ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Εισαγωγή

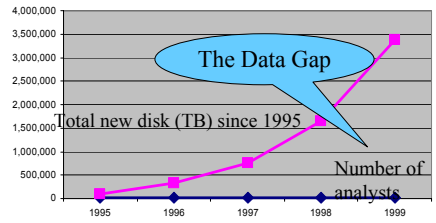
Τι είναι η Εξόρυξη Δεδομένων

(με δυο λόγια)

Αποδοτικές τεχνικές για να αναλύσουμε πολύ μεγάλες συλλογές από δεδομένα και να εξάγουμε χρήσιμες πληροφορίες από αυτά

Γιατί;

Συχνά υπάρχει πληροφορία «κρυμμένη» στα δεδομένα που δεν είναι προφανής
Οι ανθρώπινοι αναλυτές μπορεί να χρειάζονται εβδομάδες για να ανακαλύψουν χρήσιμη πληροφορία
Πολλά δεδομένα δεν αναλύονται ποτέ



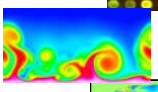
From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

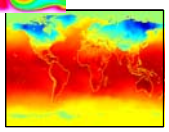
Γιατί Εξόρυξη Δεδομένων (από εμπορική πλευρά)

- Πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων
  - Web δεδομένα, e-εμπόριο
  - Αγορές σε πολύ-καταστήματα/αλυσίδες
  - Συναλλαγές με τράπεζες/πιστωτικές κάρτες
- Οι υπολογιστές γίνονται φτηνότεροι και πιο ισχυροί
- Μεγάλος ανταγωνισμός
  - Παροχή καλύτερων, προσωπικών υπηρεσιών σε κάποιο πεδίο (fraud detection, targeting marketing)

Γιατί Εξόρυξη Δεδομένων (από επιστημονική πλευρά)

- Τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες enormous speeds (GB/hour)
  - Απομακρυσμένοι αισθητήρες (remote sensors) σε δορυφόρους
  - Τηλεσκόπια στον ουρανό
  - Microarrays που παράγουν γονιδιακά δεδομένα
  - Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει τους επιστήμονες
  - Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων
  - Στην Διατύπωση Υποθέσεων





Παραδείγματα Δεδομένων

Κυβερνητικά: IRS (εφορία), δημογραφικά δεδομένα, …

Μεγάλες εταιρίες
- WALMART: 20M συναλλαγές την ημέρα
- MOBIL: 100 TB γεωλογικά σύνολα δεδομένων
- AT&T 300 M κλήσεις την ημέρα
- Εταιρίες πιστωτικών κρατών

Επιστημονικά
- NASA, EOS project: 50 GB την ώρα
- Δεδομένα για το περιβάλλον

«Κοινωνικά» - Ατομικά
- Νέα, ψηφιακές κάμερες, YouTube

Είδη/Τεχνικές Εξόρυξης Δεδομένων (συνοπτικά)

- **Ομαδοποίηση (συσταδοποίηση) – clustering**

χωρίζουμε τα δεδομένα σε ομάδες από «όμοια» σύνολα

- **Κανόνες συσχέτισης (Association rule mining)**

βρίσκουμε συσχετίσεις ανάμεσα στα δεδομένα, πχ ποια δεδομένα εμφανίζονται συχνά μαζί σε συναλλαγές

- **Κατηγοριοποίηση (Classification)**

κατηγοριοποιούμε τα δεδομένα τοποθετώντας τα σε μια (ή περισσότερες) από δοσμένες κατηγορίες

Είδη με βάση τα δεδομένα στα οποία γίνεται η εξόρυξη
- **Εξόρυξη στο διαδίκτυο**

μηχανές αναζήτησης – ενδιαφέρουσες (σημαντικές) σελίδες με βάση τους συνδέσμους

Εξόρυξη Δεδομένων (Ορισμός)

Πολύ μεγάλα σύνολα δεδομένων (data sets)

(1) η διαδικασία ανακάλυψης (discovery) προτύπων (patterns) που πριν δεν ήταν γνωστά, ισχύουν, είναι πιθανών χρήσιμα και είναι κατανοητά

(2) η ανάλυση τους για να βρούμε μη αναμενόμενες σχέσεις ανάμεσά στα δεδομένα καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες

Παραδείγματα: αγορές από πολυκαταστήματα, προσπελάσεις ιστοσελίδων, πακέτα στο δίκτυο, αποτελέσματα επιστημονικών πειραμάτων, κίνηση μετοχών, βιολογικά δεδομένα κλπ

