Δένδρα απόφασης

Ιωάννης Μαδεμλής

Δένδρα απόφασης

Δένδρα απόφασης

- Τα **δένδρα απόφασης** είναι από τους συνηθέστερους αλγορίθμους ταξινόμησης.
  - Μη παραμετρικός, δεν προκύπτει ως λύση ενός φορμαλιστικού προβλήματος βελτιστοποίησης.
- Επιχειρούν να απαντήσουν σε ερωτήματα ταξινόμησης ενός προτύπου σταδιακά.
  - «Θέτουν» διαδοχικά κατάλληλα «ερωτήματα» για την τιμή του προς ανάλυση προτύπου $\mathbf{x}_i$ σε κάποιο γνώρισμα.
  - Σε κάθε στάδιο της διαδικασίας, εξετάζεται διαφορετικό γνώρισμα.
- Η διαδικασία αυτή εκτελείται **επανειλημμένα**, ώσπου να καταλήξουμε σε τελικό συμπέρασμα για την κλάση του προτύπου/στοιχείου $\mathbf{x}_i$.

Δένδρα απόφασης

- Η διάταξη των ερωτημάτων και οι δυνατές απαντήσεις οργανώνονται σε μία κατάλληλη δομή αναπαράστασης, το *δένδρο απόφασης*.
- Αυτό αποτελείται από κόμβους (*ρίζα*, *εσωτερικούς κόμβους* και *φύλλα*) και κατευθυνόμενες ακμές που τους συνδέουν.
- Κάθε φύλλο αναπαριστά μία κλάση.
- Οι εσωτερικοί κόμβοι και η ρίζα περιέχουν κάποια συνθήκη ελέγχου γνωρίσματος.
  - Η εξέταση των συνθηκών αυτών διαχωρίζει στοιχεία με διαφορετικά χαρακτηριστικά.
  - **Σε κάθε κόμβο εξετάζεται η τιμή διαφορετικού γνωρίσματος.**

