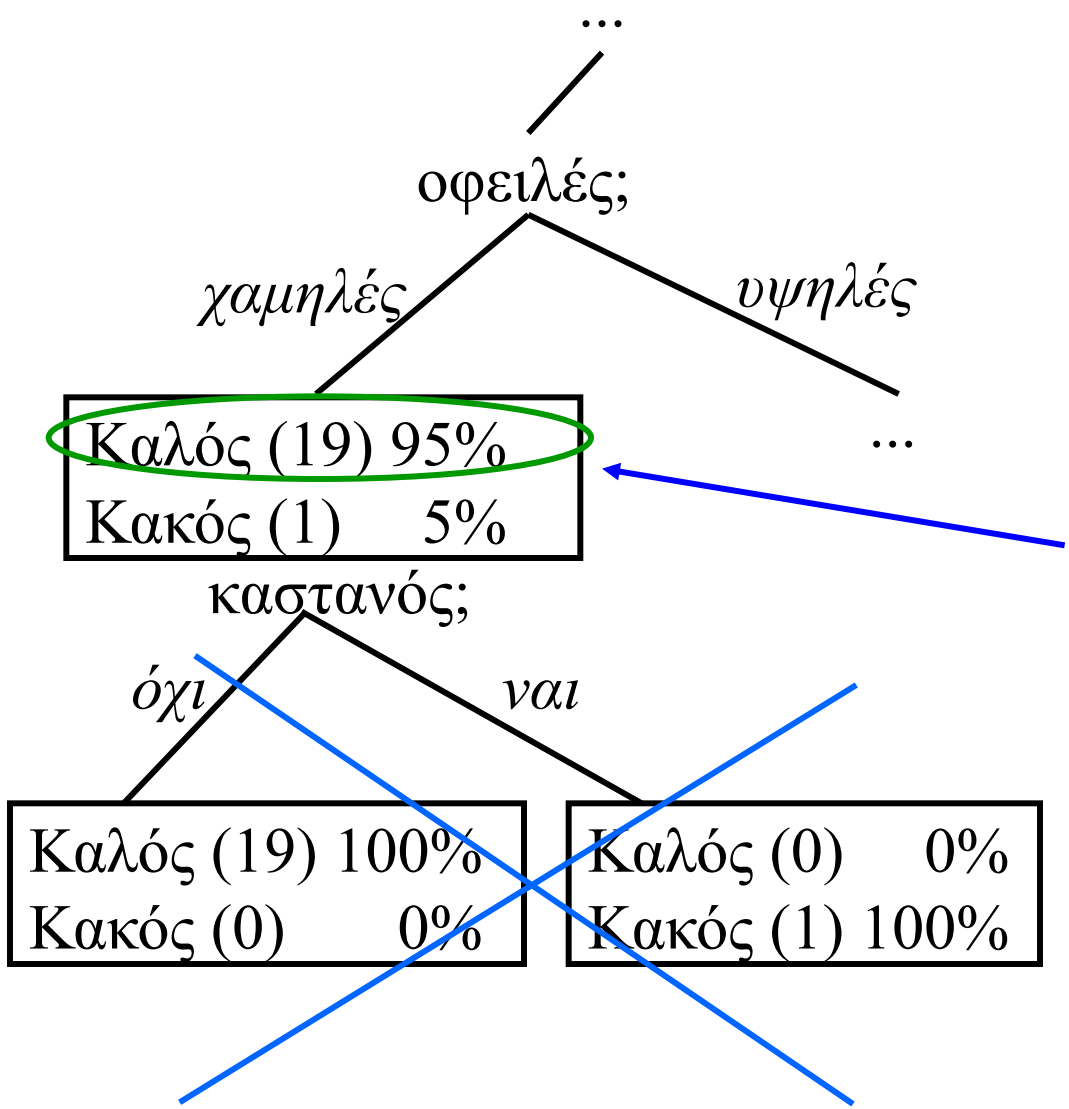
επίστρεψε το *δέντρο*

Ο πλήρης διαχωρισμός μπορεί να βλάπτει



- **Προσπαθούμε** να επιτύχουμε **πλήρη διαχωρισμό** στα φύλλα.
- Αναγκαζόμαστε να χρησιμοποιήσουμε **άσχετες ιδιότητες** που **τυχαίνει** να διαχωρίζουν τα **συγκεκριμένα παραδείγματα** εκπαίδευσης.
- Κι αυτό λόγω ενός μόνο κακού πελάτη με χαμηλές οφειλές στα παραδείγματα εκπαίδευσης. (Μπορεί να πρόκειται για λάθος!)