- **Τι δεν είναι**
  - Αναζήτηση ενός αριθμού τηλεφώνου στον τηλεφωνικό κατάλογο
  - Ερώτηση σε μια μηχανή αναζήτησης πληροφορία για το "Amazon"
- **Τι είναι**
  - Ορισμένα ονόματα είναι πιο συχνά σε κάποιες τοποθεσίες στις ΗΠΑ (πχ O'Brien, O'Rurke, O'Reilly… στην περιοχή της Βοστώνης)
  - Ομαδοποίηση όμοιων κειμένων που επιστρέφει μια μηχανή αναζήτησης με βάση τα συμφραζόμενα (πχ δάσος Amazon, Amazon.com,)

Και λίγη ιστορία …

Σε σχέση με την ιστορία των Βάσεων Δεδομένων, η Εξόρυξη Δεδομένων είναι πολύ νέα …

1960 και νωρίτερα

Συλλογή Δεδομένων - Επεξεργασία Αρχείων

1970 – αρχές του 1980

Ιεραρχικά και δικτυακά μοντέλα
Σχεσιακά συστήματα βάσεων δεδομένων
Εργαλεία μοντελοποίησης (Ο/Σ κλπ)
Μέθοδοι ευρετηριοποίησης (Β-δέντρα, κατακερματισμός, κλπ)
Γλώσσες επερωτήσεων SQL, κλπ
Διεπαφές χρήστη (πχ φόρμες και αναφορές)
Επεξεργασία και βελτιστοποίηση ερωτήσεων
Συναλλαγές, ανάκαμψη από σφάλματα, έλεγχος συγχρονικότητας
OLTP (on-line analytical processing)

Και λίγη ιστορία …

Εξελιγμένα Συστήματα Βάσεων Δεδομένων

(μέσα 1980 – σήμερα)

- Νέα μοντέλα (αντικειμενο-σχεσιακό, επεκτεταμένα σχεσιακά κλπ)
- Νέες εφαρμογές και τύποι δεδομένων (χρονικά, χωρικά, χρονο-χωρικά, δεδομένα από αισθητήρες, συνεχή, κλπ)

Εξελιγμένη Ανάλυση Δεδομένων

Αποθήκες Δεδομένων και Εξόρυξη

(1990 – σήμερα)

Διαδικτυακές Βάσεις Δεδομένων

1990 – σήμερα

IR (Ανάκτηση Πληροφορίας) + ΒΔ

Οι «ρίζες» της Εξόρυξης Δεδομένων

Πρέπει να αντιμετωπίσει:

- **Το τεράστιο μέγεθος των δεδομένων**
- **Το μεγάλο αριθμό διαστάσεων**
- **Την μη ομοιογενή και την κατανεμημένη φύση των δεδομένων**

Η προσέγγιση στο μάθημα θα είναι σε αλγορίθμους/δομές και μεγάλα σύνολα δεδομένων – από την πλευρά των συστημάτων λογισμικού

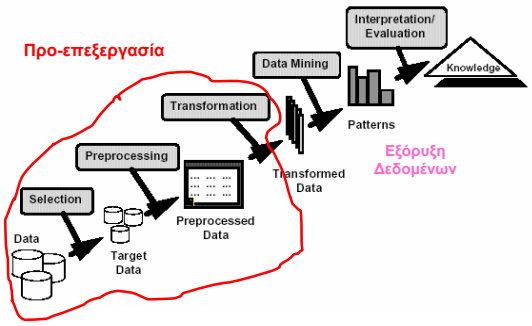
Κάποιες Πηγές – Σχετικές Κοινότητες

- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- **ACM SIGKDD** conferences since **1998** and **SIGKDD** Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Κάποιες Πηγές – Σχετικές Κοινότητες

- KDD Συνέδρια
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Άλλα Σχετικά Συνέδρια
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Περιοδικά
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Ανακάλυψη Γνώσης (Knowledge Discovery)



Ανακάλυψη Γνώσης (Knowledge Discovery)

ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ

- Data Cleaning – Καθαρισμός Δεδομένων
- Data Integration – Ενοποίηση Δεδομένων
- Data Transformation – Μετασχηματισμοί Δεδομένων

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

ΑΝΑΠΑΡΑΣΤΑΣΗ

Προ-επεξεργασία δεδομένων - Καθαρισμός

Τα δεδομένα στο πραγματικό κόσμο είναι «βρώμικα»

