

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

**Δένδρα Αποφάσεων (Decision Trees)**

**Αλγόριθμοι Διαμόρφωσης CART (Classification And  
Regression Trees), Gini Index  
Random Forests**

**Αλγόριθμοι Bagging (Bootstrap & aggregating)**

καθ. Βασίλης Μάγκλαρης

[maglaris@netmode.ntua.gr](mailto:maglaris@netmode.ntua.gr)

[www.netmode.ntua.gr](http://www.netmode.ntua.gr)

Αίθουσα 02, Νέα Κτίρια ΣΗΜΜΥ

Τρίτη 31/5/2022

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Δένδρα Αποφάσεων

James Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning", Springer 2013 <https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf> (Ch. 8)

- Το δένδρο διαμορφώνεται συνήθως με αρχική ένταξη των παραδειγμάτων σε διακριτές κατηγορίες τιμών (*categorical value ranges*) των χαρακτηριστικών τους (*features, attributes*)
- Από ένα δείγμα μάθησης (*training sample*)  $\mathcal{D} = \{(\mathbf{x}(1), d(1)), \dots, (\mathbf{x}(N), d(N))\}$  παραδειγμάτων εισόδου (*predictor variables*)  $\mathbf{x}(i)$  με γνωστά *labels*  $d(i)$ , δημιουργείται δένδρο αποφάσεων (**Decision Tree**)  $\mathbf{x}(i) \rightarrow y(i) \approx d(i)$  με αλγόριθμο *supervised learning*
- Νέα παραδείγματα εισόδου (από το *test sample*) αποφασίζουν για την πιθανότερη  $\mathbf{x}(i) \rightarrow y(i)$ , θεωρώντας πως ακολουθούν στατιστικές ιδιότητες του δείγματος μάθησης

Ορίζονται δύο τύποι *Decision Trees*:

**(1) Classification Trees** (Δένδρα Ταξινόμησης) όταν η απόφαση  $y(i)$  είναι διακριτή, συνήθως δυαδική ( $C_0 = \text{yes}$ ,  $C_1 = \text{no}$ ) ή  $y(i) \in \{0,1\}$

**(2) Regression Trees** (Δένδρα Παλινδρόμησης) όταν η  $y(i)$  είναι συνεχής

- **Εφαρμογές:** Αναγνώριση προτύπων, επεξεργασία - ταξινόμηση - αναζήτηση κειμένων και εικόνων, περιβαλλοντολογικές αναλύσεις, βιοϊατρικές προβλέψεις, ιατρικές διαγνώσεις, μηχανές αναζήτησης, διαγνώσεις/χαρακτηρισμός/ αντιμετώπιση κυβερνο-επιθέσεων (π.χ. DNS attacks, antivirus, anti-spamming...)
- **Βασικά πλεονεκτήματα:** Εύκολη *μη παραμετρική* υλοποίηση με πολυπληθή δείγματα, κατάταξη χαρακτηριστικών και ταξινόμηση με λογική παρόμοια με ανθρώπινη λειτουργία...
- **Βασικά μειονεκτήματα:** Ανάγκη για *labeled training datasets*, προβλήματα ακρίβειας, αστάθεια σε μικρές μεταβολές των χαρακτηριστικών εισόδου, *overfitting*...

Αντιμετώπιση προβλημάτων με συνδυασμό πολλαπλών decision trees (**Random Forests**)

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Διαμόρφωση Δένδρων Αποφάσεων με Supervised Learning

Μέθοδος CART - Classification And Regression Trees (*Leo Breiman, 1984*)

[https://civil.colorado.edu/~balajir/CVEN6833/lectures/cluster\\_lecture-2.pdf](https://civil.colorado.edu/~balajir/CVEN6833/lectures/cluster_lecture-2.pdf)

Η δενδρική δομή διαμορφώνεται με βάση τα **χαρακτηριστικά** (*features, attributes*) δειγματικών στοιχείων (παραδειγμάτων) μάθησης, κωδικοποιημένα σε τιμές  $x_j(i)$  (predictor value) **διανυσμάτων εισόδου**  $x(i)$ . Ορίζω πρόσθετο **χαρακτηριστικό εξόδου**  $y(i)$  με **τιμή στόχου** (*target value*). Για τα *labeled* παραδείγματα μάθησης θέλουμε  $y(i) \approx d(i)$

