

Εισαγωγή στο Data Mining
Από τα δεδομένα στη γνώση

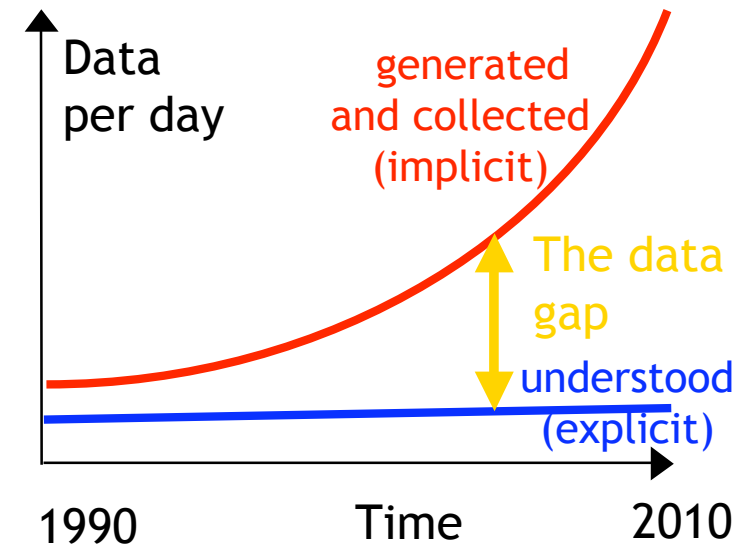
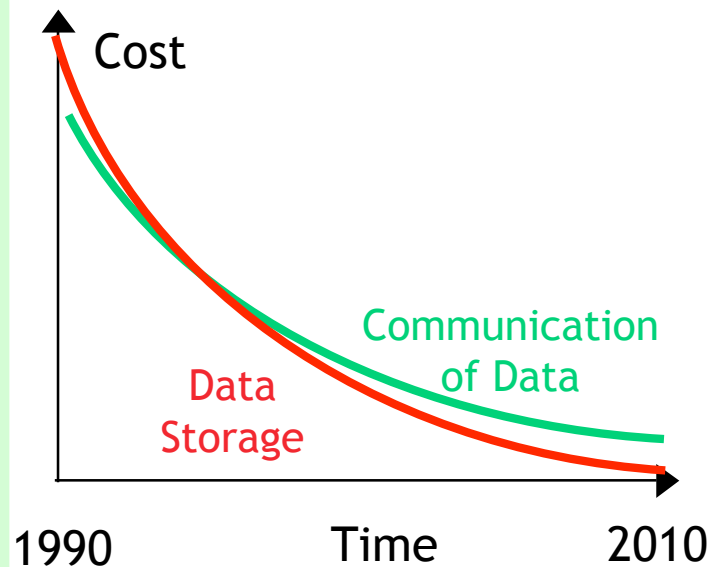
Η πληροφορία στη σύγχρονη επιχείρηση

- Εισαγωγή
 - Η Ανάγκη
 - Διαδικασία
 - Ορισμός
 - Αφετηρία
 - Πρότυπα
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Πέραν του ανθρώπινου δυναμικού, η πληροφορία αποτελεί τον πλέον πολύτιμο πόρο της σύγχρονης επιχείρησης
 - Συνιστά επομένως καθοριστικό παράγοντα για την επίτευξη των επιχειρησιακών στόχων
- Διαχείριση γνώσης (knowledge management): η εκμετάλλευση και ανάπτυξη των γνωστικών κεφαλαίων μίας επιχείρησης με αντικείμενο την προαγωγή των αντικειμενικών της επιδιώξεων
- Τα πληροφοριακά συστήματα ERP ανταποκρίνονται στην απαίτηση συλλογής και επεξεργασίας των επιχειρησιακών δεδομένων
 - Data warehouses

Τα επιχειρησιακά δεδομένα αφθονούν

- Εισαγωγή
 - Η Ανάγκη
 - Διαδικασία
 - Ορισμός
 - Αφετηρία
 - Πρότυπα
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



Malthus Law of Information:

- Το νέο πληροφοριακό περιεχόμενο διπλασιάζεται κάθε χρόνο
- Ο χρόνος που δαπανάται για την κατανάλωση πληροφοριών παραμένει σταθερός