Ο πλήρης διαχωρισμός μπορεί να βλάπτει



Αν στην πραγματικότητα όλοι οι πελάτες με χαμηλές οφειλές είναι καλοί

(ο κακός καστανός ήταν λάθος στα δεδομένα εκπαίδευσης),

τότε αν είχαμε σταματήσει σε αυτόν το κόμβο (με απόφαση δώσε) θα είχαμε ένα δέντρο με **μεγαλύτερο ποσοστό ορθότητας** στο σύνολο του πληθυσμού

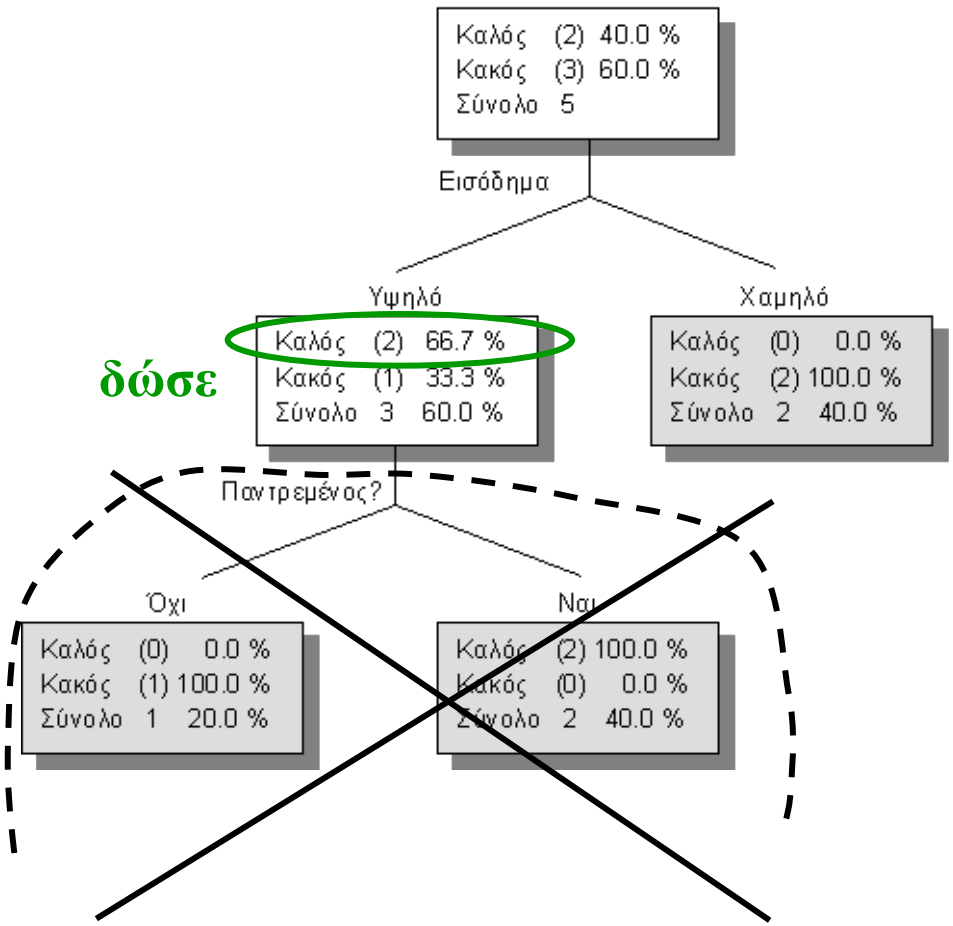
(**καλύτερη ικανότητα γενίκευσης**).

Υπερ-προσαρμογή (overfitting)

- **Υπερ-προσαρμογή** στα δεδομένα εκπαίδευσης:
  - Μάθηση **τυχαιοτήτων** των παραδειγμάτων εκπαίδευσης.
  - Η παραγόμενη υπόθεση έχει **μεγάλη συνέπεια** με τα παραδείγματα εκπαίδευσης αλλά **μικρή** ικανότητα **γενίκευσης**.
  - Πρόβλημα σε όλους τους αλγορίθμους μηχανικής μάθησης.
- Στον ID3, προσπαθώντας να διαχωρίσουμε πλήρως τα παραδείγματα εκπαίδευσης:
  - Μεγαλώνουμε το δέντρο εξετάζοντας **περισσότερες ιδιότητες**.
  - Αν δεν υπάρχουν ασυνεπή παραδείγματα και υπάρχουν αρκετές ιδιότητες, καταλήγουμε σε ένα δέντρο που είναι απολύτως **συνεπές** με τα δεδομένα εκπαίδευσης.
  - Μπορεί, όμως, το δέντρο να εξετάζει ιδιότητες που είναι **άσχετες** με το πρόβλημα (π.χ. χρώμα ματιών στα δάνεια), αλλά που **στα λίγα εναπομείναντα παραδείγματα** εκπαίδευσης ενός κόμβου **τυχαίνει** να **«προβλέπουν» σωστά** την κατηγορία.