- Ξεκινώ από τον κόμβο **ρίζα** (*Root Node*,  $m = 0$ ) που εμπεριέχει όλα τα παραδείγματα  $x(i)$  του δείγματος μάθησης  $\mathcal{D}$  με τα χαρακτηριστικά εισόδου  $x_j(i)$  και εξόδου  $y(i)$ . Διαχωρίζω το σύνολο παραδειγμάτων σε 2 υποσύνολα (υπο-δένδρα) με βάση την **κατηγοριοποίηση** κάποιου χαρακτηριστικού εισόδου (*attribute*) σε περιοχές ορισμένου εύρους (*range*)
- Συνεχίζω τον διαχωρισμό σε 2 υπο-δένδρα στον κόμβο  $m$  μέχρι την διαμόρφωση τελικών φύλλων συμμετοχής όλων των παραδειγμάτων μάθησης

**Τελικός στόχος Αλγορίθμου Διαμόρφωσης Δένδρου CART:**

- (1) Για *classification trees* στόχος είναι η ένταξη στα τελικά φύλλα του δένδρου των παραδειγμάτων μάθησης με ενιαία απόκριση  $y(i)$  κατά μεγάλη πλειοψηφία
- (2) Για *regression trees* στόχος είναι να ενταχθούν στα φύλλα του δένδρου ομάδες παραδειγμάτων μάθησης με μικρές αποκλίσεις εξόδου από μέσο όρο αποκρίσεων (π.χ. ελάχιστης τετραγωνικής απόκλισης ή με κριτήριο greedy)
- (3) Συνήθως απαιτείται απλοποίηση του τελικού δένδρου με κλάδεμα (*pruning*) κλαδιών μικρής πιθανότητας συμμετοχής ενός στοιχείου μάθησης για αποφυγή *overfitting*

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Διαμόρφωση Classification Trees Δυαδικής Εξόδου $y(i) \in \{0,1\}$

Έστω ότι ο κόμβος  $m$  του δένδρου αποτελεί υπο-δένδρο για περιοχή (*region*)  $R_m$  όπως ορίζεται από το εύρος (*range*) τιμών ενός χαρακτηριστικού (*attribute*) στοιχείου εισόδου, κωδικοποιημένο στις διαστάσεις (*predictor variables*) του διανύσματος  $\mathbf{x}(i) \in R_m$ . Η αναλογία των αποκρίσεων (*target values*)  $y(i)$  στοιχείων εισόδου  $\mathbf{x}(i) \rightarrow (y(i) = k)$  στα  $N_m$  στοιχεία του  $m$  είναι

$$\hat{p}_{mk} \triangleq \frac{1}{N_m} \sum_{\mathbf{x}(i) \in R_m} I(y(i) = k)$$

- Η **πλειοψηφία** ταξινομήσεων του κόμβου  $m$  γίνονται στη κλάση  $k(m) = \operatorname{argmax}_k \hat{p}_{mk}$  που καθορίζει τη ταξινόμηση όλων των  $\mathbf{x}(i) \in R_m$ . Η κάθε διαφορετική ταξινόμηση θεωρείται παραφωνία (**ακαθαρσία**, *impurity*)
- Για 2 κλάσεις ( $C_0$  και  $C_1$ ) μέτρο **ακαθαρσίας** (*impurity*) ορίζεται ο **συντελεστής Gini Index**:  $GI(p) = 2p(1 - p) = 1 - (1 - p)^2 - p^2$  όπου  $p$  η αναλογία ασύμβατης (λάθος) ταξινόμησης δειγματικού σημείου μάθησης  $\{\mathbf{x}(i), d(i) = 0\}$  του υπο-δένδρου  $m$  σε έξοδο  $y(i) = 1 \neq d(i)$ . Σε περίπτωση **απόλυτης συμβατότητας**,  $p = 0$  ή  $p = 1$  και  $GI(p) = 0$  (**ελαχίστη τιμή**). Η **μέγιστη ασυμβατότητα**  $p = 0.5$  δίνει τη **μέγιστη τιμή**  $GI(p) = 0.5$
- Το χαρακτηριστικό εισόδου (*attribute*) για επόμενο διαχωρισμό σε υπο-δένδρα προκύπτει από σύγκριση των εναλλακτικών και επιλογή της **ελάχιστης** τιμής του **Gini Index**  
**Σημείωση**: Παρόμοιο μέτρο επιλογής χαρακτηριστικού αφορά στη ελαχιστοποίηση του **Δείκτη Εντροπίας**:  $-p \log_2(p) - (1 - p) \log_2(1 - p)$

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα Διαμόρφωσης Δένδρου Ταξινόμησης με Χρήση του Gini Index (1/2)