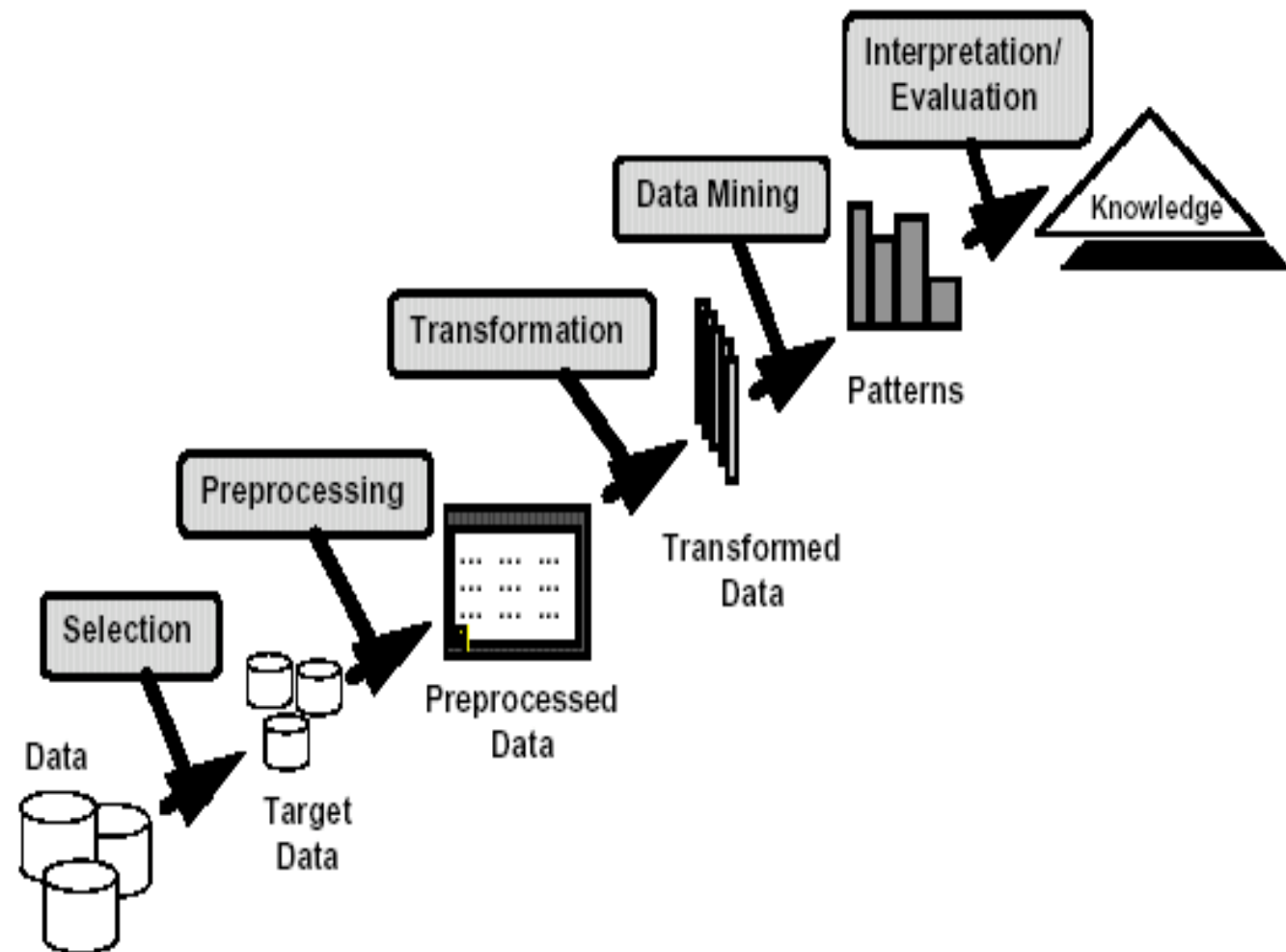
Τα δεδομένα αφθονούν

- Εισαγωγή
 - Η Ανάγκη
 - Διαδικασία
 - Ορισμός
 - Αφετηρία
 - Πρότυπα
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Μόνο ένα μικρό ποσοστό (5-10%) των συλλεγόμενων δεδομένων τυγχάνει ανάλυσης
- Μία τυπική επιχειρησιακή βάση δεδομένων σήμερα περιέχει συχνά μεγάλο αριθμό εγγραφών (10^8 - 10^{12}) δεδομένων πολλών διαστάσεων (10 - 10^4 μεταβλητές)
- Τελικά: ***“We are drowning in data, but starving for knowledge!”***
- Πώς μπορούν να εξερευνηθούν εκατομμύρια εγγραφών εκατοντάδων μεταβλητών, ώστε να ανακαλυφθούν πρότυπα (patterns)?

Από τα δεδομένα στην πληροφορία και τη γνώση

- Εισαγωγή
 - Η Ανάγκη
- Διαδικασία
 - Ορισμός
 - Αφετηρία
 - Πρότυπα
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



Ορισμός

- Εισαγωγή

- Η Ανάγκη

- Διαδικασία

- Ορισμός

- Αφετηρία

- Πρότυπα

- Case Studies

- Μορφή
Δεδομένων

- Μορφή
Εξαγομένων

- Αλγόριθμοι

- Λογισμικό

- Επίδειξη

- **Data mining:** όρος που χρησιμοποιείται –λανθασμένα– για να περιγράψει το σύνολο της διαδικασίας εξόρυξης γνώσης από βάσεις δεδομένων (*Knowledge Discovery in Databases*)
- Ορισμός: *The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.*
- Εναλλακτικά: *Statistics at scale, speed and simplicity.*
- Στο εξής, ορίζουμε: *Εξόρυξη Πληροφορίας / Γνώσης από Δεδομένα* $\equiv$ *Data Mining*

Αφετηρία

- Εισαγωγή

- Η Ανάγκη
- Διαδικασία
- Ορισμός
- Αφετηρία
- Πρότυπα

- Case Studies

- Μορφή
Δεδομένων

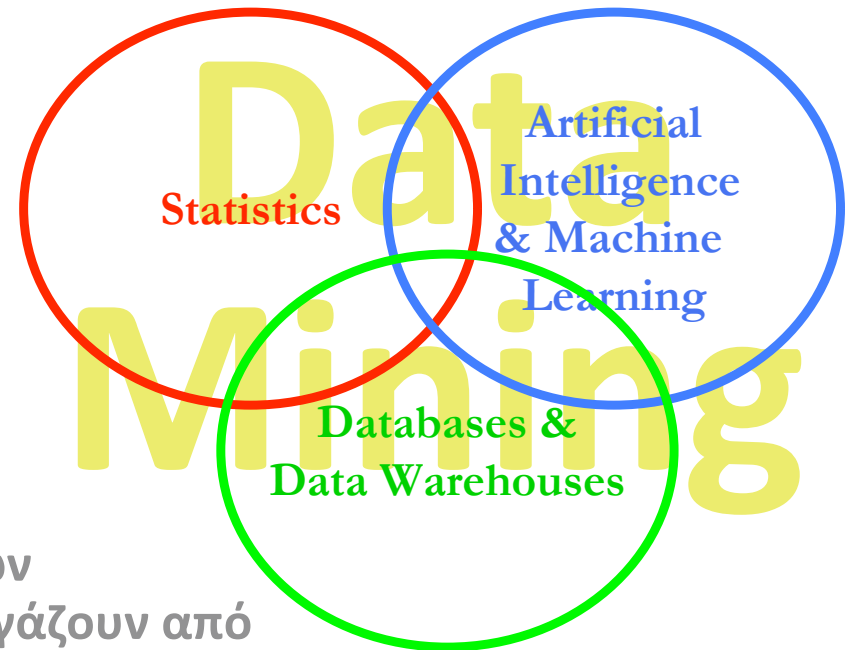
- Μορφή
Εξαγομένων

- Αλγόριθμοι

- Λογισμικό

- Επίδειξη

- Το ερευνητικό πεδίο αποτελεί τομή μεθόδων και εργαλείων που πηγάζουν από
 - Στατιστική
 - Μηχανική Μάθηση
 - Βάσεις & αποθήκες δεδομένων
- Αποτελεί σύγχρονη εξέλιξη, το πρώτο σχετικό συνέδριο πραγματοποιήθηκε το 1995.
- Πειραματική επιστήμη!



Πρότυπα

- Εισαγωγή

- Η Ανάγκη
- Διαδικασία
- Ορισμός
- Αφετηρία

- **Πρότυπα**

- Case Studies

- Μορφή
Δεδομένων

- Μορφή
Εξαγομένων

- Αλγόριθμοι

- Λογισμικό

- Επίδειξη

- Επομένως αναζητούνται *αλγόριθμοι εντοπισμού προτύπων (patterns) και κανονικοτήτων σε δεδομένα*
- Περιγραφή προτύπων
 - Μαύρο κουτί: μη κατανοητοί μηχανισμοί
 - Διαφανές κουτί: αποκαλύπτει τη δομή του προτύπου
→ δομική περιγραφή
- Οι δομικές (structural) περιγραφές αναπαριστούν τα πρότυπα με σαφώς ορισμένο (ρητό, explicit) τρόπο, με σκοπό την
 - Πρόβλεψη
 - Κατανόηση και επεξήγηση μοντέλου πρόβλεψης

Case study: VISA

- Εισαγωγή
- Case Studies
 - VISA
 - American Express
 - Softmap
 - Walmart
 - Victoria Secret
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Απάτη κατά τη χρήση πιστωτικών καρτών
 - 0,93% σε off-line συναλλαγές
 - 1,97% σε on-line
- Κλασσικές στατιστικές μέθοδοι: 97 λανθασμένοι συναγερμοί για κάθε αληθή
- Ανάλυση δεδομένων συμπεριφοράς με τεχνικές μηχανικής μάθησης
 - μείωση της αναλογίας λανθασμένων συναγερμών στους 10 ανά αληθή
 - ελάττωση περιστατικών απάτης από 1576 σε 478, σε δείγμα 7 ευρωπαϊκών τραπεζών