Κλάδεμα του παραγόμενου δέντρου

- Αν **κλαδέψουμε** το δέντρο που παράγει ο ID3, ενδέχεται να έχει καλύτερη ικανότητα γενίκευσης.
  - Παράγουμε το δέντρο από τα **παραδείγματα εκπαίδευσης**.
  - Εξετάζουμε μήπως **αφαιρώντας ένα υποδέντρο** έχουμε υψηλότερο ποσοστό ορθότητας σε **ξεχωριστά παραδείγματα επικύρωσης** (validation data).
  - Αφού καταλήξουμε στο «καλύτερο» δέντρο, το αξιολογούμε σε **παραδείγματα ελέγχου** (test data).
- Ή **σταματάμε πρόωρα** την επέκταση του δέντρου.
  - Π.χ. όταν το 95% των παραδειγμάτων κάθε κόμβου ανήκει στην ίδια κατηγορία.

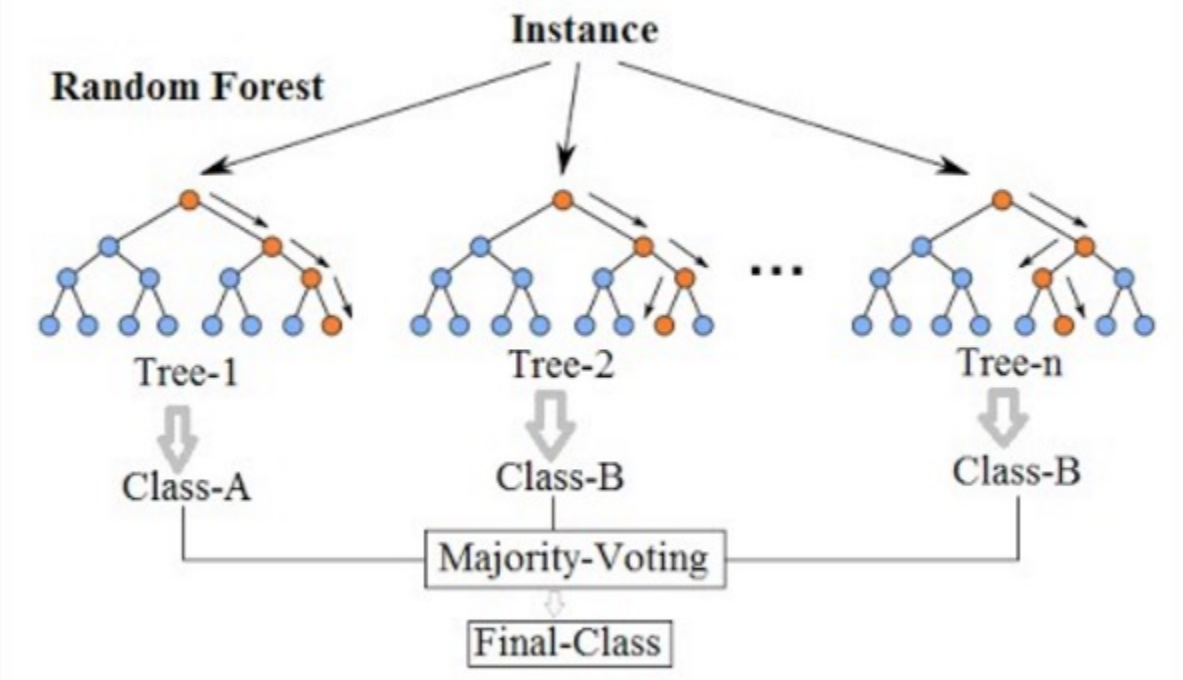


Άλλες βελτιώσεις του ID3

- **Ιδιότητες** με **πάρα πολλές τιμές**.
  - Στην ακραία περίπτωση, πρόβλημα αν υπάρχει **ιδιότητα-κλειδί**: προβλέπει πλήρως τη σωστή απόκριση στα δεδομένα εκπαίδευσης, αλλά ένα δέντρο που βασίζεται σε αυτή δεν έχει καμία ικανότητα γενίκευσης.
  - Υπάρχουν συναρτήσεις αξιολόγησης ιδιοτήτων που **τιμωρούν** ιδιότητες που οδηγούν σε υπερβολικό **κατακερματισμό** των παραδειγμάτων εκπαίδευσης (π.χ. gain ratio: κέρδος πληροφορίας που παρέχει η ιδιότητα / εντροπία ιδιότητας).
- **Ιδιότητες** με **μη διακριτές τιμές**.
  - Π.χ. ιδιότητες με τιμές **πραγματικούς** αριθμούς.
  - Μετατροπή σε **ιδιότητες διακριτών τιμών**.