Το δείγμα μάθησης (*training sample*) αποτελείται από 16 στοιχεία (*records*) με 4 αριθμητικά συνεχή χαρακτηριστικά εισόδου (*attribute values, predictors*) A, B, C, D και γνωστή δυαδική έξοδο (*label, target*):  
 $E \in \{\text{positive, negative}\}$

Επιλέγεται *αυθαίρετα* κατηγοριοποίηση των *attributes* σε 2 κατηγορίες ανάλογα με το ύψος της τιμής τους:

A	B	C	D
$\geq 5.0$	$\geq 3.0$	$\geq 4.2$	$\geq 1.4$
$< 5.0$	$< 3.0$	$< 4.2$	$< 1.4$

### Gini Index $GI$ , Attribute A

$A \geq 5.0$ : 12/16,  $A < 5.0$ : 4/16

$A \geq 5.0$  & E positive: 5/12

$A \geq 5.0$  & E negative: 7/12

$$GI(5,7) = 1 - (5/12)^2 - (7/12)^2 = 0.486$$

$A < 5.0$  & E positive: 3/4

$A < 5.0$  & E negative: 1/4

$$GI(3,1) = 1 - (3/4)^2 - (1/4)^2 = 0.375$$

$$GI(A) = 12/16 \times 0.486 + 4/16 \times 0.375 = 0.45825$$

	A	B	C	D	E
1	4.8	3.4	1.9	0.2	positive
2	5	3	1.6	0.2	positive
3	5	3.4	1.6	0.4	positive
4	5.2	3.5	1.5	0.2	positive
5	5.2	3.4	1.4	0.2	positive
6	4.7	3.2	1.6	0.2	positive
7	4.8	3.1	1.6	0.2	positive
8	5.4	3.4	1.5	0.4	positive
9	7	3.2	4.7	1.4	negative
10	6.4	3.2	4.5	1.5	negative
11	6.9	3.1	4.9	1.5	negative
12	5.5	2.3	4	1.3	negative
13	6.5	2.8	4.6	1.5	negative
14	5.7	2.8	4.5	1.3	negative
15	6.3	3.3	4.7	1.6	negative
16	4.9	2.4	3.3	1	negative



# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα Διαμόρφωσης Δένδρου Ταξινόμησης με Χρήση του Gini Index (2/2)

