ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό

Συγκεντρωτικές μέθοδοι

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

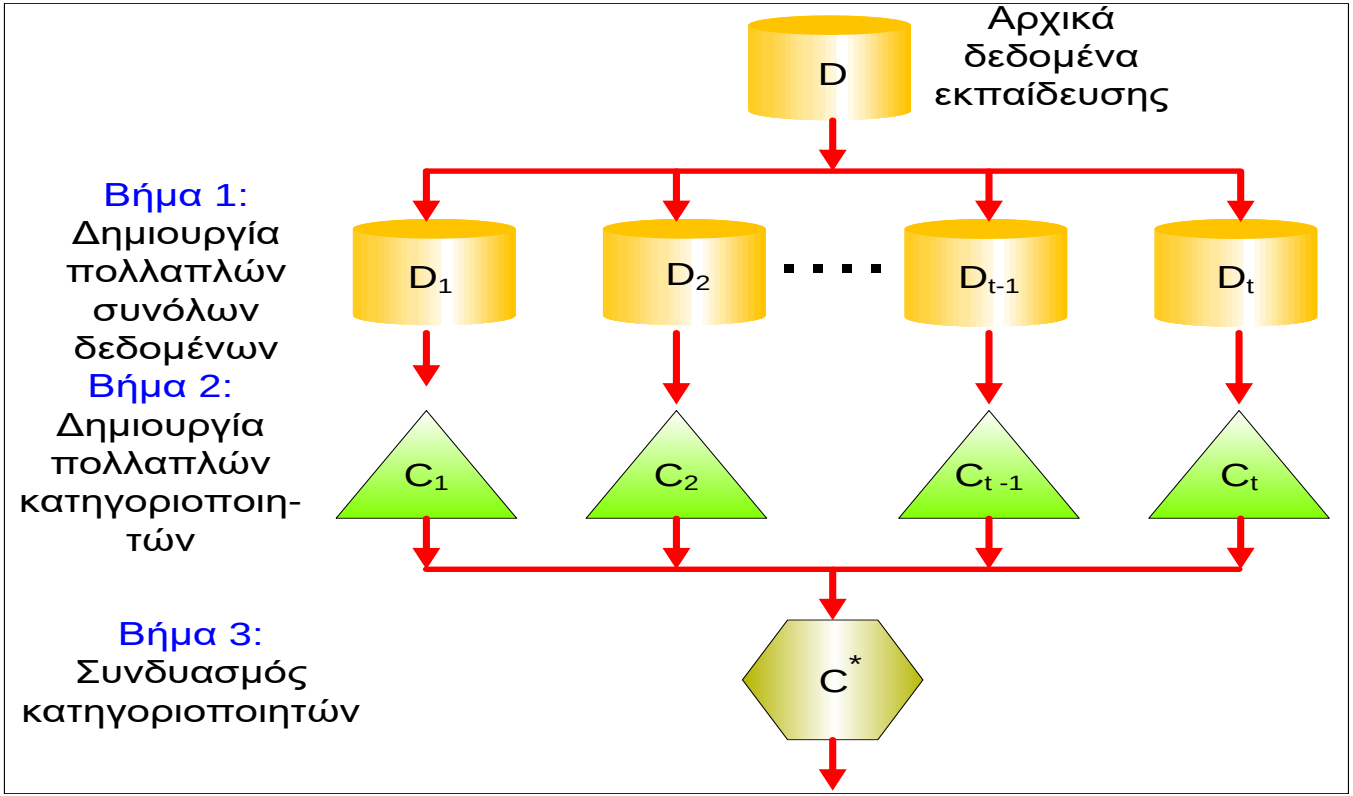
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Συγκεντρωτικές μέθοδοι

- Κατασκευή ενός συνόλου κατηγοριοποιητών από τα σώματα δεδομένων

- Η πρόβλεψη της κλάσης ενός νέου παραδείγματος γίνεται αθροίζοντας τις προβλέψεις των διαφόρων κατηγοριοποιητών