Case study: American Express

- Εισαγωγή
- Case Studies
 - VISA
 - American Express
 - Softmap
 - Walmart
 - Victoria Secret
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Πρόβλημα: Παροχή δανείου
 - Ερωτηματολόγιο για σχετικές οικονομικές και προσωπικές πληροφορίες → Αποδοχή / απόρριψη αίτησης δανειοδότησης
 - Συνήθεις στατιστικές τεχνικές καλύπτουν αποτελεσματικά το 90% των περιπτώσεων
 - Οι υπόλοιπες ασαφείς περιπτώσεις αξιολογούνται από ειδικούς
 - Ωστόσο το 50% των αποδεκτών ασαφών περιπτώσεων δεν αποπληρώνουν
 - 1000 παραδείγματα εκπαίδευσης ασαφών περιπτώσεων
 - 20 χαρακτηριστικά (ηλικία, οικογενειακή κατάσταση, διάρκεια εργασίας υπό τον ίδιο εργοδότη, άλλα δάνεια ...)
- Οι κανόνες που προέκυψαν αποδείχθηκαν σωστοί στο 70% των περιπτώσεων

Case study: Softmap

- Εισαγωγή
- Case Studies
 - VISA
 - American Express
 - Softmap
 - Walmart
 - Victoria Secret
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Κορυφαία εταιρεία πωλήσεων PC & software στην Ιαπωνία
- Οι on-line πωλήσεις εμφάνισαν συνεχή μείωση
 - Αδυναμία προσαρμογής στις ειδικές ανάγκες του πελάτη;
- Ανάλυση χαρακτηριστικών 2 εκ. παλαιών πελατών & ανάπτυξη διαδικτυακού συστήματος υποβολής πρότασης μετά από ερωτηματολόγιο
 - αύξηση επισκεψιμότητας ιστοσελίδας 67%
 - τριπλασιασμός εσόδων ιστοσελίδας το έτος εισαγωγής του συστήματος

Case study: Walmart

- Εισαγωγή
- Case Studies
 - VISA
 - American Express
 - Softmap
 - Walmart
 - Victoria Secret
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Τεράστια βάση δεδομένων αγοραστικής συμπεριφοράς
 - Η ανάλυση με μεθόδους μηχανικής μάθησης ανέδειξε συμπεράσματα της μορφής
 - Οι πελάτες που αγοράζουν μύρα συχνά προμηθεύονται και πίτσα
 - Κάθε Πέμπτη, οι πελάτες που αγοράζουν μύρα αγοράζουν επίσης πάνες
- Τέτοιες πληροφορίες έχουν δυνητικά μεγάλη προστιθέμενη αξία
 - Αναδιάταξη προϊόντων,
 - Διαχείριση αποθεμάτων...
- Συνήθης εφαρμογή: Τα δεδομένα που αποκτώνται μέσω προσωπικών εκπτώτικών καρτών είναι συχνά μεγαλύτερης αξίας από την παρεχόμενη έκπτωση

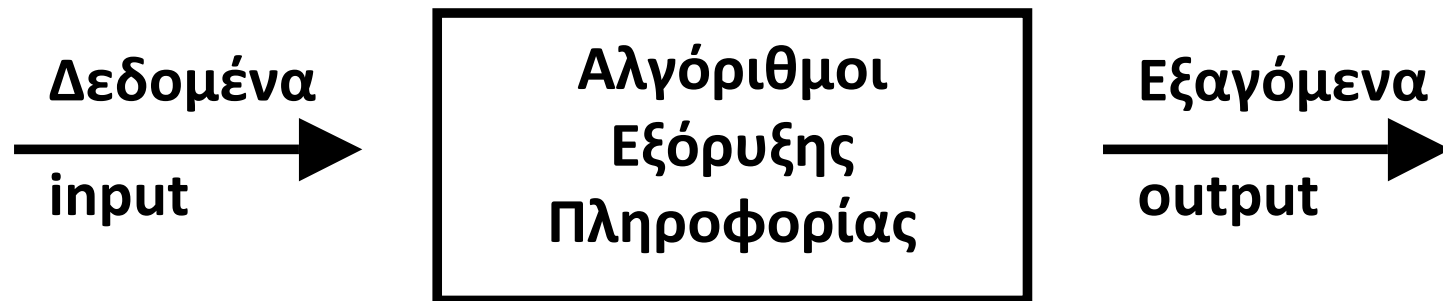
Case study: Victoria Secret

- Εισαγωγή
- Case Studies
 - VISA
 - American Express
 - Softmap
 - Walmart
 - Victoria Secret
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- **Direct Marketing:** Οι προσφορές προώθησης προϊόντος είναι συχνά κοστοβόρες και έχουν ένα πολύ χαμηλό –αλλά ιδιαίτερα προσοδοφόρο – ποσοστό απόκρισης
- **Victoria Secret:** φθίνουσα πορεία εσόδων από ταχυδρομική αποστολή καταλόγου
- Μετά από την **ανάλυση δημογραφικών δεδομένων** των πελατών της, η στόχευση εστιάστηκε σε διαφοροποιημένο υποσύνολο πελατών και προϊόντων
 - Η πορεία των εσόδων αντιστράφηκε, ενώ παράλληλα το κόστος αποστολής μειώθηκε κατά 15%

Η μορφή των δεδομένων και η σημασία της

- Εισαγωγή
- Case Studies
- **Μορφή
Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- Μορφή
Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



- Η κατανόηση της μορφής των δεδομένων και εξαγομένων είναι ίσως *περισσότερο σημαντική* από τα ενδιάμεσα βήματα σε έναν αλγόριθμο εξόρυξης πληροφορίας