Δένδρα απόφασης

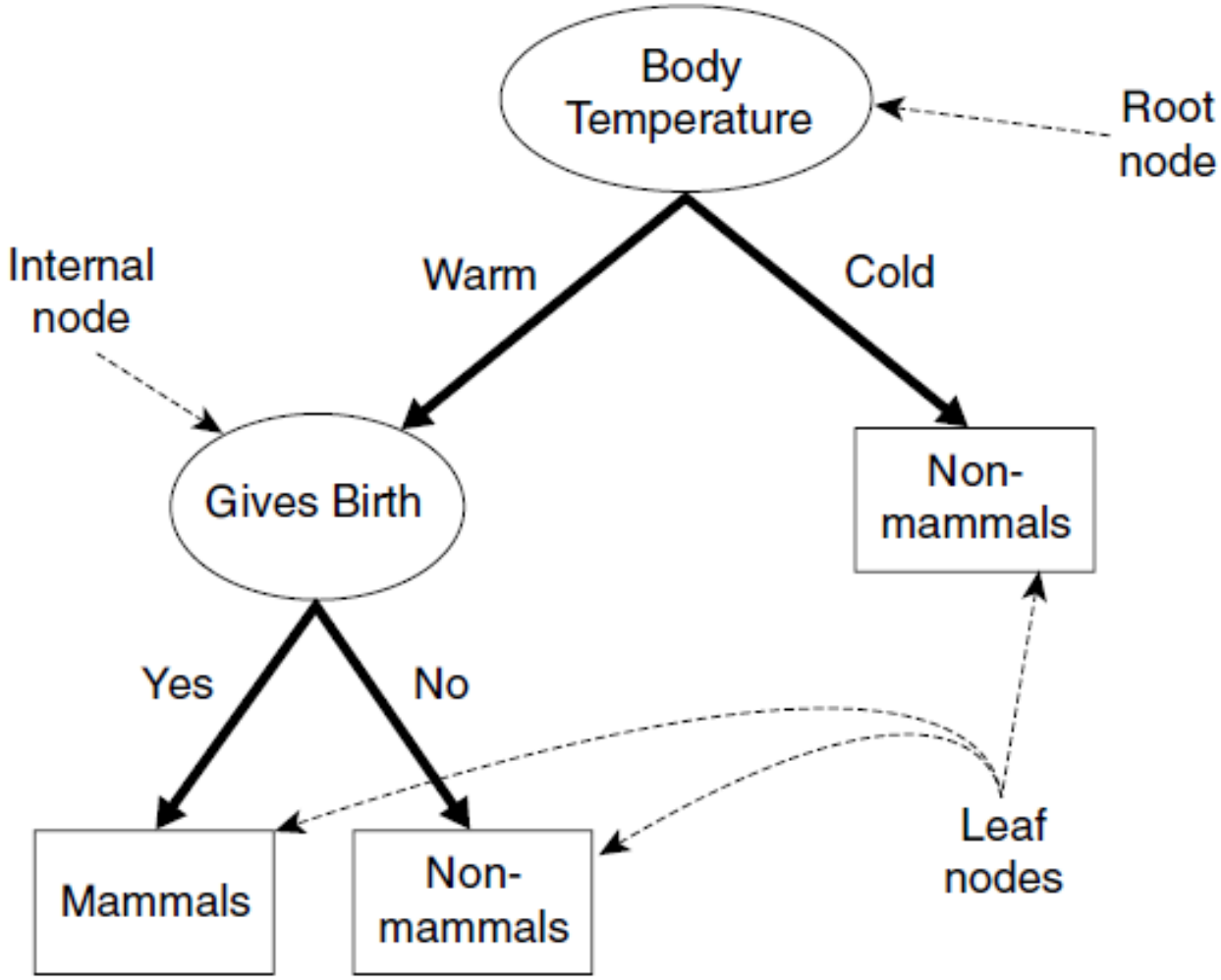
Δένδρα απόφασης

- Η ταξινόμηση ενός νέου προτύπου με δοθέν δένδρο απόφασης εκκινεί από τη ρίζα, εφαρμόζοντας στο στοιχείο τη συνθήκη ελέγχου τιμής του αντίστοιχου γνωρίσματος.
- Στη συνέχεια, ακολουθείται το κλαδί του δένδρου που αντιστοιχεί στο αποτέλεσμα του ελέγχου.
- Η διαδικασία επαναλαμβάνεται διαδοχικά με νέα συνθήκη ελέγχου σε κάθε εσωτερικό κόμβο.
- Τερματίζει όταν καταλήγουμε σε φύλλο.
  - Στο στοιχείο ανατίθεται τελικά η κλάση η οποία συσχετίζεται με το φύλλο τερματισμού.

Δένδρα απόφασης

Δένδρα απόφασης

Δένδρα απόφασης



Πηγή: Tan, “Introduction to Data Mining”, 2006.

Δένδρα απόφασης

Εκπαίδευση δένδρων

- Η χρήση του δένδρου για την πρόβλεψη της κλάσης ενός νέου στοιχείου ελέγχου προϋποθέτει να έχει κατασκευαστεί ήδη το δένδρο απόφασης.

- Η σπουδαιότερη μέθοδος κατασκευής δένδρου απόφασης είναι ο **αλγόριθμος του Hunt**.

- Το δένδρο παράγεται σταδιακά, διαμερίζοντας **αναδρομικά** το σύνολο εκπαίδευσης σε όλο και πιο καθαρά υποσύνολα.

Δένδρα απόφασης

Εκπαίδευση δένδρων

- Η ανάθεση μίας συνθήκης ελέγχου τιμής κάποιου γνωρίσματος σε έναν κόμβο, αμέσως τον «**διασπά**»:
  - Δημιουργούνται οι θυγατρικοί του κόμβοι, ένας για κάθε δυνατό αποτέλεσμα του ελέγχου τιμής.
  - Τα πρότυπα εκπαίδευσης τα οποία είχαν προηγουμένως ανατεθεί στον γονικό κόμβο **διαμερίζονται** στους θυγατρικούς του, με κριτήριο το αποτέλεσμα του ελέγχου τιμής στο επίμαχο γνώρισμα ανά στοιχείο/πρότυπο.

- Ο αλγόριθμος εκκινεί με έναν κόμβο-ρίζα στον οποίον περιέχονται όλα τα στοιχεία.