Γενική ιδέα

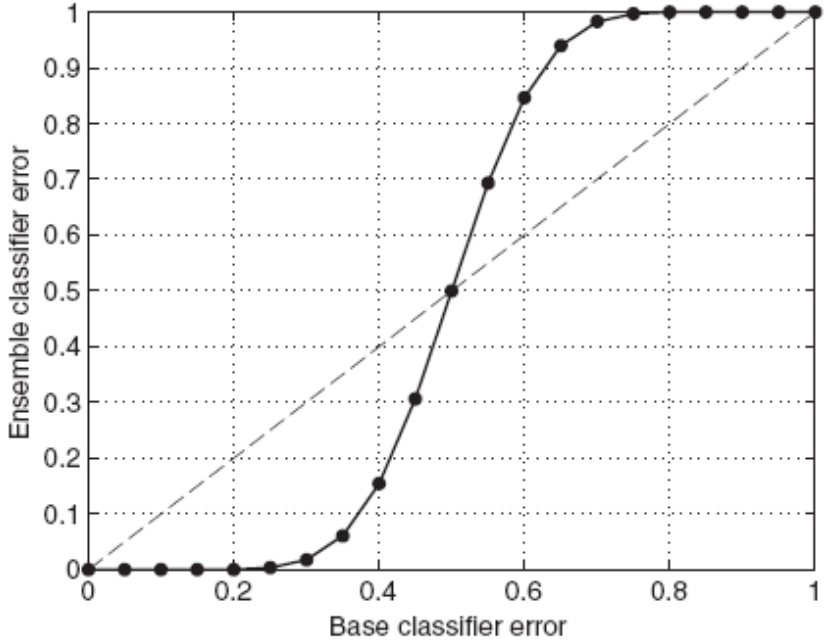

D
Αρχικά δεδομένα εκπαίδευσης
Βήμα 1: Δημιουργία πολλαπλών συνόλων δεδομένων
D1
D2
Dt-1
Dt
Βήμα 2: Δημιουργία πολλαπλών κατηγοριοποιητών
C1
C2
Ct -1
Ct
Βήμα 3: Συνδυασμός κατηγοριοποιητών
C*

Για ποιο λόγο λειτουργεί;

- Έστω ότι υπάρχουν 25 κατηγοριοποιητές
  - Κάθε ένας με σφάλμα (error rate), $\varepsilon = 0.35$
  - Υποθέτουμε ότι είναι ανεξάρτητοι
  - Πιθανότητα ενός συγκεντρωτικού κατηγοριοποιητή να κάνει εσφαλμένη πρόβλεψη:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Σύγκριση σφάλματος



Παραδείγματα

- Πως φτιάχνουμε ένα συγκεντρωτικό κατηγοριοποιητή;
- Πειράζοντας τα Δεδομένα
  - Bagging
  - Boosting
- Πειράζοντας τις Ιδιότητες
  - Random Forests

Bagging

- Δειγματοληψία με αντικατάσταση

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Κατασκευή ενός κατηγοριοποιητή σε κάθε δείγμα εκκίνησης (bootstrap)

- Κάθε δείγμα έχει πιθανότητα $(1 – 1/n)^n$ να επιλεχθεί

Παράδειγμα

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1
True Class	1	1	1	-1	-1	-1	-1	1	1	1

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.8	0.9	1	1	1
y	1	1	1	-1	-1	1	1	1	1	1

x <= 0.65 ==> y = 1
x > 0.65 ==> y =

-1

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.3 ==> y = 1
x > 0.3 ==> y = -1

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

x <= 0.05 ==> y = -1
x > 0.05 ==> y = 1

Boosting

- Μια επαναληπτική διαδικασία που προσαρμοστικά αλλάζει την κατανομή των δεδομένων εκπαίδευσης επικεντρώνοντας σε παραδείγματα που προηγουμένως έχουν ταξινομηθεί λανθασμένα
  - Αρχικά, αναθέτουμε ίσα βάρη σε όλες τις N εγγραφές
  - Αντίθετα με το Bagging, τα βάρη μπορεί να αλλάζουν στο τέλος κάθε γύρου

Boosting

- Οι εγγραφές που κατηγοριοποιούνται λανθασμένα **αυξάνουν** τα βάρη τους
- Οι εγγραφές που κατηγοριοποιούνται ορθά **μειώνουν** τα βάρη τους

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Το παράδειγμα 4 είναι δύσκολο να κατηγοριοποιηθεί
- Το βάρος του αυξάνεται, επομένως είναι πιο πιθανό να επιλεγεί στον επόμενο γύρο