### Gini Index $GI$ , Attribute B

Όπως και για το **Attribute A** προκύπτει

$$GI(B) = 0.3345$$

### Gini Index $GI$ , Attribute C

Όπως και για το **Attribute A** προκύπτει

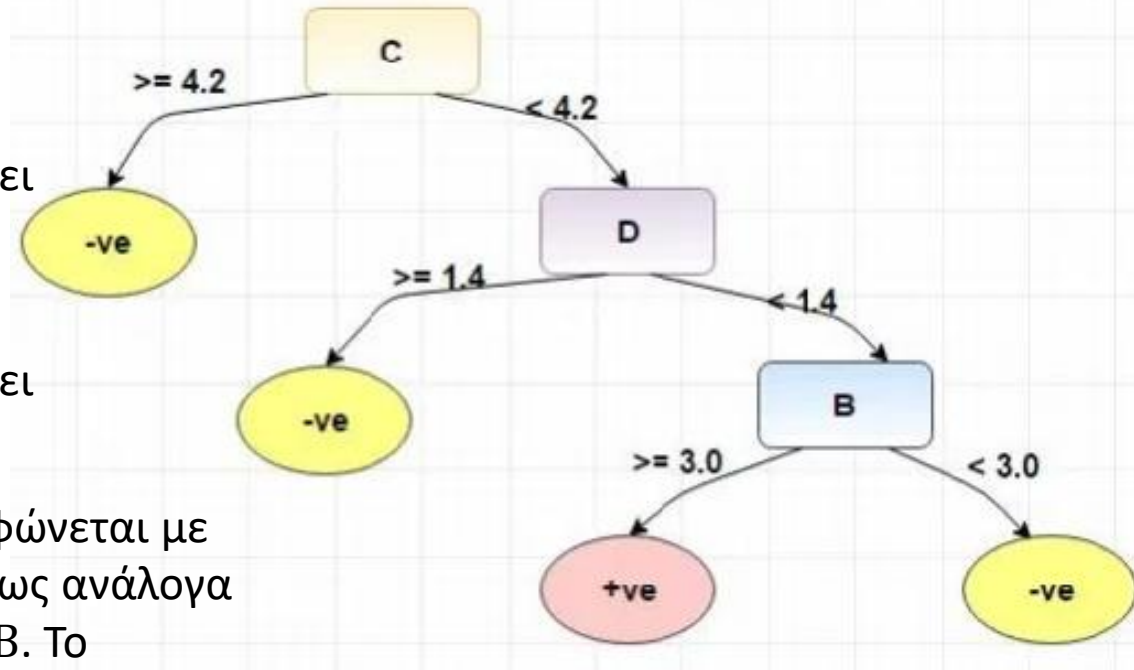
$$GI(C) = 0.2$$

### Gini Index $GI$ , Attribute D

Όπως και για το **Attribute A** προκύπτει

$$GI(D) = 0.273$$

Άρα το **Δένδρο Αποφάσεων** διαμορφώνεται με επιλογή χαρακτηριστικών αντιστρόφως ανάλογα με τις τιμές των  $GI$  κατά σειρά: C, D, B. Το χαρακτηριστικό A δεν επηρεάζει τη διαδικασία διαμόρφωσης



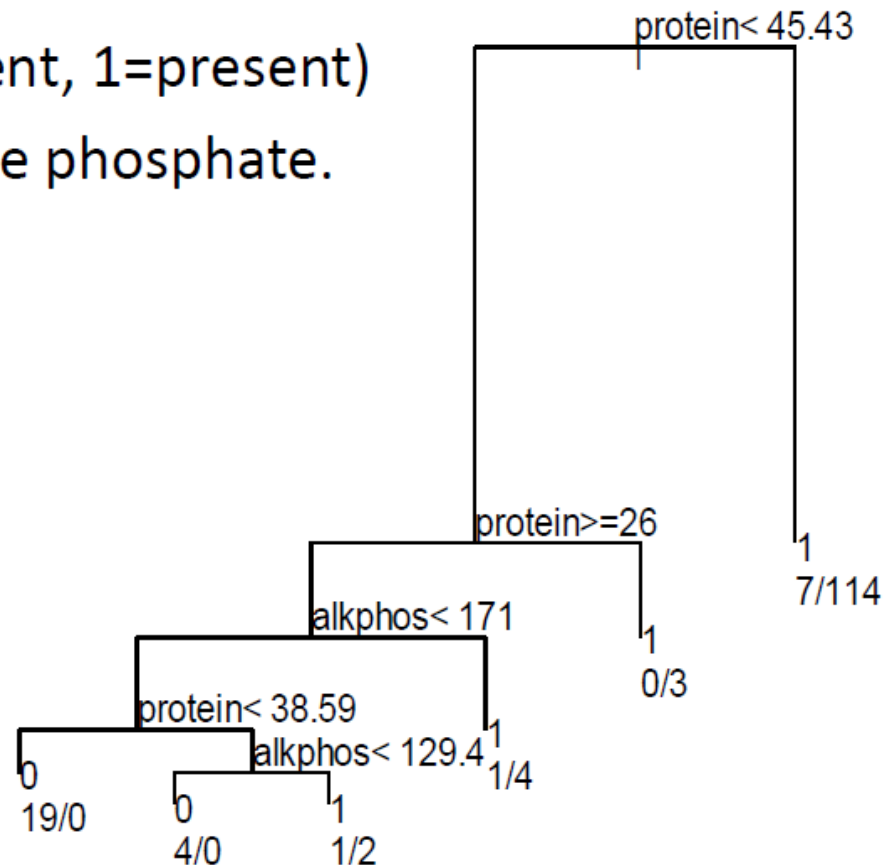
Attribute	A	B	C	D
Lower Value	$\geq 5.0$	$\geq 3.0$	$\geq 4.2$	$\geq 1.4$
Upper Value	$< 5.0$	$< 3.0$	$< 4.2$	$< 1.4$
Gini Index	<b>0.45825</b>	<b>0.3345</b>	<b>0.2</b>	<b>0.273</b>

Βασισμένο στη παρουσίαση της **Adele Cutler** (Utah State University), “Random Forests for Regression and Classification”, Ovronnaz, Switzerland, Sep. 2010 <https://math.usu.edu/adele/randomforests/ovronnaz.pdf>

## A Classification Tree

Predict hepatitis (0=absent, 1=present)  
using protein and alkaline phosphate.

“Yes” goes left.