- Ελλειπή - incomplete: μπορεί να λείπουν κάποιες τιμές γνωρισμάτων (να μην καταγράφηκαν, να καταγράφηκαν λανθασμένα λόγω μη συνεννόησης ή λανθασμένης λειτουργίας), να λείπουν κάποια ενδιαφέροντα γνωρίσματα (που να μην θεωρήθηκαν σημαντικά ή να μην ήταν διαθέσιμα), ή να περιέχουν μόνο συναθροιστικά (aggregate) δεδομένα
  - Συμπλήρωση των γνωρισμάτων και τιμών που λείπουν
- Με θόρυβο - noisy: περιέχουν λάθη ή outliers (περιθωριακές τιμές – τιμές που διαφέρουν πολύ από την πλειοψηφία)
  - Εύρεση των περιθωριακών τιμών και απομάκρυνση θορύβου
- Ασυνεπή - inconsistent: περιέχουν ασυνέπειες, διπλότιμα
  - Διόρθωση ασυνεπών τιμών

Προ-επεξεργασία δεδομένων

Επιλογή Δεδομένων και Γνωρισμάτων και εφαρμογή κατάλληλων Μετασχηματισμών

- Συνάθροιση – Aggregation: συνδυασμούς δεδομένων από πολλές πηγές
- Sampling – δειγματοληψία: χρήση αντιπροσωπευτικού δείγματος των δεδομένων για βελτίωση της απόδοσης
- Dimensionality reduction – Κατάρα της διάστασης (curse of dimensionality)

Πολλές τεχνικές για την ανάλυση δεδομένων γίνονται δυσκολότερες με την αύξηση της διάστασης των δεδομένων (αυξάνει εκθετικά η πολυπλοκότητα ή τα δεδομένα γίνονται πολύ αραιά)

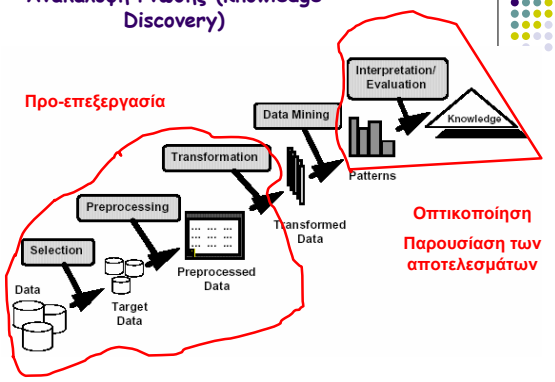
Τεχνικές της γραμμικής άλγεβρας (SVD, PCA)

Απεικόνιση σε άλλο χώρο με μικρότερο αριθμό διαστάσεων

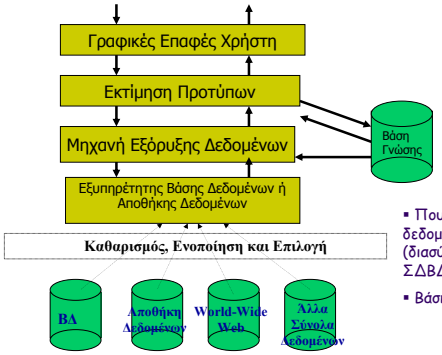
Προ-επεξεργασία δεδομένων

- Discretization (μετασχηματιμός σε μια διακριτή τιμή) ή binarization (μετασχηματισμός σε δυαδική τιμή)
- Variable transformation – μετασχηματισμοί των τιμών των μεταβλητών
  - Πχ Κανονικοποίηση

Ανακάλυψη Γνώσης (Knowledge Discovery)



Αρχιτεκτονική του Συστήματος



- Που είναι τα δεδομένα (διασύνδεση με ΣΔΒΔ)
- Βάση Γνώσης

Αποθήκες Δεδομένων

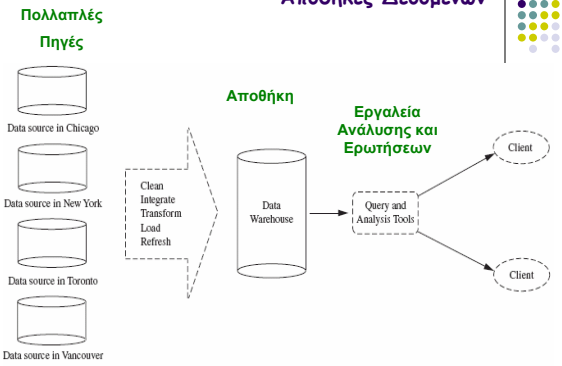
Αποθήκη δεδομένων είναι μια συλλογή από δεδομένα που συλλέγονται από διάφορες πηγές δεδομένων, αποθηκεύονται με βάση ένα κοινό σχήμα (συνήθως) σε έναν κόμβο

ExtractTransformLoad διαδικασίες – τα δεδομένα παίρνονται από τις βάσεις, μετασχηματίζονται και φορτώνονται στην αποθήκη

Οι μετασχηματισμοί μπορεί να είναι επιλογές συγκεκριμένων πεδίων και τιμών, αλλαγή μονάδων μέτρησης, καθαρισμός, κλπ