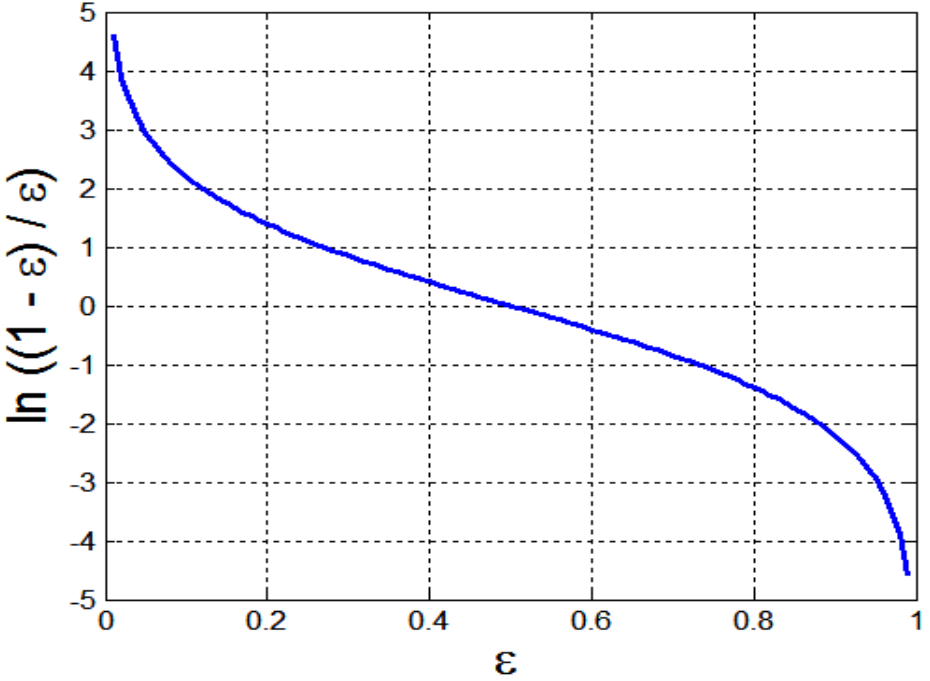
Παράδειγμα: Adaboost

- Κατηγοριοποιητές βάσης: C_1, C_2, …, C_T
- Error rate:

$$\varepsilon_i = \frac{1}{N}\sum_{j=1}^{N} w_j \delta\left(C_i(x_j) \neq y_j\right)$$

- Σημαντικότητα ενός κατηγοριοποιητή:

$$\alpha_i = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$



Παράδειγμα: Adaboost

- Αναβάθμιση βαρών:

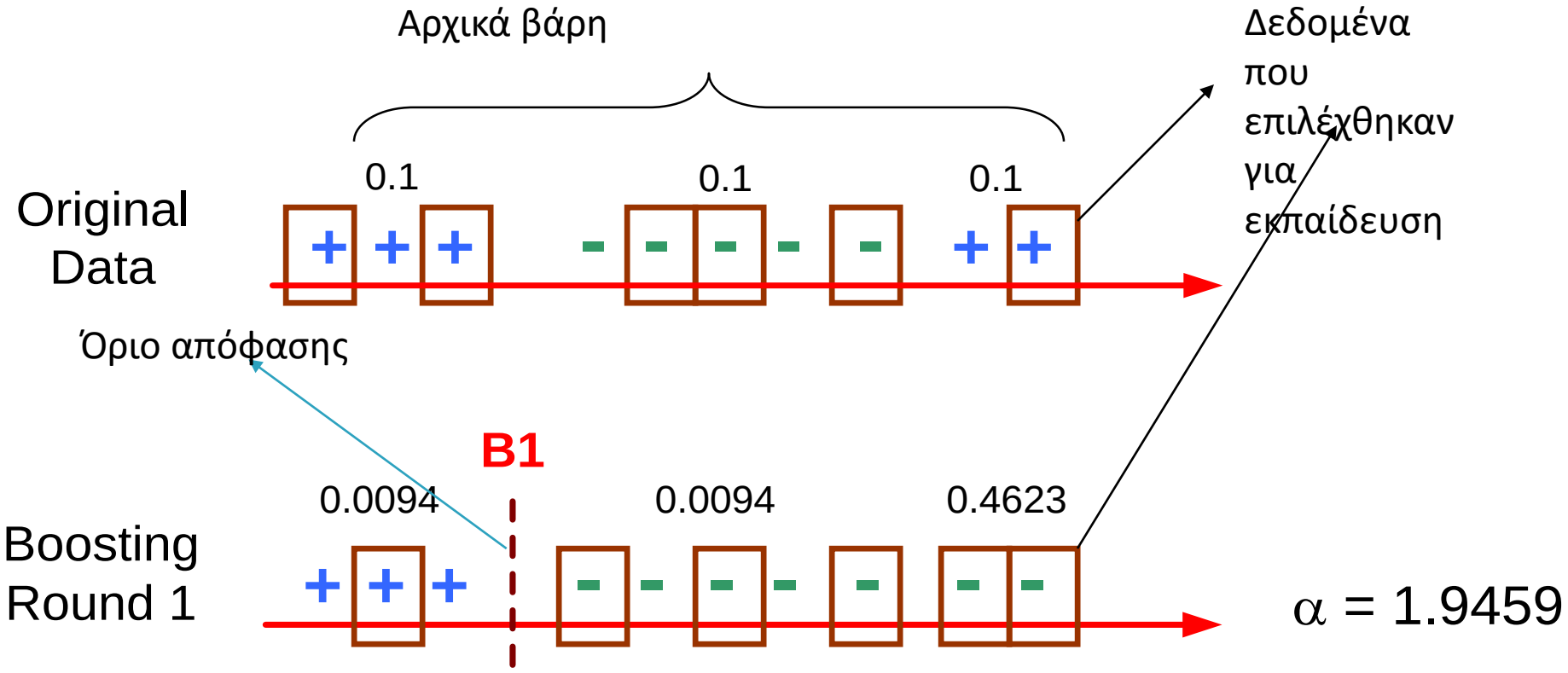
$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \begin{cases} \exp^{-\alpha_j} & \text{αν } C_j(x_i) = y_i \\ \exp^{\alpha_j} & \text{αν } C_j(x_i) \neq y_i \end{cases}$$