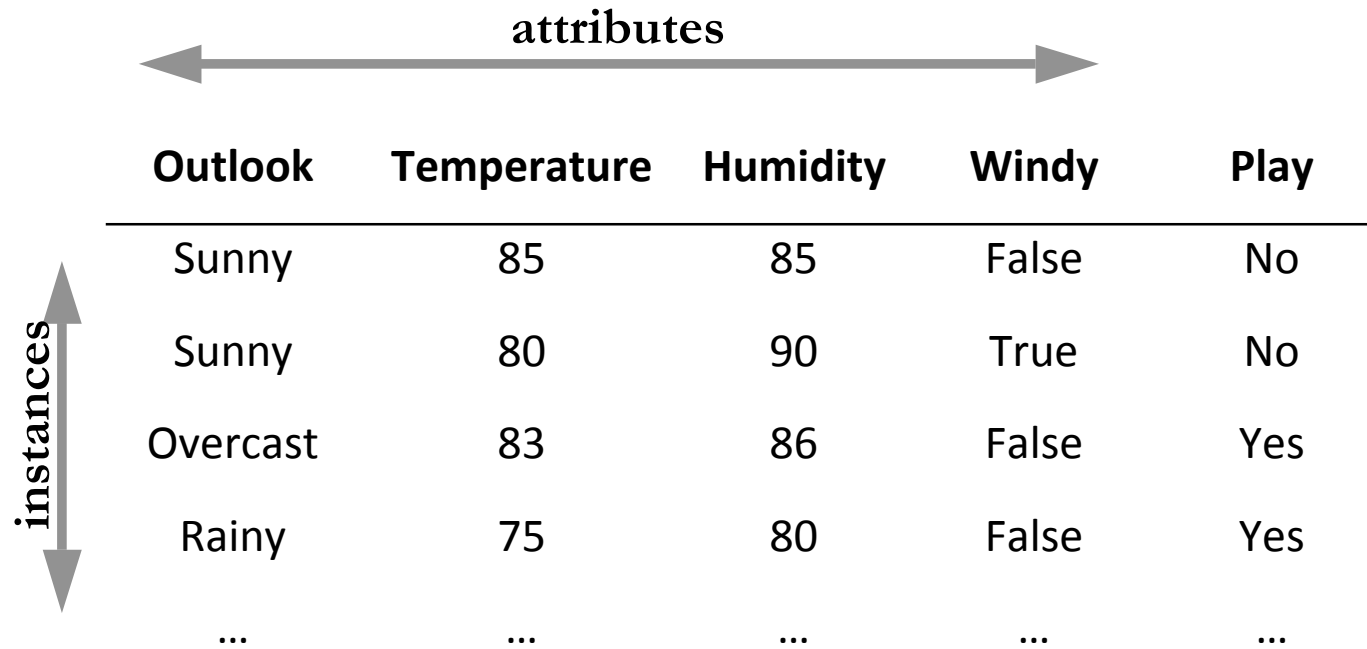
Συνιστώσες Δεδομένων, Ορολογία

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - **Ορολογία**
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- **Αντίληψη (concept):** *το αντικείμενο της μάθησης*
 - Στόχος: εύρεση κατανοητής και λειτουργικής περιγραφής ενός concept
- **Υπόδειγμα (instance):** *το ξεχωριστό και ανεξάρτητο παράδειγμα (example) ενός concept*
 - Σημείωση: Περισσότερο σύνθετα σχήματα εισόδου είναι πιθανά
- **Χαρακτηριστικό (attribute):** *η μετρήσιμη συνιστώσα ενός υποδείγματος*
 - Θα εστιάσουμε σε ονομαστικές και αριθμητικές συνιστώσες

Συνιστώσες Δεδομένων, Παράδειγμα

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



attributes				
Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

concept description:

**If outlook = sunny and humidity = high
then play = no**

Concept (Αντίληψη)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- **Είδη μάθησης:**
 - Ταξινόμηση (classification):
πρόβλεψη διακριτής κατηγορίας
 - Συσχέτιση (association):
εντοπισμός συσχετίσεων μεταξύ χαρακτηριστικών
 - Ομαδοποίηση (clustering):
ανάδειξη ομάδων όμοιων υποδειγμάτων
 - Αριθμητική πρόβλεψη (numeric prediction):
πρόβλεψη αριθμητικής ποσότητας
- **Concept (αντίληψη):** *το αντικείμενο της μάθησης*
- **Περιγραφή αντίληψης (concept description):**
το προϊόν / σχήμα της μαθησιακής διαδικασίας

Ταξινόμηση (Classification)

- Εισαγωγή
- Case Studies
- **Μορφή
Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή
Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Ταξινόμηση υποδείγματος σε προκαθορισμένη *τάξη (class)*
 - Για παράδειγμα: play/no-play
- Μάθηση **με επίβλεψη (supervised)**
 - Οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές
 - Το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό
- Ο **βαθμός αξιοπιστίας** μετριέται
 - σε μη χρησιμοποιημένα για τη διαμόρφωση του concept description δεδομένα (*test data*) είτε
 - υποκειμενικά, ανάλογα με το βαθμό αποδοχής της περιγραφής

Συσχέτιση (Association)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Ανακάλυψη συσχετίσεων μεταξύ των διάφορων χαρακτηριστικών
- Διαφορές με τη μάθηση ταξινόμησης:
 - Μπορεί να αναδείξουν τη συσχέτιση είτε να προβλέψουν την τιμή **οποιουδήποτε χαρακτηριστικού** και όχι μόνο της τάξης
 - Συνδέουν πιθανόν **περισσότερα από ένα χαρακτηριστικά** κάθε φορά
 - Επομένως, προκύπτουν **πολλοί περισσότεροι κανόνες** συσχέτισης από ότι κανόνες ταξινόμησης
 - Άρα περιορισμοί κατά την αναζήτηση είναι αναγκαίοι, όπως ελάχιστη κάλυψη & ελάχιστη ακρίβεια

Ομαδοποίηση (Clustering)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Εύρεση ομάδων αντικειμένων με υψηλό βαθμό ομοιότητας και εκχώρηση υποδειγμάτων στις ομάδες αυτές
- *Χωρίς επίβλεψη (unsupervised)*
 - η τάξη του υποδείγματος δεν ανήκει σε γνωστό σύνολο

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Αριθμητική πρόβλεψη (Numeric prediction)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Όμοια με ταξινόμηση, αλλά τώρα η 'τάξη' είναι αριθμητική
- Μάθηση με επίβλεψη (*supervised*)
 - Η τιμή – στόχος κάθε υποδείγματος είναι εκ των προτέρων γνωστή

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
...	...	...	...	...

Παράδειγμα (example)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - **Παράδειγμα**
 - Χαρακτηριστικό
 - Προπαρασκευή
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Υπόδειγμα (instance): ειδικός τύπος παραδείγματος
 - Προς ταξινόμηση, συσχέτιση ή ομαδοποίηση
 - *Ιδιαίτερο & ανεξάρτητο παράδειγμα της αντίληψης προς εκμάθηση*
 - *Χαρακτηρίζεται από προκαθορισμένο σύνολο χαρακτηριστικών*
- Συνήθης μορφή δεδομένων προς εξόρυξη γνώσης: σύνολο υποδειγμάτων (dataset)
 - *Πίνακας υποδειγμάτων (instances) – χαρακτηριστικών (attributes) (επίπεδο αρχείο, flat file)*
 - Μάλλον περιοριστική μορφή δεδομένων (καμία συσχέτιση μεταξύ των αντικειμένων)

Χαρακτηριστικό (attribute)

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - **Χαρακτηριστικό**
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Κάθε παράδειγμα αποτελείται από προκαθορισμένο σύνολο στοιχείων, τα οποία καλούνται χαρακτηριστικά (attributes)
- Στην πράξη, ο αριθμός των χαρακτηριστικών μπορεί να ποικίλει
- Επίσης, η ύπαρξη ενός χαρακτηριστικού μπορεί να εξαρτάται από την τιμή ενός άλλου
- **Πιθανοί τύποι χαρακτηριστικών (“levels of measurement”):**
 - Ονομαστικά (*nominal*),
 - τακτικά (*ordinal*),
 - περιοδικά (*interval*) και
 - αναλογικά (*ratio*)