Περιοδική ενημέρωση της αποθήκης

Αποθήκες Δεδομένων



Αποθήκες Δεδομένων

Συνήθως ακολουθείται ένα πολύ-διάστατο σχήμα, όπου κάθε διάσταση αντιστοιχεί και σε ένα γνώρισμα (ή σύνολο) γνωρισμάτων του σχήματος και κάθε κελί σε μια μέτρηση

Το φυσικό σχήμα είναι συνήθως ένας πολυδιάστατος κύβος

Υποστηρίζουν OLAP (online analytical processing) λειτουργίες σε διαφορετικά επίπεδα λεπτομέρειας

Drill-down και roll-up

Αποθήκες Δεδομένων

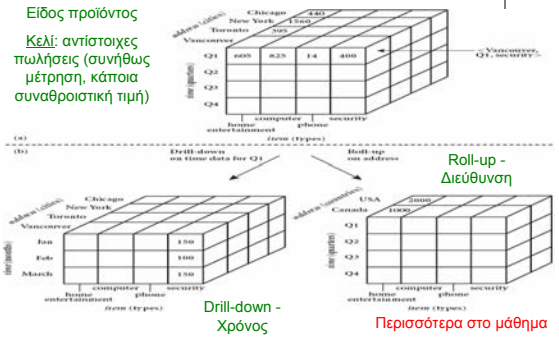
3-διαστάσεις

Διεύθυνση (πόλεις)

Χρόνος (τετράμηνα)

Είδος προϊόντος

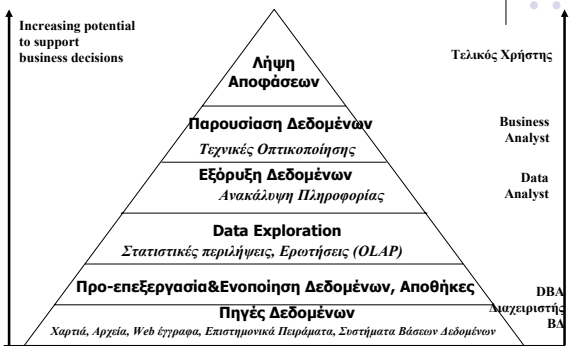
Κελί: αντίστοιχες πωλήσεις (συνήθως μέτρηση, κάποια συναθροιστική τιμή)



Roll-up - Διεύθυνση

Drill-down - Χρόνος

Περισσότερα στο μάθημα



Για το μάθημα

Ιστοσελίδα

http://www.cs.uoi.gr/~pitoura/courses/dm

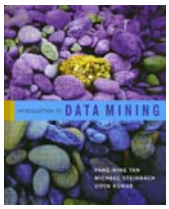
Βιβλία

Υπάρχουν 2 ελληνικά

- Μ. Βαζιργιάννης και Μ. Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων. Τυποθήτω, Νοέμβριος 2003
- M. H. Dunham, Data Mining, Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα.
  Επιμέλεια Ελληνικής Έκδοσης: Β. Βερύκιος και Γ. Θεοδωρίδης. Εκδόσεις Νέων Τεχνολογιών, 2004.

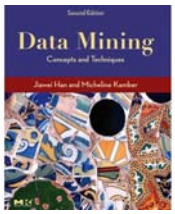
Για το μάθημα

2 «κλασικά» βιβλία στα αγγλικά



P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006

J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006



Αρκεί το υλικό στις διαφάνειες

Για το μάθημα

- 2 σύνολα ασκήσεων (κάποιες θεωρητικές και προγραμματιστικές ασκήσεις) – 50%
- Τελικό διαγώνισμα *(πιθανό)* – *50%*

Βασικοί Όροι

- **Δεδομένα (data)**
  Ένα σύνολο από στοιχεία (γεγονότα) D συνήθως αποθηκευμένα σε μια βάση δεδομένων
- **Γνωρίσματα (attributes)**
  Ένα πεδίο ενός στοιχείου I στο D
- **Πρότυπο (pattern)**
  Μια έκφραση E σε μια γλώσσα L που περιγράφει ένα υποσύνολο των δεδομένων του D
- **Βαθμός ενδιαφέροντος (Interestingness)**
  Μια συνάρτηση $I_{D,L}$ που απεικονίζει μια έκφραση E της L σε ένα πεδίο μετρήσεων M

Βασικοί Όροι

Το έργο της εξόρυξης δεδομένων (data mining task)

Δοσμένου
του συνόλου δεδομένων D,
μιας γλώσσας γεγονότων L,
μια συνάρτησης βαθμού ενδιαφέροντος $I_{D,L}$ και
ενός κατωφλίου c,
Βρες αποδοτικά
την έκφραση E τέτοια ώστε $I_{D,L}(E) > C$