όπου Z_j παράγοντας κανονικοποίησης

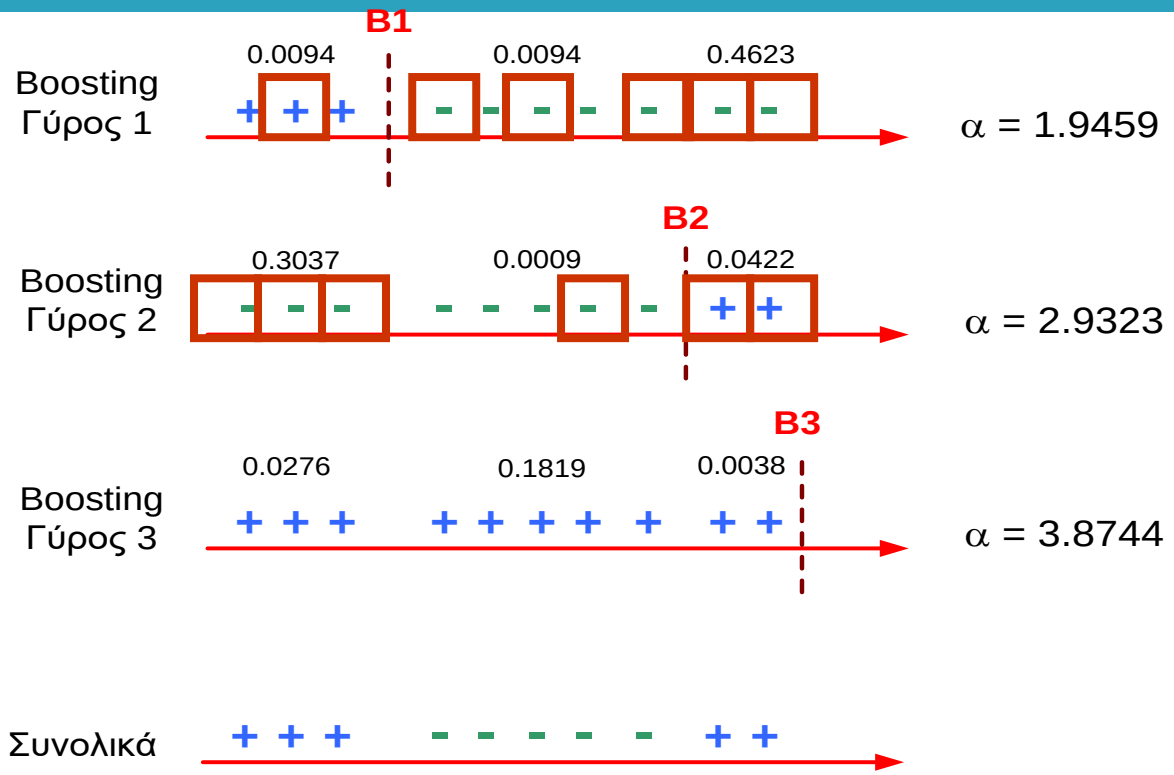
- Αν κάποιος ενδιάμεσος γύρος παράγει σφάλμα μεγαλύτερο του 50%, τα βάρη ξαναγίνονται ίσα με 1/n και επαναλαμβάνεται η δειγματοληψία
- Κατηγοριοποίηση:

$$C^*(x) = \underset{y}{\operatorname{argmax}} \sum_{j=1}^{T} \alpha_j \delta\left(C_j(x) = y\right)$$