Προπαρασκευή δεδομένων εισόδου

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- **Πρόβλημα01: διαφορετικές πηγές δεδομένων** (για παράδειγμα, σε μία επιχείρηση, εγγραφές τμήματος πωλήσεων, λογιστηρίου, ...)
 - Διαφορές: τρόπος αποθήκευσης εγγραφών, παραδοχές, χρονικές περίοδοι, άθροιση δεδομένων, σφάλματα
 - Τα δεδομένα πρέπει να συγκεντρωθούν σε ενιαίο σύνολο, με ενιαία λιτή δομή → **data warehouse: συνεπές σημείο πρόσβασης**
 - Κρίσιμος παράγοντας: τύπος και επίπεδο άθροισης δεδομένων
 - Για παράδειγμα δεδομένα ανά συναλλαγή, ανά πελάτη, ανά ημέρα κτλ.
- **Πρόβλημα02: ανακριβείς τιμές**
- **Πρόβλημα03: άγνωστες τιμές**

Εξερεύνηση των δεδομένων

- Εισαγωγή
- Case Studies
- **Μορφή
Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή
Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Ορισμένα απλά εργαλεία οπτικοποίησης είναι πολύ χρήσιμα
 - Ονομαστικά χαρακτηριστικά: **ιστογράμματα** (είναι η κατανομή συμβατή με τη γνώση πεδίου;)
 - Αριθμητικά χαρακτηριστικά: **γραφήματα** (υπάρχουν εμφανείς τιμές προς εξαίρεση (outliers);)
- Διαγράμματα 2 & 3 διαστάσεων υποδεικνύουν εξαρτήσεις & αλληλοσυσχετίσεις
- Αναγκαία η γνώση των ειδικών
- Αχανής όγκος δεδομένων; **Δειγματοληψία!**

Προεπεξεργασία δεδομένων

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Ο μετασχηματισμός των δεδομένων σε μορφή κατάλληλη και αποδοτική για την επιλεγμένη μέθοδο μάθησης
 - *Σύνολο τεχνασμάτων που εφαρμόζονται με κύριο στόχο την αύξηση του βαθμού αξιοπιστίας*
- Κάποιες φορές λειτουργούν, κάποιες όχι, και, μέχρι σήμερα, είναι δύσκολο να γνωρίζει κανείς εκ των προτέρων την αποτελεσματικότητά τους
- Καθώς ο πλέον αξιόπιστος οδηγός είναι η μέθοδος ‘δοκιμής & σφάλματος’ (‘trial & error’), η γνώση και κατανόηση των τεχνασμάτων αυτών είναι εξαιρετικά σημαντική

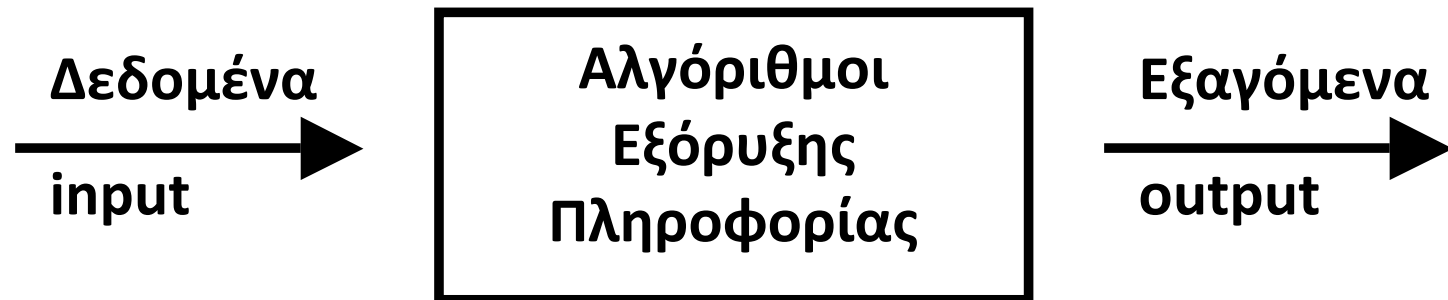
Τεχνοτροπίες προεπεξεργασίας δεδομένων

- Εισαγωγή
- Case Studies
- **Μορφή Δεδομένων**
 - Ορολογία
 - Αντίληψη & είδη της
 - Παράδειγμα
 - Χαρακτηριστικό
 - Προπαρασκευή
- **Μορφή Εξαγομένων**
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- **Προεπιλογή χαρακτηριστικών (attributes)**
 - Κατά κανόνα, τα περισσότερα χαρακτηριστικά – συχνά η συντριπτική πλειοψηφία τους – είναι ευκρινώς μη συσχετιζόμενα ή περιττά → αποκοπή
- **Διακριτοποίηση χαρακτηριστικών**
 - Αναγκαία όταν κάποια χαρακτηριστικά είναι αριθμητικά αλλά ο επιλεγμένος αλγόριθμος μπορεί να χειριστεί μόνο ρητά (categorical) χαρακτηριστικά
- **Μετασχηματισμός δεδομένων**
 - Μαθηματικοί μετασχηματισμοί, βασισμένοι στη γνώση πεδίου, λογικοί μετασχηματισμοί, αλλαγή δομής / μορφής δεδομένων
- **Καθαρισμός δεδομένων**
 - Οπτικοποίηση και αυτοματοποιημένες μέθοδοι

Η μορφή των εξαγόμενων και η σημασία της

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- **Μορφή Εξαγόμενων**
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



- Οι τεχνικές Μηχανικής Μάθησης παρέχουν διάφορες δομικές περιγραφές των εξαγόμενων
- Καθεμία από αυτές υπαγορεύει το είδος του αλγορίθμου που πρέπει να χρησιμοποιηθεί για το σχηματισμό της από τα δεδομένα

Εξαγόμενα: Απεικόνιση γνώσης

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Πίνακες απόφασης (*decision tables*)
- Δένδρα απόφασης (*decision trees*)
- Κανόνες απόφασης (*decision rules*)
- Κανόνες συσχέτισης (*association rules*)
- Κανόνες με εξαιρέσεις (*rules with exceptions*)
- Κανόνες με συσχετίσεις (*rules involving relations*)
- Γραμμική παλινδρόμηση (*linear regression*)
- Δένδρα για αριθμητική πρόβλεψη (*trees for numeric prediction*)
- Απεικόνιση με βάση υποδείγματα (*instance-based representation*)
- Ομάδες (*clusters*)