Εκπαίδευση δένδρων

Δένδρα απόφασης

- **Ψευδοκώδικας**:
- 1: *Αν* όλα τα στοιχεία του τρέχοντος κόμβου ανήκουν στην ίδια κλάση Α, τότε ο κόμβος είναι φύλλο κλάσης Α.
- 2: *Διαφορετικά*:
- 3: Επιλέγεται μία συνθήκη ελέγχου γνωρίσματος η οποία διαμερίζει τα στοιχεία του κόμβου σε μικρότερα υποσύνολα.
- 4: Κατασκευάζεται ένας θυγατρικός κόμβος για κάθε δυνατό αποτέλεσμα της συνθήκης ελέγχου (τιμή γνωρίσματος).
- 5: Τα στοιχεία του αρχικού συνόλου κατανέμονται καταλλήλως στους θυγατρικούς κόμβους.
- 6: Για κάθε θυγατρικό κόμβο, επιστροφή στο Βήμα 1.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Εμφανίζονται κάποιες προβληματικές περιπτώσεις:
  - Ένας υπό κατασκευή θυγατρικός κόμβος ενδέχεται να **μην περιέχει καθόλου στοιχεία**.
    - Πότε; Αν δεν υπάρχουν διανύσματα στο σύνολο εκπαίδευσης με τον αντίστοιχο συνδυασμό τιμών γνωρισμάτων.
    - Λύση: ο κόμβος θεωρείται **φύλλο** και του ανατίθεται η κλάση στην οποία ανήκουν τα περισσότερα στοιχεία του γονικού του.
  - Όλα τα στοιχεία ενός υπό κατασκευή θυγατρικού κόμβου **ταυτίζονται στις τιμές όλων των γνωρισμάτων τους**, μα έχουν διαφορετική ετικέτα κλάσης.
    - Άρα ο κόμβος αυτός δεν μπορεί να διαιρεθεί περαιτέρω.
    - Λύση: ο κόμβος θεωρείται **φύλλο** και του ανατίθεται η κλάση στην οποία ανήκουν τα περισσότερα στοιχεία του ίδιου.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Όμως:
  - Πώς επιλέγεται η συνθήκη ελέγχου γνωρίσματος σε κάθε κόμβο;
  - Ποιο γνώρισμα αφορά κάθε φορά;
  - Ποιοι θυγατρικοί κόμβοι δημιουργούνται;
- Αν το γνώρισμα είναι δυαδικό:
  - Ο ένας θυγατρικός κόμβος περιλαμβάνει τα στοιχεία που έχουν τη μία δυνατή τιμή και ο δεύτερος αυτά που έχουν την άλλη.
- Αν το γνώρισμα έχει πολλαπλές δυνατές τιμές, υπάρχουν διάφορες επιλογές.
  - Η κατάλληλη εξαρτάται από το εκάστοτε πρόβλημα.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Η απλούστερη λύση είναι να δημιουργηθούν τόσοι κόμβοι όσες οι δυνατές τιμές.
  - Τα συνεχή γνωρίσματα υφίστανται πρώτα διακριτοποίηση.
  - Όμως κάθε κόμβος μπορεί να συγκεντρώνει έτσι πολύ λίγα στοιχεία, καθιστώντας αναξιόπιστη την ταξινόμηση.
- Εναλλακτικές λύσεις:
  - Κατηγορικά γνωρίσματα: κάθε θυγατρικός κόμβος περιλαμβάνει **στοιχεία με πολλαπλές διαφορετικές τιμές** στο επίμαχο γνώρισμα.
    - Αρκεί να μην υφίσταται επικάλυψη μεταξύ των κόμβων.
  - Αριθμητικά γνωρίσματα: τα στοιχεία διασπώνται με μία συγκριτική συνθήκη ελέγχου η οποία δίνει δύο μόνο θυγατρικούς κόμβους.
    - «*Το γνώρισμα Α έχει τιμή μεγαλύτερη ή μικρότερη του κατωφλίου t;*»
    - Η επιλογή του κατάλληλου κατωφλίου γίνεται με αναζήτηση σε όλα τα δυνατά *t* και αξιολόγησή τους με κάποιο κριτήριο.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Πώς επιλέγεται όμως το γνώρισμα στο οποίο θα γίνει η διάσπαση;
  - Αυτό το οποίο θα αφορά η συνθήκη ελέγχου στον γονικό κόμβο;
- Με αναζήτηση στο χώρο των δυνατών διασπάσεων και αξιολόγησή τους με βάση ένα κριτήριο **καθαρότητας των θυγατρικών κόμβων** οι οποίοι προκύπτουν από την εν λόγω διάσπαση.
  - Ένα μέτρο του κατά πόσον θα περιέχουν στοιχεία διαφορετικών κλάσεων.
- Στόχος είναι η επιλογή της διάσπασης η οποία **μεγιστοποιεί την καθαρότητα** των προκυπτόντων θυγατρικών κόμβων.
  - Έτσι είναι περισσότερο πιθανή η σύντομη κατάληξη σε φύλλο.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Σύνηθες μέτρο καθαρότητας ενός κόμβου είναι η **εντροπία**.
  - Υπολογίζεται ως εντροπία μίας διακριτής κατανομής με δειγματικό χώρο τις δυνατές κλάσεις.
  - Μεγαλύτερη εντροπία συνεπάγεται μικρότερη καθαρότητα.
- Επιλέγεται η διάσπαση η οποία μεγιστοποιεί το **πληροφοριακό κέρδος Δ**.
  - Τη διαφορά της καθαρότητας του γονικού κόμβου από τον σταθμισμένο μέσο της εντροπίας των προκυπτόντων θυγατρικών κόμβων.
- Όμως συνήθως θέλουμε μικρό πλήθος θυγατρικών κόμβων. Η λύση είναι ο **λόγος κέρδους**.
  - Κανονικοποιούμε το προκαταρκτικό κέρδος διαιρώντας το με το πλήθος των προκυπτόντων θυγατρικών κόμβων.
  - Μεγάλο πλήθος κόμβων σημαίνει μικρός λόγος κέρδους.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- Το πληροφοριακό κέρδος μπορεί να αξιοποιηθεί και για την αξιολόγηση των δυνατών τιμών κατωφλίου t.
- Τα δένδρα απόφασης κινδυνεύουν από **υπερεκπαίδευση**.
  - Μεγάλο πλήθος κόμβων και κλαδιών, άρα πολύπλοκο μοντέλο.
- Μία λύση είναι το **κλάδεμα του δένδρου μετά την κατασκευή του**, με στόχο την απλοποίησή του.
  - Κάποια υποδένδρα αντικαθίστανται από φύλλα.
  - Η κλάση των εν λόγω φύλλων είναι η κλάση της πλειονότητας των προτύπων που εμπεριέχονταν στο υποδένδρο που αντικαθιστούν.
- Έτσι, εν τέλει, οι δυνατές διαδρομές από τη ρίζα έως τα φύλλα του δένδρου δεν καλύπτουν επακριβώς όλα τα στοιχεία του συνόλου εκπαίδευσης.
  - Αυτό θα οδηγούσε σε υψηλή διακύμανση και υπερεκπαίδευση.