Πως χρησιμοποιείται

1. Κατανόηση του προβλήματος
2. Χρήση τεχνικών εξόρυξης δεδομένων για να πάρουμε πληροφορία από τα δεδομένα
3. Χρήση αυτής της πληροφορίας
4. Μέτρηση των αποτελεσμάτων

Στη συνέχεια σήμερα, θα δούμε τα βασικά θέματα που θα μας απασχολήσουν

- Λειτουργικότητα/Είδη Εξόρυξης – Τι είδους πρότυπα μπορούν να εξορυχθούν
- Τεχνική/Μέθοδος για να πετύχουμε αυτήν την εξόρυξη

(συνοπτικά)

1. Ταξινόμηση - Classification: εκμάθηση μια συνάρτησης – κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο σε μια από ένα σύνολο από προκαθορισμένες κλάσεις
2. Συσταδοποίηση - Clustering: εύρεση ενός συνόλου από ομάδες με όμοια στοιχεία
3. Εύρεση Συχνών Προτύπων, Εξαρτήσεων και Συσχετίσεων – Dependencies and associations: εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ γνωρισμάτων
5. Συνοψίσεις - Summarization: εύρεση μιας συνοπτικής περιγραφής του συνόλου δεδομένων ή ενός υποσυνόλου του
6. Άλλα

Predictive Methods – Μέθοδοι πρόβλεψης
Χρήση κάποιων μεταβλητών για να προβλέψουν άγνωστες ή μελλοντικές τιμές κάποιων άλλων μεταβλητών

Descriptive Methods - Περιγραφικοί Μέθοδοι
Στόχος να βρεθούν κατανοητά πρότυπα που περιγράφουν τα δεδομένα – τις ιδιότητες τους

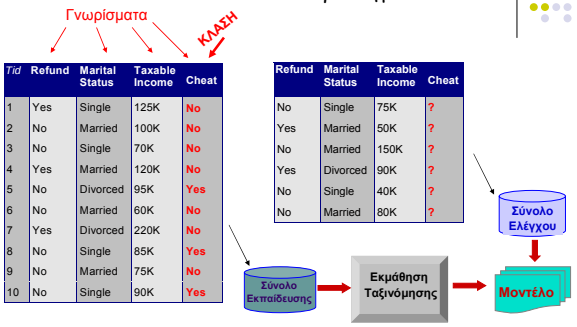
- Ταξινόμηση [Predictive]
- Συσταδοποίηση [Descriptive]
- Εύρεση Κανόνων Συσχέτισης [Descriptive]

- Sequential Pattern Discovery [Descriptive]
- Regression – Συνοψίσεις [Predictive]
  ένα συνοπτικό μοντέλο για τα δεδομένα (πχ μια συνάρτηση)
- Deviation/Anomaly Detection [Predictive]
  outlier analysis (στατιστικοί έλεγχοι για σπάνια σημεία), evolution analysis (πχ ανάλυση χρονοσειρών – πχ μετοχές) κλπ

Ορισμός

- Δοθέντος ενός συνόλου από εγγραφές (σύνολο εκπαίδευσης - training set)
  - Κάθε εγγραφή έχει ένα σύνολο από γνωρίσματα, ένα από αυτά είναι η κλάση (ή κατηγορία)
- Εύρεση ενός μοντέλου για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων.
- Στόχος: να αναθέτει σε εγγραφές που δεν έχουμε δει μια κλάση με την μεγαλύτερη δυνατή ακρίβεια
  - Για να χαρακτηρίσουμε την ακρίβεια του μοντέλου χρησιμοποιούμε ένα σύνολο ελέγχου (test set)
  - Συνήθως, το δοθέν σύνολο δεδομένο χωρίζεται σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου – το πρώτο χρησιμοποιείται για την κατασκευή του μοντέλου και το δεύτερο για τον έλεγχο του

Παράδειγμα



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

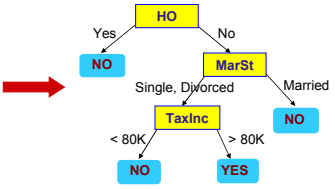
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Πως μοιάζει το μοντέλο;

Παράδειγμα Μοντέλου: Δέντρο Απόφασης – Decision tree

Tid	Home Owner	Marital Status	Taxable Income	Default
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ΚΛΑΣΗ

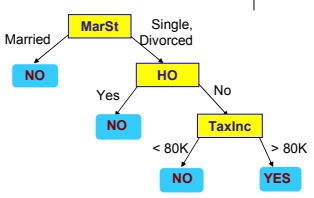


Σύνολο Εκπαίδευσης

Μοντέλο: Δέντρο Απόφασης

Tid	Home Owner	Marital Status	Taxable Income	Default
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ΚΛΑΣΗ