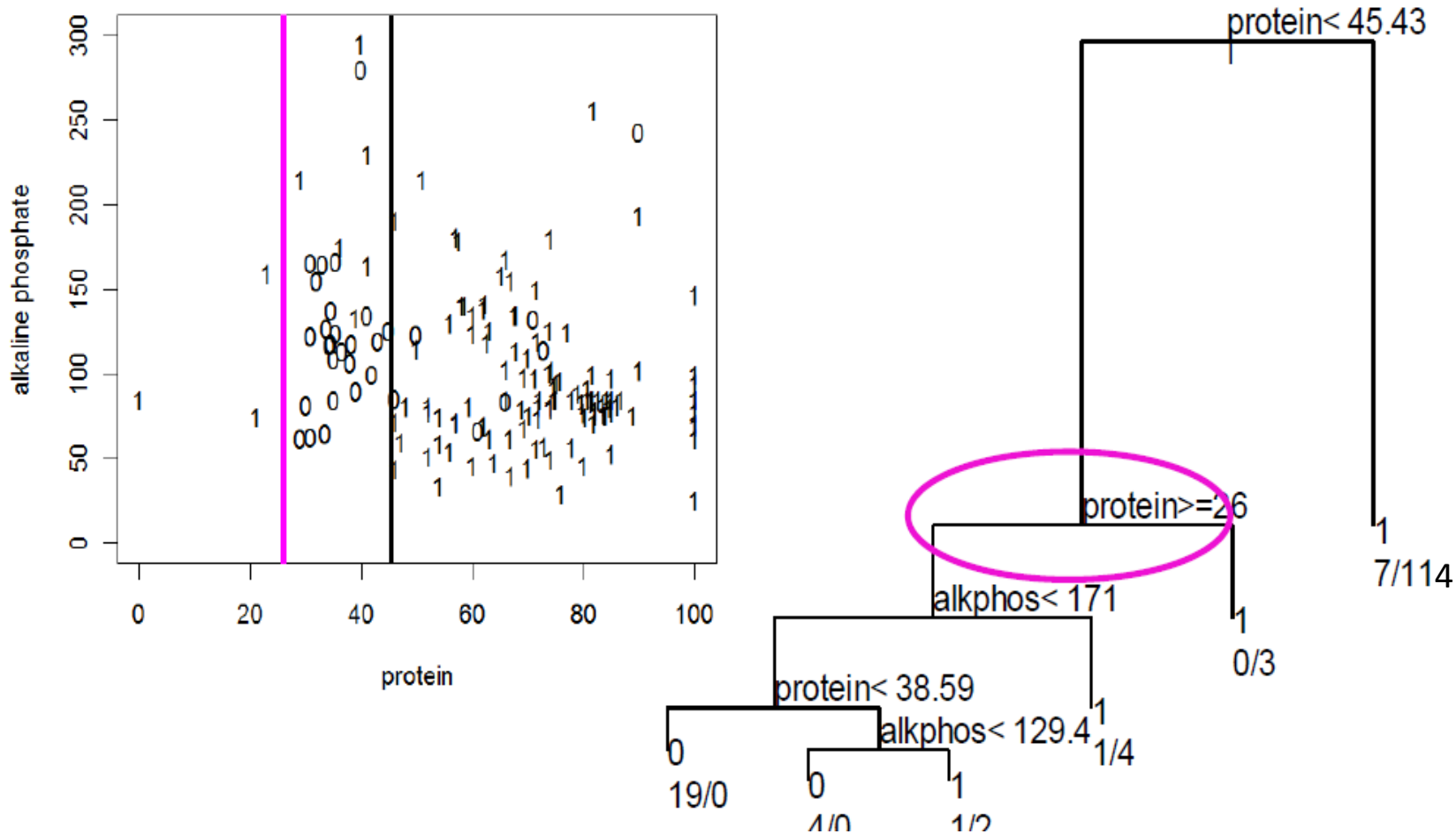






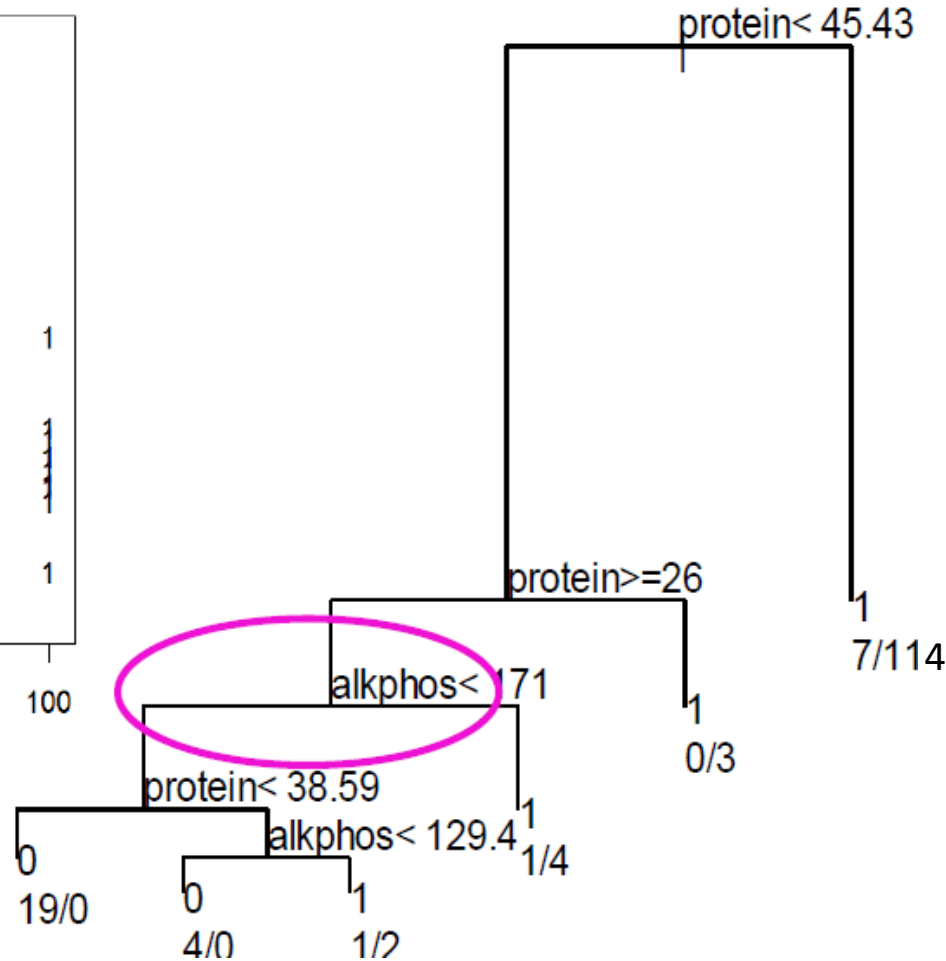
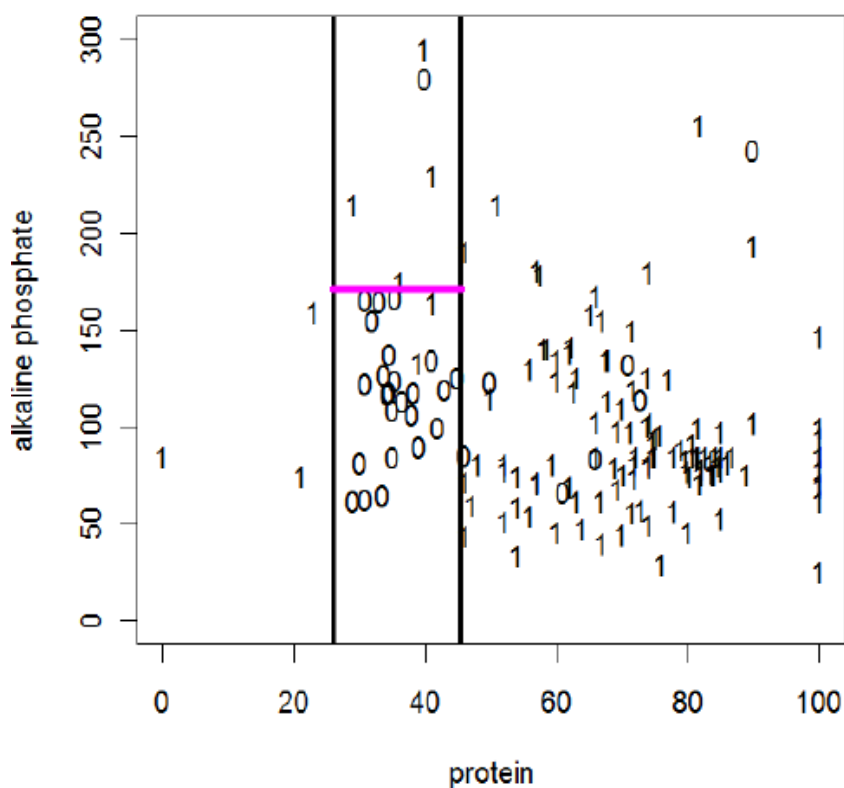
# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα: Πρόβλεψη Ηπατίτιδας με Δένδρο Ταξινόμησης (3/5)