Ζητήματα στην εκπαίδευση δένδρων

- Μία εναλλακτική λύση είναι το **πρόωρο σταμάτημα** της κατασκευής του αρχικού δένδρου.
  - Μπορεί να σταματήσει πρόωρα η διάσπαση εσωτερικών κόμβων.
  - Έτσι, ένας κόμβος σηματοδοτείται ως φύλλο χωρίς να περιέχει στοιχεία μόνο μίας κλάσης.
    - Αρκεί η εντροπία του να είναι μικρότερη από ένα κατώφλι-υπερπαράμετρο.
- Με το πρόωρο σταμάτημα απαιτείται μία εκτίμηση του **σφάλματος επικύρωσης** σε κάθε βήμα εκπαίδευσης, δηλαδή κατασκευής του δένδρου.
  - Απαιτείται, ώστε να ξέρουμε πότε να σταματήσουμε πρόωρα.
    - Όταν αρχίζει να αυξάνεται το σφάλμα επικύρωσης.
  - Ίσως να απαιτείται και κατά το κλάδεμα, αφού έχει τερματιστεί η κατασκευή του δένδρου, ώστε να αξιολογείται κάθε υποψήφια απόφαση κλαδέματος.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- **Πλεονεκτήματα** των δένδρων απόφασης:
  - Παράγουν προβλέψεις σε μορφή κανόνων και, άρα, ερμηνεύσιμα από τον χρήστη αποτελέσματα.
    - Ο κανόνας προκύπτει από το μονοπάτι το οποίο ακολουθείται από τη ρίζα έως το τελικό φύλλο, για δοθέν πρότυπο ελέγχου.
    - Οι επιλογές οι οποίες γίνονται κατά μήκος του μπορούν να αξιολογηθούν από τον χρήστη ως εύλογες ή μη.
  - Υλοποιούνται εύκολα.
  - Είναι αξιόπιστα παρουσία θορύβου.
  - Δεν υποθέτουν παραδοχές για τα δεδομένα και τη δομή τους.
  - Δεν επηρεάζονται αρνητικά από πλεονάζοντα γνωρίσματα.
    - Απλώς αυτά δεν χρησιμοποιούνται σε συνθήκες ελέγχου.
  - Η συναγωγή μίας πρόβλεψης ταξινόμησης για δοθέν πρότυπο ελέγχου είναι πολύ γρήγορη.
    - Υπολογιστική πολυπλοκότητα γραμμική ως προς το βάθος του δένδρου.