Για τα ίδια δεδομένα μπορεί να υπάρχουν παραπάνω από ένα δέντρα απόφασης (μοντέλα)

Regression analysis – ανάλυση παλινδρόμησης: στατιστική εκμάθηση μια συνάρτησης που απεικονίζει ένα στοιχείο σε μια πραγματική τιμή, χρήση για αριθμητικές προβλέψεις

Ανάλυση σχετικότητας (relevance analysis): ποια γνωρίσματα επηρεάζουν την ταξινόμηση

Άλλα είδη μοντέλων πλην των Δέντρων Απόφασης, νευρωνικά δίκτυα, κ-ποιο κοντινοί γείτονες, support vector machines κλπ

- *Στο μάθημα θα δούμε μόνο τα δέντρα απόφασης (αναλυτικά) + δομές για κοντινότερους γείτονες (πιθανόν)*

Ταξινόμηση: Εφαρμογή 1

Direct Marketing

Στόχος: Μείωση των ταχυδρομικών εξόδων για την αποστολή διαφημιστικών με τη στοχοποίηση *targeting* του συνόλου των πελατών που είναι πιο πιθανόν να αγοράσουν ένα κινητό τηλέφωνο

Προσέγγιση:
- Χρησιμοποίηση των δεδομένων από ένα παρόμοιο προϊόν που βγήκε στην αγορά πρόσφατα
- Για αυτό το προϊόν ξέρουμε ποιοι αποφάσισαν να το αγοράσουν και ποιοι όχι -> γνώρισμα της κλάσης {buy, don't buy}.
- Συλλογή ποικίλων δημογραφικών δεδομένων κλπ για αυτούς τους πελάτες
- Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.

Ταξινόμηση: Εφαρμογή 2

Fraud Detection – Αναγνώριση Απάτης σε Πιστωτικές Κάρτες

Στόχος: Να βρούμε ποιες συναλλαγές μιας πιστωτικής κάρτας δεν είναι από τον ιδιοκτήτη της

Προσέγγιση:
- Χρησιμοποίηση των δεδομένων από προηγούμενες συναλλαγές μα αυτήν την κάρτα και πληροφορίες για τον κάτοχο της (τι αγοράζει, πότε, από πού, πόσο συχνά πληρώνει)
- Χαρακτηρισμός κάθε προηγούμενης συναλλαγής ως απάτη ή όχι -> γνώρισμα της κλάσης {fraud, fair}.
- Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.
- Χρήση του μοντέλου για τον χαρακτηρισμό μελλοντικών συναλλαγών

Ταξινόμηση: Εφαρμογή 3

Customer Attrition

Στόχος: Να εκτιμήσουμε να ένας πελάτης θα προτιμήσει μια ανταγωνιστική εταιρεία

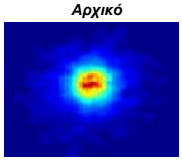
Προσέγγιση:
- Χρησιμοποίηση των δεδομένων από παλιές και νέες συναλλαγές πελατών (πόσο συχνά τηλεφωνούν, που πότε, την οικονομική του κατάσταση, την οικογενειακή του κατάσταση κλπ)
- Χαρακτηρισμός κάθε πελάτη ως πιστού ή όχι -> γνώρισμα της κλάσης {loyal, disloyal}.
- Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.

Ταξινόμηση: Εφαρμογή 4

Ταξινόμηση Γαλαξιών

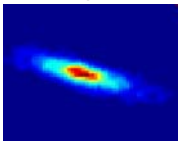
Courtesy: http://aps.umn.edu

Αρχικό



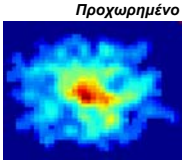
Κλάση:
- **Στάδιο δημιουργίας**

Ενδιάμεσο



Γνωρίσματα:
- **Χαρακτηριστικά της εικόνας,**
- **Χαρακτηριστικά του κυμάτων φωτός που ελήφθησαν, κλπ.**

Προχωρημένο



Μέγεθος Δεδομένων:
- **72 εκατ. άστρα, 20 εκατ. γαλαξίες**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Συσταδοποίηση

Ορισμός

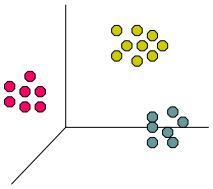
- Δοθέντων
  - Ενός συνόλου από σημεία που το καθένα έχει κάποια γνωρίσματα
  - Μιας μέτρηση ομοιότητας μεταξύ τους
- Εύρεση συστάδων (clusters) τέτοιων ώστε:
  - Τα σημεία σε μία συστάδα είναι πιο όμοια μεταξύ τους
  - Τα σημεία σε διαφορετικές συστάδες είναι λιγότερα όμοια μεταξύ τους