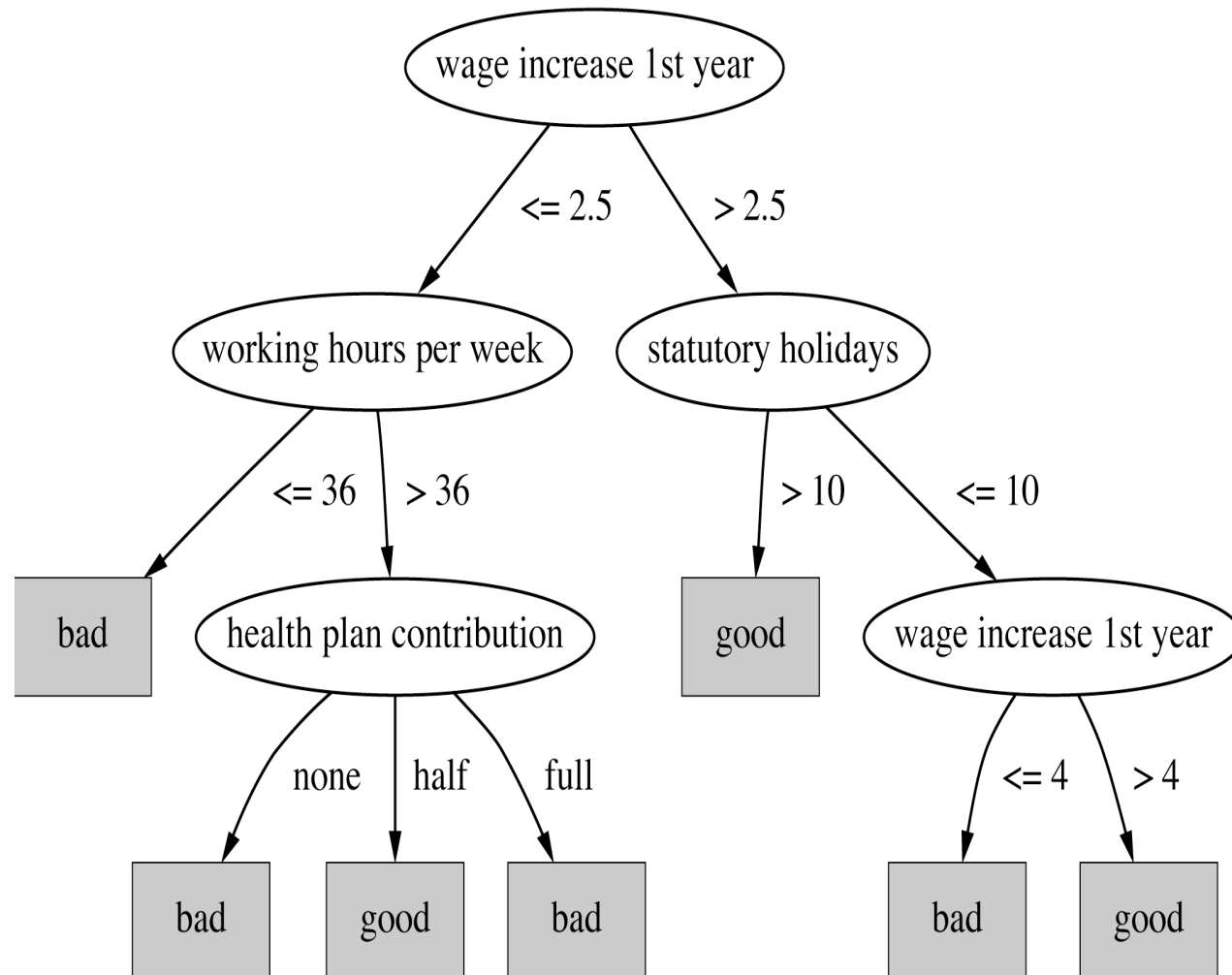
Decision tables

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

Outlook	Humidity	Play
Sunny	High	No
Sunny	Normal	Yes
Overcast	High	Yes
Overcast	Normal	Yes
Rainy	High	No
Rainy	Normal	No