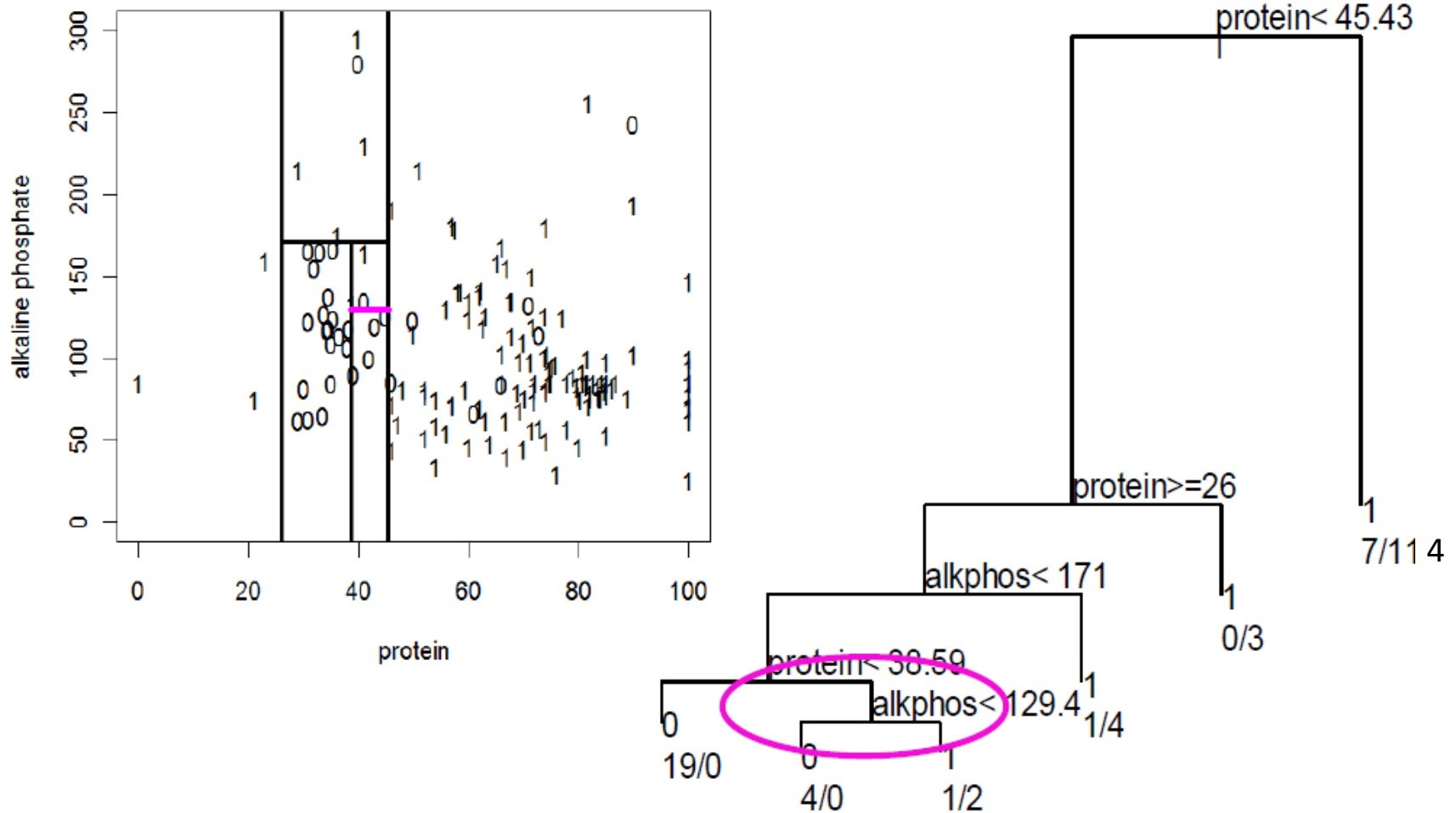
# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα: Πρόβλεψη Ηπατίτιδας με Δένδρο Ταξινόμησης (4/5)



# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα: Πρόβλεψη Ηπατίτιδας με Δένδρο Ταξινόμησης (5/5)



Το χαρακτηριστικό *protein* ορίζει αποκλειστικά τον διαχωρισμό για μεγάλες και μικρές τιμές του  
Το χαρακτηριστικό *alkaline phosphate* υπεισέρχεται μόνο για ενδιάμεσες τιμές του *protein*