Σε αντίθεση με την ταξινόμηση, οι συστάδες δεν είναι γνωστές από πριν

Παράδειγμα

Οι αποστάσεις μέσα στη συστάδα ελαχιστοποιούνται

Οι αποστάσεις ανάμεσα στις συστάδες μεγιστοποιούνται



- 3-διάστατα σημεία, ευκλείδεια απόσταση

Συσταδοποίηση: Εφαρμογή 1

Market Segmentation

Στόχος: Χωρισμός των καταναλωτών σε ομάδες έτσι ώστε τα μέλη κάθε ομάδας να είναι ο στόχος για μια συγκεκριμένη πολιτική marketing

Προσέγγιση:
- Συγκέντρωση διαφορετικών γνωρισμάτων για τους καταναλωτές
- Ορισμός «ομοιότητας» ανάμεσα στους πελάτες
- Δημιουργία ομάδων με όμοιους πελάτες
- Μέτρηση της ποιότητας της ομαδοποίησης (πχ παρατηρώντας τις αγοραστικές συνήθειες στην ίδια ομάδα και ανάμεσα σε διαφορετικές ομάδες)

Συσταδοποίηση: Εφαρμογή 2

Συσταδοποίηση Εγγράφων

Στόχος: Εύρεση ομάδων από έγγραφα που είναι όμοια μεταξύ τους με βάση τους σημαντικούς όρους που εμφανίζονται σε αυτά

Προσέγγιση: Εύρεση για κάθε έγγραφο των όρων που εμφανίζονται συχνά σε αυτό.
Μέτρηση ομοιότητας με βάση τη συχνότητα των διαφορετικών όρων, Χρήση της για τη δημιουργία συστάδων

Όφελος: Μέθοδοι Ανάκτησης Πληροφορία (Information Retrieval) μπορεί να χρησιμοποιήσουν τις συστάδες για να συσχετίσουν έναν καινούργιο έγγραφο ή έναν όρο αναζήτησης με τα έγγραφα κάθε συστάδας

Συσταδοποίηση: Εφαρμογή 2

Σημεία: 3204 Άρθρα των Los Angeles Times.
Μέτρηση Ομοιότητας: Πόσες κοινές λέξεις έχουν

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Συσταδοποίηση

Στο μάθημα

Θα δούμε ενδιαφέροντες τρόπους να ορίσουμε ομοιότητα/απόσταση και τους θεμελιώδεις (και απλούς) αλγορίθμους συσταδοποίησης

Κανόνες Συσχέτισης

Ορισμός (συχνών στοιχειοσυνόλων)

- Δοθέντος
  - Ενός συνόλου από εγγραφές που η κάθε μία έχει έναν αριθμό από στοιχεία από κάποιο δοσμένο σύνολο
- Εύρεση κανόνων εξάρτησης που προβλέπουν την παρουσία ενός στοιχείου με βάση την παρουσία άλλων στοιχείων

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Κανόνες που βρέθηκαν:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Κανόνες Συσχέτισης: Εφαρμογή 1

Για marketing και προώθηση πωλήσεων:

Έστω ότι ο κανόνας που ανακαλύφθηκε είναι ο:
{Bagels, … } --> {Potato Chips}

Potato Chips στα δεξιά του κανόνα => Τι πρέπει να γίνει για να αυξηθούν οι πωλήσεις.

Bagels στα αριστερά => Μπορεί να χρησιμοποιηθεί για να εκτιμηθεί ποια προϊόντα θα επηρεαστούν αν πχ ένα μαγαζί σταματήσει να τα πουλάει.

Bagels στα αριστερά and Potato chips στα δεξιά => Ποια προιόντα πρέπει να πουληθούν μαζί με Bagels για την προώθηση των Potato chips!

Κανόνες Συσχέτισης: Εφαρμογή 2

Πως θα φτιάξουμε τα ράφια στα super-markets!

«Θρυλικός» κανόνας --

Αν ο καταναλωτής αγοράσει πάνες, πολύ πιθανών να αγοράσει και μπύρα!