Decision trees

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



Classification rules

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

If outlook = sunny and humidity > 83
then play = no

If outlook = rainy and windy = true
then play = no

If outlook = overcast
then play = yes

If humidity < 85
then play = yes

If none of the above
then play = yes

Association rules

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

If temperature = cool

then humidity = normal

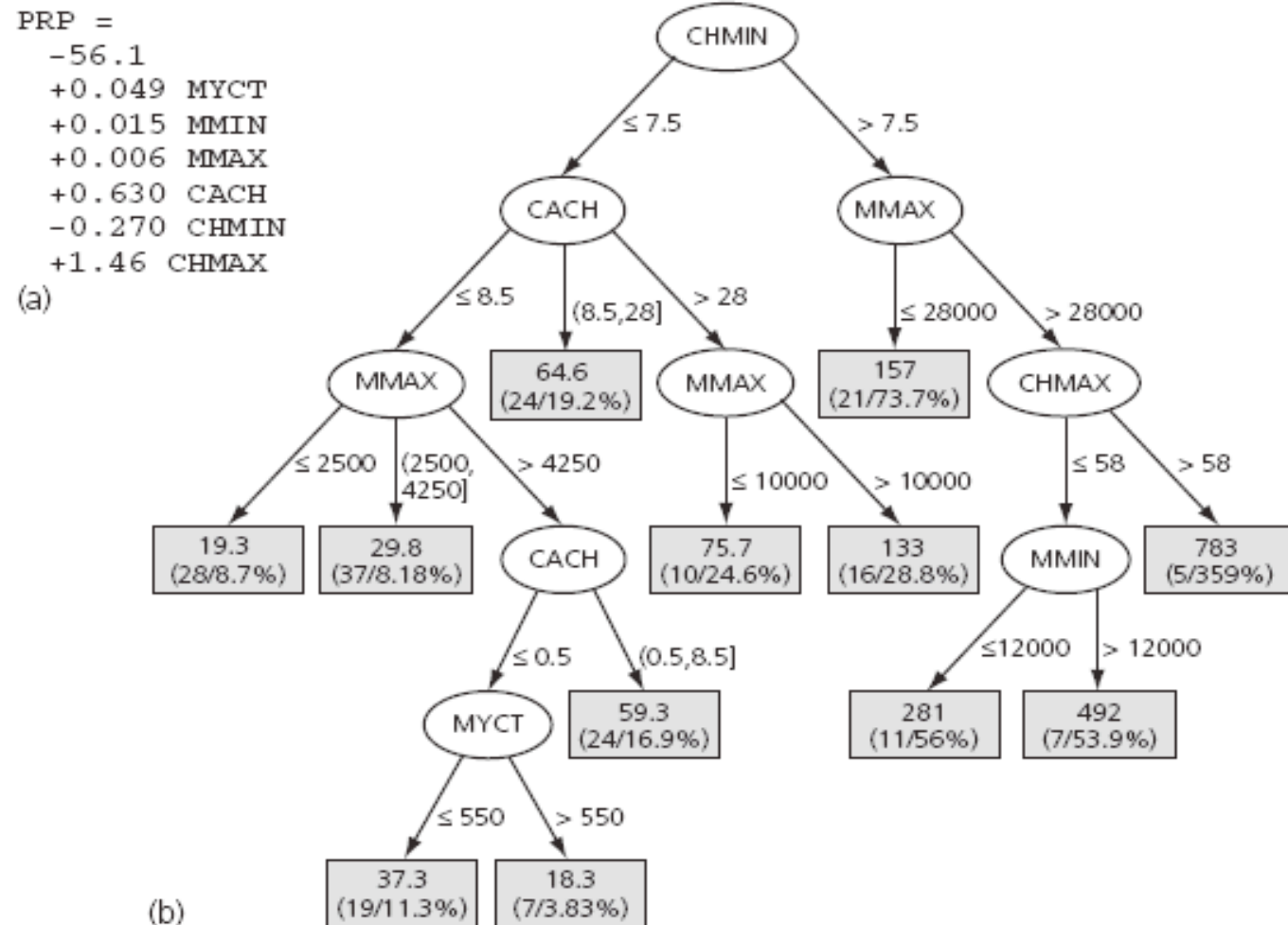
If windy = false and play = no

then outlook = sunny and humidity = high

- Συσχετίζουν τις τιμές των διάφορων attributes μεταξύ τους

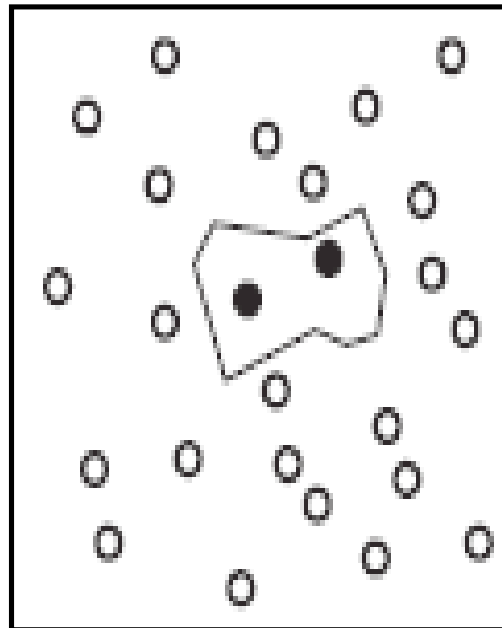
Regression tree

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
- Πίνακες
- Δένδρα
- Κανόνες
- Παλινδρόμηση
- Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



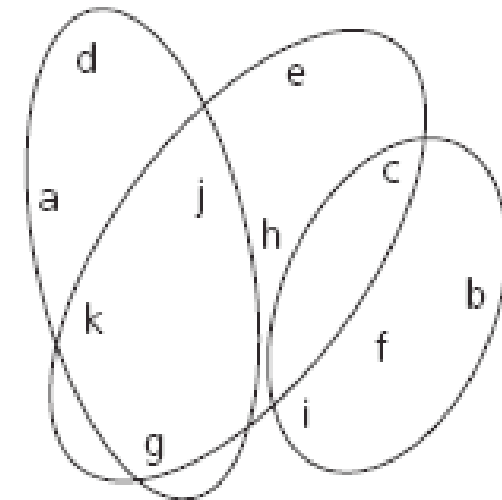
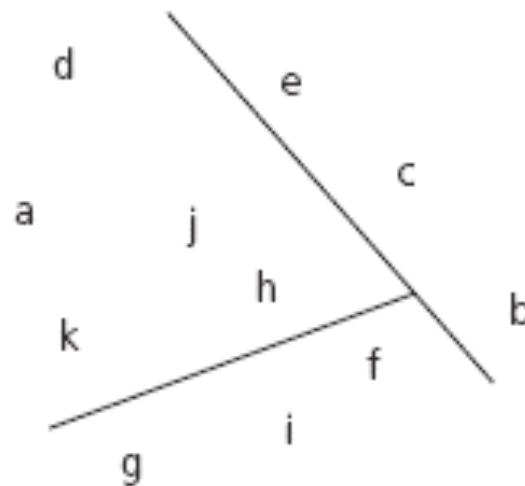
Instance-based representation

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
- Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



Clustering

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγόμενων
 - Πίνακες
 - Δένδρα
 - Κανόνες
 - Παλινδρόμηση
 - Ομάδες
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη



Αποτίμηση: το κλειδί της επιτυχίας

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
 - Αποτίμηση
 - Εκπαίδευση
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Πλήθος μεθόδων εξαγωγής δομικών περιγραφών από δεδομένα
 - Ποιες είναι οι βέλτιστες για συγκεκριμένο πρόβλημα;
 - Απαιτείται συστηματική προσέγγιση για την αποτίμηση και τη σύγκριση αποδοτικότητας των περιγραφών αυτών
- Αναφέρεται ρητά ότι το σφάλμα στα δεδομένα εκπαίδευσης δεν αποτελεί αξιόπιστο δείκτη απόδοσης σε μελλοντικά δεδομένα
- Λύση: διάσπαση των δεδομένων σε *σύνολο εκπαίδευσης (training set)* και *σύνολο ελέγχου (test set)*
 - Απαιτείται μεγάλος όγκος (ταξινομημένων) δεδομένων
 - Ωστόσο, ο όγκος των δεδομένων είναι συνήθως περιορισμένος
 - Αναγκαία η χρήση περισσότερο εκλεπτυσμένων τεχνικών

Εκπαίδευση & έλεγχος

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- **Αλγόριθμοι**
 - Αποτίμηση
 - **Εκπαίδευση**
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Μέτρο απόδοσης σε προβλήματα ταξινόμησης: **τιμή σφάλματος (error rate)**
 - **Επιτυχία (success)**: σωστή πρόβλεψη της τάξης του υποδείγματος
 - **Σφάλμα (error)**: λανθασμένη πρόβλεψη της τάξης του υποδείγματος
 - **Τιμή σφάλματος (error rate)**: αναλογία σφαλμάτων στο σύνολο των υποδειγμάτων
- Ωστόσο, ενδιαφέρει η πιθανή μελλοντική απόδοση σε νέα παραδείγματα και όχι η απόδοση στα ήδη δεδομένα υποδείγματα εκπαίδευσης...
- **Αποτελεί η τιμή σφάλματος σε ήδη γνωστά δεδομένα αξιόπιστη ένδειξη της τιμής σφάλματος σε νέα δεδομένα;**

Εκπαίδευση & έλεγχος

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- **Αλγόριθμοι**
 - Αποτίμηση
 - **Εκπαίδευση**
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Η απάντηση συνιστά ένα ηχηρό ΟΧΙ
 - στην περίπτωση που τα παλαιά αυτά δεδομένα έχουν χρησιμοποιηθεί στη διαμόρφωση του μοντέλου κατά τη διαδικασία εκπαίδευσης
- Καθώς η δομική περιγραφή έχει σχεδιαστεί με τρόπο ώστε να ελαχιστοποιεί το συγκεκριμένο σφάλμα, *οποιαδήποτε εκτίμηση της αποδοτικότητάς της βασισμένη στα υποδείγματα αυτά αποτελεί αισιόδοξη, αν όχι ανέλπιστα αισιόδοξη εκτίμηση*

Κριτήρια επιλογής μοντέλου

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- **Αλγόριθμοι**
 - Αποτίμηση
 - Εκπαίδευση
 - **Επιλογή**
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Αποπειρώνται να επιτύχουν θετικό συμβιβασμό μεταξύ:
 - Της πολυπλοκότητας του μοντέλου
 - Της ακρίβειας των προβλέψεών του στα δεδομένα εκπαίδευσης
- **Συλλογιστική:** άρτιο θεωρείται το μοντέλο που αφενός είναι απλό και αφετέρου επιτυγχάνει υψηλή ακρίβεια στα διαθέσιμα δεδομένα
 - Επίσης γνωστή και ως *Ξυράφι του Όκκαμ (Occam's Razor)*: βέλτιστη είναι η μικρότερη θεωρία που περιγράφει όλες τις πληροφορίες
- **Κατά τον Αλβέρτο Αϊνστάιν:** “Everything should be made as simple as possible, but no simpler.”