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Διαμόρφωση Random Forests – Bootstrap Aggregating (Bagging)

Leo Breiman, “Random Forests”, Machine Learning, 45, 5-32, Kluwer Academic Publishers 2001

<https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf>

Leo Breiman, “Bagging Predictors”, Machine Learning, 24, 123-140, Kluwer Academic Publishers 1996

<https://link.springer.com/content/pdf/10.1007/BF00058655.pdf>

- Ένα **Random Forest** αποτελείται από πολλά ( $R$ ) **decision trees** τα οποία αποφασίζουν **ανεξάρτητα** για μια παράμετρο εξόδου  $y(i)$ . Η τελική απόφαση προκύπτει από την πλειοψηφία (**aggregation**) των  $y(i)$ : **Voting** για **classification**, **averaging** για **regression**
- Η διαδικασία επιβλεπόμενης μάθησης ακολουθεί τον αλγόριθμο **Bootstrap Aggregating (Bagging)** του **Leo Breiman (1996)**
- Με τυχαίο τρόπο επιλέγουμε  $R$  δειγματικά υποσύνολα (**bootstrap samples**) του **labeled** δείγματος μάθησης  $\mathcal{D} = \{(\mathbf{x}(1), d(1)), \dots, (\mathbf{x}(N), d(N))\}$ . Default επιλογή  $R = 500$  αλλά τα αποτελέσματα είναι συνήθως ικανοποιητικά για μικρές τιμές (π.χ.  $R = 20$ )
- Η επιλογή **παραδειγμάτων** σε υποσύνολα γίνεται με **ανεξάρτητες** τυχαίες κλήσεις και επανατοποθέτηση των επιλογών προηγούμενων κλήσεων (**sampling with replacement**). Ένα παράδειγμα (**observation**) σε **bootstrap sample** μπορεί να επιλεγεί πάνω από μια φορά. Τα παραδείγματα του δείγματος που δεν έχουν επιλεγεί σε ένα υποσύνολο αποτελούν **out of bag (oob) observations** (πολλά **oob** οδηγούν σε αδυναμία πρόβλεψης)
- Τα **bootstrap trees** διαμορφώνονται σε όλο τους το βάθος με αλγόριθμο τύπου **CART (Classification And Regression Trees)** αλλά με **μειωμένα χαρακτηριστικά (prediction attributes)** που επιλέγονται **τυχαία** και **ανεξάρτητα** σε κάθε κόμβο των δένδρων

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Αξιολόγηση & Επιδόσεις Random Forests – Αλγορίθμου Bagging

- Τρόπος αξιολόγησης **crossvalidation**: Ένα μεγάλο **labeled** δείγμα μπορεί να χωριστεί σε υποσύνολα  $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots$ . Κάθε  $\mathcal{D}_k$  μπορεί εναλλακτικά να θεωρηθεί **test set** και τα γνωστά **labels** των στοιχείων του να συγκριθούν με τις προβλέψεις που προκύπτουν με σύστημα που εκπαιδεύεται με βάση τα υπόλοιπα υποσύνολα σαν **supervised training sets**
- Για τα **Random Forests (RF)** η επίδοση μπορεί να αξιολογηθεί από τα σφάλματα προβλέψεων δένδρων αποφάσεων (**RF Predictors**) με *out of bag (oob) observations*
- Από πολλές αξιολογήσεις προκύπτει πως τα **RF** δεν παρουσιάζουν τυπικά προβλήματα των **Decision Trees**: Έχουν καλές επιδόσεις ακρίβειας προβλέψεων με μικρή διακύμανση (**variance**), δεν έχουν υπερβολική ευαισθησία σε ασταθείς μετρήσεις χαρακτηριστικών των **predictors**, δεν προκαλούν **overfitting**
- Γενικά θεωρούνται πολλά υποσχόμενη λύση για προβλήματα ταξινόμησης, πρόβλεψης και συγκριτικής αξιολόγησης των χαρακτηριστικών (**attributes**) πολυδιάστατων δειγμάτων. Προαπαιτούν βέβαια την διαθεσιμότητα αξιόπιστων **labeled** δειγμάτων μάθησης για την εφαρμογή μεθόδων **supervised learning**
- Για πολύ μεγάλα datasets οι απαιτήσεις μνήμης – επεξεργασίας των RF οδηγούν σε χρήση εξειδικευμένου H/W (**GPU**) ή λύσεις **cloud**