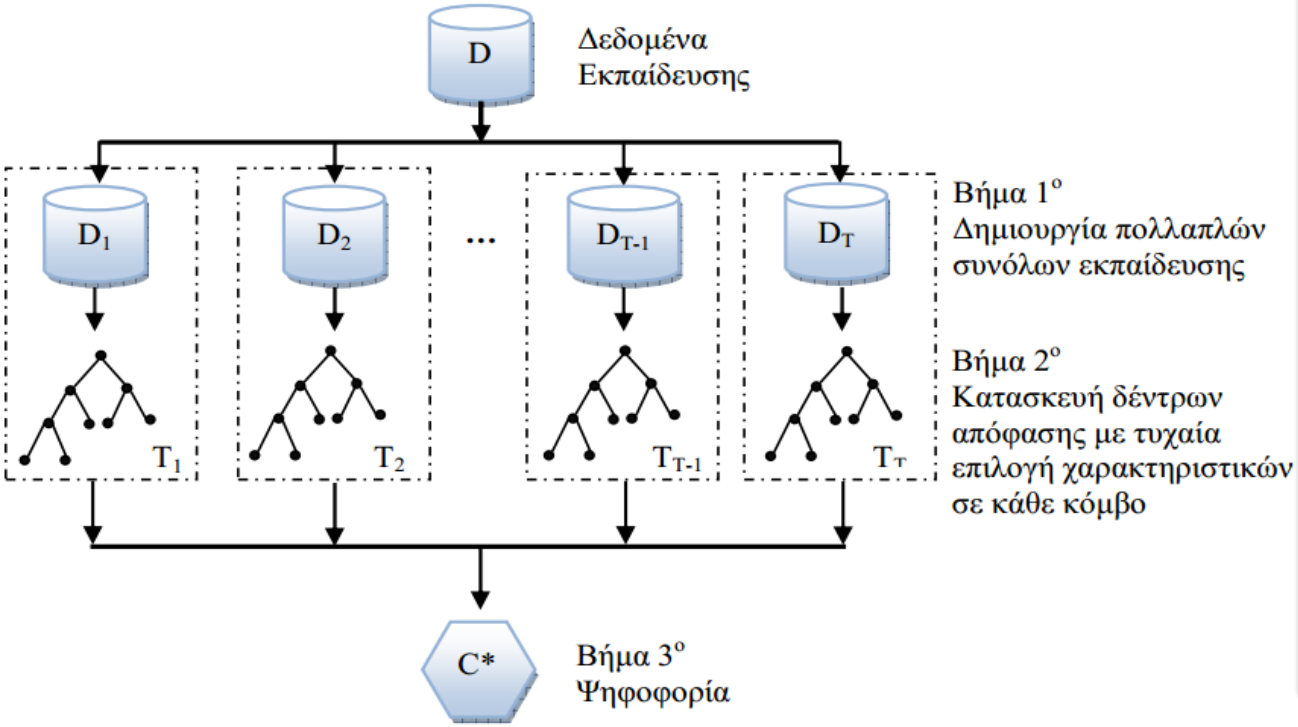
AdaBoost


Αρχικά βάρη
Δεδομένα που επιλέχθηκαν για εκπαίδευση
Original Data
0.1
0.1
0.1
+ + + - - - - - + +
Όριο απόφασης
B1
Boosting Round 1
0.0094
0.0094
0.4623
+ + + - - - - - - -
α = 1.9459

AdaBoost


B1
0.0094
0.0094
0.4623
Boosting
Γύρος 1
α = 1.9459
B2
0.3037
0.0009
0.0422
Boosting
Γύρος 2
α = 2.9323
B3
0.0276
0.1819
0.0038
Boosting
Γύρος 3
α = 3.8744
Συνολικά

Random Forests



- Μπορούν να εκπαιδευτούν σε σύνολα δεδομένων υψηλής διάστασης όπως είναι τα κείμενα και οι εικόνες, χωρίς να εμφανίσουν σημαντικό βαθμό overfitting
- Παρουσιάζουν ανεκτικότητα ως προς το όρυβο και αριθμητικών σφαλμάτων στα δεδομένα εκπαίδευσης (π.χ. απόκρυψη μέρους του αντικειμένου, ελλιπή δεδομένα).
- Για την επαγωγή κάθε δέντρου περίπου το 1/3 των παραδειγμάτων δεν επιλέγεται για εκπαίδευση. Αυτά τα παραδείγματα καλούνται Out-of-Bag παραδείγματα και μπορούν να χρησιμοποιηθούν για την εκτίμηση της πιθανότητας σφάλματος, εξαλείφοντας την ανάγκη ύπαρξης ενός συνόλου ελέγχου ή εφαρμογής της τεχνικής cross-validation