Κομψότητα vs. σφάλμα

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
 - Αποτίμηση
 - Εκπαίδευση
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Θεωρία Α: πολύ απλή, κομψή θεωρία που περιγράφει τα δεδομένα με σχεδόν απόλυτο τρόπο
- Θεωρία Β: σημαντικά περισσότερο πολύπλοκη θεωρία που περιγράφει τα δεδομένα χωρίς λάθη
- *Η Θεωρία Α είναι προφανώς προτιμώμενη*
 - Κλασικό παράδειγμα: οι τρεις νόμοι του Κέπλερ για την κίνηση των πλανητών
 - Λιγότερο ακριβείς από την τελευταία βελτίωση του Κοπέρνικου στη θεωρία των επικύκλων του Πτολεμαίου

Πρώτα τα βασικά

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
 - Αποτίμηση
 - Εκπαίδευση
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Πολύ συχνά, απλοί κανόνες επιτυγχάνουν μη αναμενόμενα υψηλή ακρίβεια
 - Αίτιο η απλοϊκή δομή που συχνά υποβόσκει πίσω από πραγματικά σύνολα δεδομένων
- Σε κάθε περίπτωση, συνίσταται αρχικά ο πειραματισμός με τις πλέον απλές των μεθόδων
 - The KISS principle: *Keep It Simple, Stupid!*

“Μετα”-μαθησιακά σχήματα

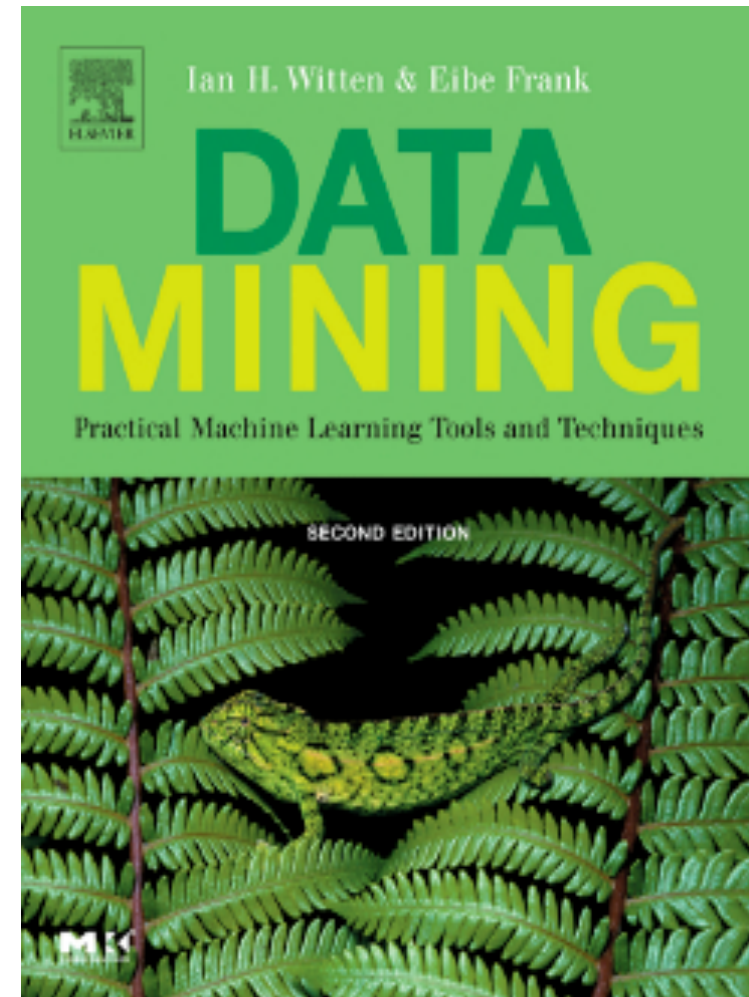
- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
 - Αποτίμηση
 - Εκπαίδευση
 - Επιλογή
 - Πρώτα τα βασικά
 - Μετα-αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Θεμελιώδης ιδέα: κατασκευή πολλών “εμπειρογνομόνων” (“experts”), ανάδειξη πλειοψηφούσας γνώμης
 - “εμπειρογνώμονας”: ένα μοντέλο που δημιουργήθηκε με τεχνικές μηχανική μάθησης
- Πλεονέκτημα:
 - Συχνά βελτιώνει σημαντικά την προβλεπτική ικανότητα
- Μειονέκτημα:
 - Τα εξαγόμενα είναι πολύ δύσκολο να αναλυθούν
- Σχήματα:
 - Εμφωλίαση (*bagging*)
 - Ενδυνάμωση (*boosting*)
 - Συσώρευση (*stacking*)
 - Κώδικες διόρθωσης σφαλμάτων εξόδου (*error-correcting output codes*)

Βιβλίο

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
 - Βιβλίο
 - WEKA
 - .arff
 - .xls σε .arff
- Επίδειξη

“Data Mining, Practical Machine Learning Tools and Techniques”,
Witten & Frank,
Morgan Kaufmann,
Ιούνιος 2005.



Λογισμικό

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
 - Βιβλίο
 - WEKA
 - .arff
 - .xls σε .arff
- Επίδειξη

WEKA: Waikato Environment for Knowledge Analysis

<http://www.cs.waikato.ac.nz/ml/weka/>



Η τυποποίηση .ARFF

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
 - Βιβλίο
 - WEKA
 - .arff
 - .xls σε .arff
- Επίδειξη

```
%  
% ARFF file for weather data with some numeric features  
%  
@relation weather  
  
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature numeric  
@attribute humidity numeric  
@attribute windy {true, false}  
@attribute play? {yes, no}  
  
@data  
sunny, 85, 85, false, no  
sunny, 80, 90, true, no  
overcast, 83, 86, false, yes  
...
```

Μετατροπή .xls σε .arff

- Εισαγωγή
- Case Studies
- Μορφή Δεδομένων
- Μορφή Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
 - Βιβλίο
 - WEKA
 - .arff
 - .xls σε .arff
- Επίδειξη

- Έστω πίνακας σε αρχείο .xls
- Save as → save as type → name.csv
- Open name.csv με .txt editor, για παράδειγμα notepad
- Πρόσθεσε το όνομα του dataset (*@relation*), τις πληροφορίες των χαρακτηριστικών (*@attribute*, μία σειρά για κάθε χαρακτηριστικό) και τη σειρά *@data*
- Save as type: all files & filename: dataset.arff

Επίδειξη

- Εισαγωγή
- Case Studies
- Μορφή
Δεδομένων
- Μορφή
Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

- Τρίτη 17 Ιουλίου, 35°C, 49 διαφάνειες μετά:
 - ...
 - Στην πράξη; Πώς τα εφαρμόζω όλα αυτά;



Τέλος

- Εισαγωγή
- Case Studies
- Μορφή
Δεδομένων
- Μορφή
Εξαγομένων
- Αλγόριθμοι
- Λογισμικό
- Επίδειξη