Στις ΗΠΑ, Πέμπτη και Σάββατο, άντρες που αγοράζουν πάνες αγοράζουν και μπύρα

Εύρεση Ακολουθιακών Προτύπων

Ακολουθιακές εξαρτήσεις: μας ενδιαφέρει *η σειρά* εμφάνισης των στοιχείων (γεγονότων)

Παραδείγματα

- Ακολουθία από προσπελάσεις σελίδων στο διαδίκτυο
- Ακολουθία στο δανεισμό βιβλίων από μια βιβλιοθήκη
- Ακολουθία πακέτων που οδήγησαν σε επίθεση σε κάποιον υπολογιστή
- Σε χωρικά δεδομένα, πχ δεδομένα από την κίνηση ενός αυτοκινήτου

Κανόνες Συσχέτισης

Στο μάθημα

Θα μελετήσουμε ένα διάσημο αλγόριθμο τον a-priori

Και έναν ενδιαφέρον αλγόριθμο (FPGrowth) βασισμένο σε tries

Και πιθανών την εφαρμογή του a-priori σε γραφήματα

Η γενική εικόνα

- Εκμάθηση του πεδίου εφαρμογής
  - Σχετική προηγούμενη γνώση και τους στόχους της εφαρμογής
- Δημιουργία του συνόλου δεδομένων: data selection
- Καθαρισμός και προ-επεξεργασία των δεδομένων: (έως και 60% της συνολικής προσπάθειας)
- Ελάττωση δεδομένων και μετασχηματισμοί
  - Χρήσιμα χαρακτηριστικά, ελάττωση διαστάσεων κλπ
- Επιλογή λειτουργίας εξόρυξης δεδομένων
  - πχ, συσταδοποίηση, ταξινόμηση, κλπ
- Επιλογή του αλγορίθμου εξόρυξης δεδομένων
- Εξόρυξη Δεδομένων: αναζήτηση προτύπων ενδιαφέροντος
- Εκτίμηση προτύπων και αναπαράσταση γνώσης
  - οπτικοποίηση, μετασχηματισμοί, απομάκρυνση περιττών προτύπων, κλπ
- Χρήση της γνώσης

Εκτίμηση ενδιαφέροντος

Χαρακτηρισμό του «ενδιαφέροντος» ενός προτύπου:

(1) Εύκολα κατανοητό

(2) Να ισχύει σε δεδομένα ελέγχου ή σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας

(3) Πιθανών χρήσιμο

(4) Νέα πληροφορία

Υπάρχουν υποκειμενικά (αναμενόμενα και μη αναμενόμενα) και αντικειμενικά κριτήρια – Κάποιες τιμές κατωφλίου

Πληρότητα (όλα τα ενδιαφέροντα πρότυπα)

Βελτιστοποίηση (μόνο τα ενδιαφέροντα πρότυπα)

Web mining

Διάφορες τεχνικές (πχ ομαδοποίηση, ταξινόμηση) και

Διαφορετικά δεδομένα

Δομή ιστοσελίδων (συνδέσεις)

Web logs

ανάλυση κοινοτήτων στο web

Στο μάθημα θα δούμε κάποια γενικά στοιχεία και δυο διάσημους αλγόριθμους πίσω από τις μηχανές αναζήτησης

Υπάρχει σχετικό λογισμικό

Κάτι αντίστοιχο ενός ΣΔΒΔ;

Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL[3] as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
        and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```


OLEDB για DM (Microsoft'2000) και πιο πρόσφατα DMX (Microsoft SQLServer 2005)

- Βασισμένη σε OLE, OLE DB, OLE DB για OLAP, C#
- Συνδυασμός ΣΔΒΔ, Αποθηκών και εξόρυξης δεδομένων

DMML (Data Mining Mark-up Language) από την DMG (www.dmg.org)

Οι 10 καλύτεροι αλγόριθμοι ΕΔ (ICDM 2006)

#1: C4.5 (61 votes) – ταξινόμηση (δέντρο απόφασης)
#2: K-Means (60 votes) - συσταδοποίηση
#3: SVM (58 votes) – ταξινόμηση (support vector machine)
#4: Apriori (52 votes) – κανόνες συσχέτισης
#5: EM (48 votes) – στατιστική, συσταδοποίηση (expectation maximization)
#6: PageRank (46 votes) – ιστοσελίδες
#7: AdaBoost (45 votes) – μετα-ταξινομητής
#7: kNN (45 votes) – συσταδοποίηση (κοντινότερος γείτονας)
#7: Naive Bayes (45 votes) – στατιστική, ταξινόμηση
#10: CART (34 votes) – ταξινόμηση (δέντρο απόφασης)

ΣΥΝΟΨΗ: Τι θα καλύψουμε στο μάθημα (με τη σειρά)

- Ομαδοποίηση (συσταδοποίηση)
- Κανόνες Συσχέτισης
- Κατηγοριοποίηση (δέντρα απόφασης)
- Γραφήματα (πιθανών)
- Παγκόσμιο Ιστό
  - HITS, PageRank
- ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ

ΜΑΘΗΜΑ ΕΠΟΜΕΝΗΣ ΕΒΔΟΜΑΔΑΣ

Συσταδοποίηση

+ είδη δεδομένων και αποστάσεις (ομοιότητα)

ΠΟΤΕ;