Δένδρα απόφασης

Ζητήματα στην εκπαίδευση δένδρων

- **Προβλήματα** των δένδρων απόφασης:
  - Η εύρεση του βέλτιστου δένδρου απόφασης για δεδομένο σύνολο εκπαίδευσης συνήθως είναι υπολογιστικά αδύνατη.
    - NP-πλήρες πρόβλημα.
    - Επιστρατεύονται άπληστοι αλγόριθμοι κατασκευής οι οποίοι δίνουν υποβέλτιστα δένδρα.
  - Ίσως κάποιο υποδένδρο να επαναλαμβάνεται σε διαφορετικά κλαδιά, καθιστώντας το δένδρο περισσότερο πολύπλοκο από όσο είναι απαραίτητο και επιρρεπές στην υπερεκπαίδευση.
    - Αποτέλεσμα του γεγονότος ότι σε κάθε εσωτερικό κόμβο αντιστοιχίζεται μία συνθήκη ελέγχου ενός μόνο γνωρίσματος η οποία δεν λαμβάνει υπ' όψιν της τα υπόλοιπα γνωρίσματα.

Ζητήματα στην εκπαίδευση δένδρων

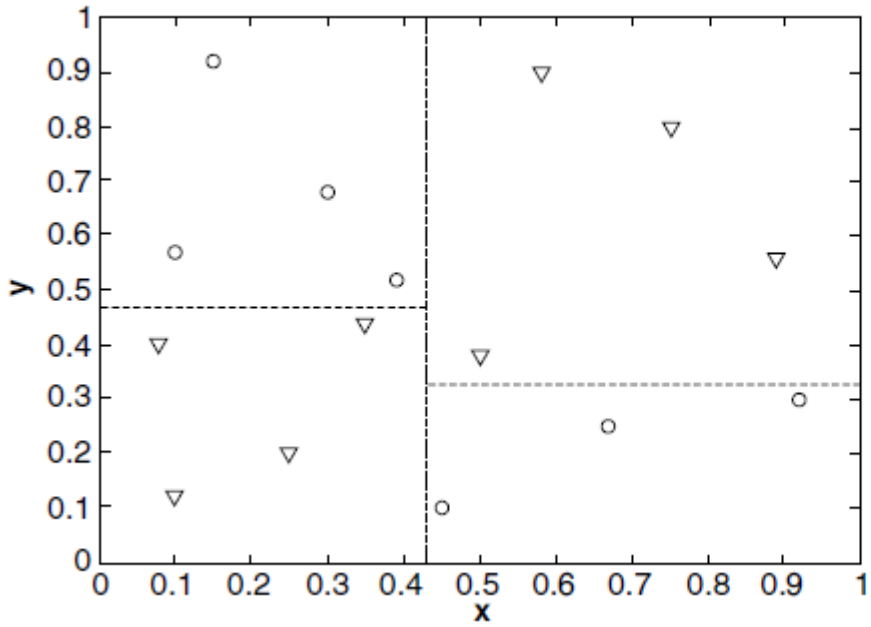
Δένδρα απόφασης

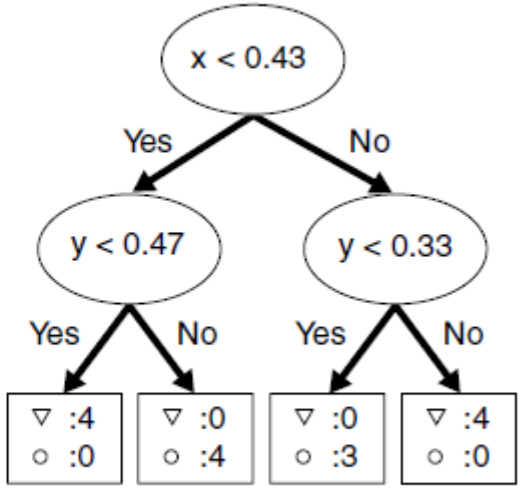
- **Προβλήματα** των δένδρων απόφασης:
  - Τα δένδρα απόφασης αδυνατούν να ανακαλύψουν κατάλληλη διάσπαση σε συνεχή γνωρίσματα, όταν τα όρια μεταξύ γειτονικών περιοχών με στοιχεία διαφορετικών κλάσεων δεν είναι παράλληλα με τους άξονες του χώρου των προτύπων.