    π.χ. X' = 1 ανν $0 \leq X < 1.2$, X' = 2 ανν $1.2 \leq X < 2.7$, …
  - Υπάρχουν αλγόριθμοι που εντοπίζουν τα καλύτερα **σημεία διαχωρισμού** των πεδίων τιμών συνεχών ιδιοτήτων.

Αλγόριθμος τυχαίου δάσους (random forest)



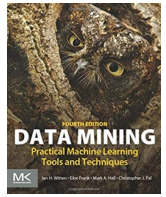
- Κατασκευάζουμε **πολλά δέντρα απόφασης** (δάσος):
  - **Κάθε δέντρο** εκπαιδεύεται σε **διαφορετική παραλλαγή του συνόλου εκπαίδευσης**, χρησιμοποιώντας **διαφορετικό υποσύνολο των διαθέσιμων ιδιοτήτων**.
  - Ακολουθούμε τη γνώμη της **πλειοψηφίας των δέντρων**.

Αλγόριθμος τυχαίου δάσους – συνέχεια

- **Σύνολα εκπαίδευσης** των δέντρων του δάσους:
  - Αν το **αρχικό σύνολο εκπαίδευσης** έχει ***N* παραδείγματα**, **κάθε παραλλαγή** του συνόλου εκπαίδευσης έχει **και αυτή *N* παραδείγματα**, που επιλέγονται **τυχαία με επανατοποθέτηση** από το αρχικό σύνολο εκπαίδευσης («bagging»).
- **Σύνολα ιδιοτήτων** των δέντρων του δάσους:
  - Αν το **αρχικό σύνολο διαθέσιμων ιδιοτήτων** έχει ***M* ιδιότητες**, δίνουμε **σε κάθε δέντρο** ένα **τυχαίο υποσύνολο *m* < *M* διαθέσιμων ιδιοτήτων** (*m* σταθερό για όλα τα δέντρα).
- Μορφή **συλλογικής μάθησης** (ensemble learning):
  - **Συνδυάζουμε ταξινομητές** που κάνουν **διαφορετικά λάθη**.
  - **Το Τυχαίο Δάσος έχει συνήθως καλύτερες επιδόσεις από μεμονωμένα δέντρα** (και συνήθως μικρότερη υπερ-εφαρμογή) αλλά γίνεται **πιο δύσκολη η εξήγηση των αποφάσεων**.

Βιβλιογραφία

- Russel & Norvig (4η έκδοση): ενότητες 19.3, 19.8.2, 20.2.2 (μόνο όσα αναφέρουν οι διαφάνειες).
  - Όσοι ενδιαφέρονται μπορούν να διαβάσουν προαιρετικά και τις υπόλοιπες ενότητες των κεφαλαίων 19 και 20.
- Βλαχάβας κ.ά: ενότητες 18.5, 18.8 (μόνο όσα αναφέρουν οι διαφάνειες).
  - Όσοι ενδιαφέρονται μπορούν να διαβάσουν προαιρετικά (εκτός εξεταστέας ύλης) και τις υπόλοιπες ενότητες του κεφαλαίου 18.
- Η γνωστότερη παραλλαγή του ID3 (με κλάδεμα) είναι ο C4.5.
  - Αντιστοιχεί στη μέθοδο J48 του Weka.
  - Βλ. βιβλίο των Witten & Frank.



Βιβλιογραφία – συνέχεια

- Υπάρχουν πολλές μορφές του αφελούς ταξινομητή Bayes.
  - Η μορφή που εξετάσαμε χρησιμοποιεί **δίτιμες (Boolean) ιδιότητες** και λέγεται «**πολυμεταβλητή μορφή Bernoulli**» (multivariate Bernoulli Naive Bayes).
  - Η **πολυωνυμική** (multinomial) μορφή του NB μπορεί να λάβει υπόψη της και τις **συχνότητες των λέξεων** σε κάθε κείμενο κατά την κατάταξη κειμένων. Για περισσότερες πληροφορίες, βλ. http://www.aueb.gr/users/ion/docs/ceas2006_paper.pdf.
  - Περισσότερες πληροφορίες για τις μορφές του NB και άλλα θέματα που αναφέραμε (π.χ. κέρδος πληροφορίας) παρέχονται στο κεφάλαιο 13 του βιβλίου «An introduction to Information Retrieval» των C.D. Manning, P. Raghavan και H. Schütze. Διατίθεται ελεύθερα: https://nlp.stanford.edu/IR-book/information-retrieval-book.html