    - Τα δένδρα απόφασης, δηλαδή, ορίζουν μόνο ορθογώνιες περιοχές απόφασης στον χώρο των προτύπων.
- Το ζήτημα αυτό λύνεται από τα **κεκλιμένα δένδρα απόφασης** (oblique decision trees).
  - Μία συνθήκη ελέγχου σε έναν εσωτερικό κόμβο μπορεί να αφορά περισσότερα από ένα γνωρίσματα.
  - Έχουν αυξημένη ευελιξία στον ορισμό περιοχών απόφασης.
- Εναλλακτική λύση είναι η κατάλληλη προεπεξεργασία των γνωρισμάτων, ώστε να είναι συμβατά με ορθογώνιες περιοχές απόφασης.

Δένδρα απόφασης

Δένδρα απόφασης

Παράδειγμα σχηματισμού περιοχών απόφασης σε διδιάστατα δεδομένα.



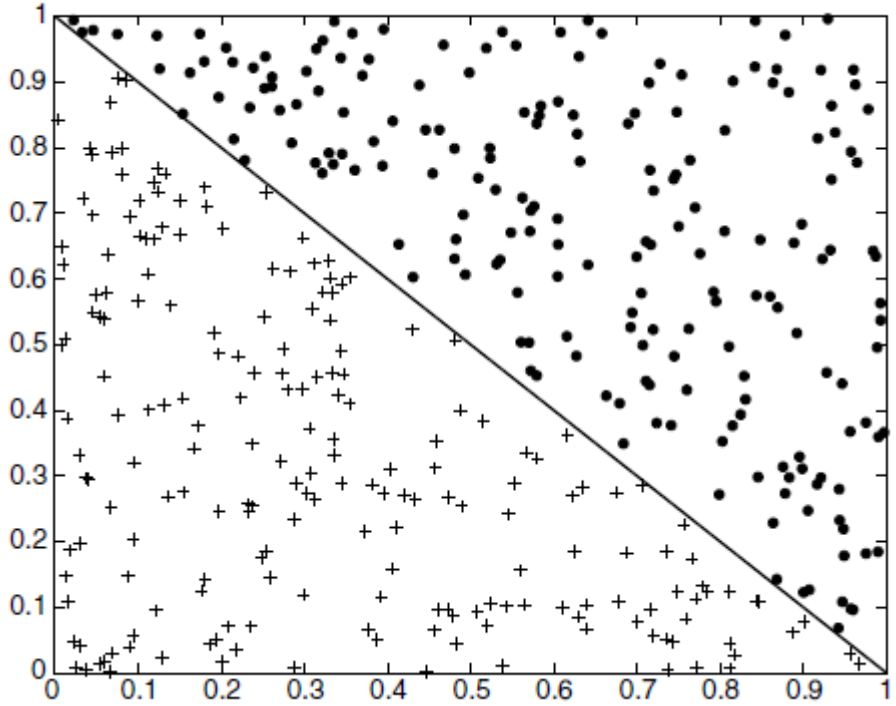


Πηγή: Tan, “Introduction to Data Mining”, 2006.

Δένδρα απόφασης

Δένδρα απόφασης

Παράδειγμα γραμμικά διαχωρίσιμων διδιάστατων δεδομένων τα οποία μπορούν να ταξινομηθούν ορθά μόνο από κεκλιμένο δένδρο απόφασης.



Πηγή: Tan, “Introduction to Data Mining”, 2006.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr