Εξόρυξη Δεδομένων

2: Μέθοδοι εξόρυξης δεδομένων

Περιεχόμενα

- Επισκόπηση μεθόδων
  - Ταξινόμηση
  - Παλινδρόμηση
  - Συσταδοποίηση
  - Κανόνες συσχέτισης
  - Ακολουθιακά πρότυπα
- Εκτίμηση αποτελεσμάτων
- Η μάθηση ως αναζήτηση
- Επιρροή

Μέθοδοι εξόρυξης δεδομένων

Ανάλογα με τον τρόπο εξόρυξης

- Ταξινόμηση - Classification: εκμάθηση μια συνάρτησης – κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο σε μια (συνήθως) κλάση επιλέγοντας από ένα σύνολο από προκαθορισμένες κλάσεις
- Συσταδοποίηση - Clustering: εύρεση ενός συνόλου από ομάδες με όμοια στοιχεία
- Εύρεση Συχνών Προτύπων, Εξαρτήσεων και Συσχετίσεων – Dependencies and associations: εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ γνωρισμάτων
- Συνοψίσεις - Summarization: εύρεση μιας συνοπτικής περιγραφής του συνόλου δεδομένων ή ενός υποσυνόλου του
- …και άλλες

Ανάλογα με το στόχο

- 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 μεθόδων με βάση το αποτέλεσμα

- 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

Ταξινόμηση

Διαδικασία ταξινόμησης

- Αρχικά διαθέτουμε ένα σύνολο από εγγραφές
- Κάθε εγγραφή έχει ένα σύνολο από γνωρίσματα, ένα από αυτά είναι η κλάση (ή κατηγορία)
- Σκοπός: εύρεση ενός μοντέλου για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων.
  - να αναθέτει σε εγγραφές που δεν έχουμε δει μια κλάση με την μεγαλύτερη δυνατή ακρίβεια
- Το αρχικό σύνολο δεδομένων χωρίζεται σε ένα σύνολο εκπαίδευσης (training set) και σε ένα σύνολο ελέγχου (test set)
  - το training set χρησιμοποιείται για την κατασκευή του μοντέλου
  - το test set χρησιμοποιείται για να ελέγξει την ακρίβεια του μοντέλου

Παράδειγμα

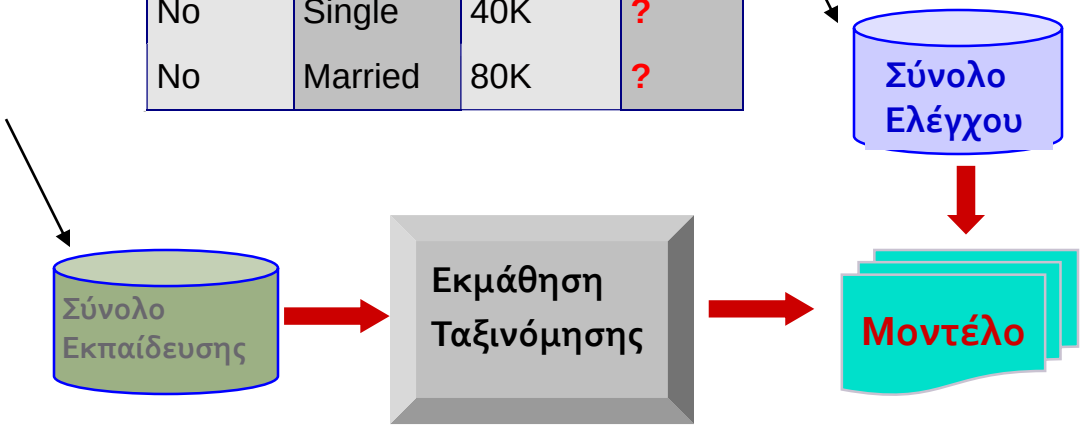
Γνωρίσματα

ΚΛΑΣΗ

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

Η τιμή του cheat είναι γνωστή για όλα τα δείγματα

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



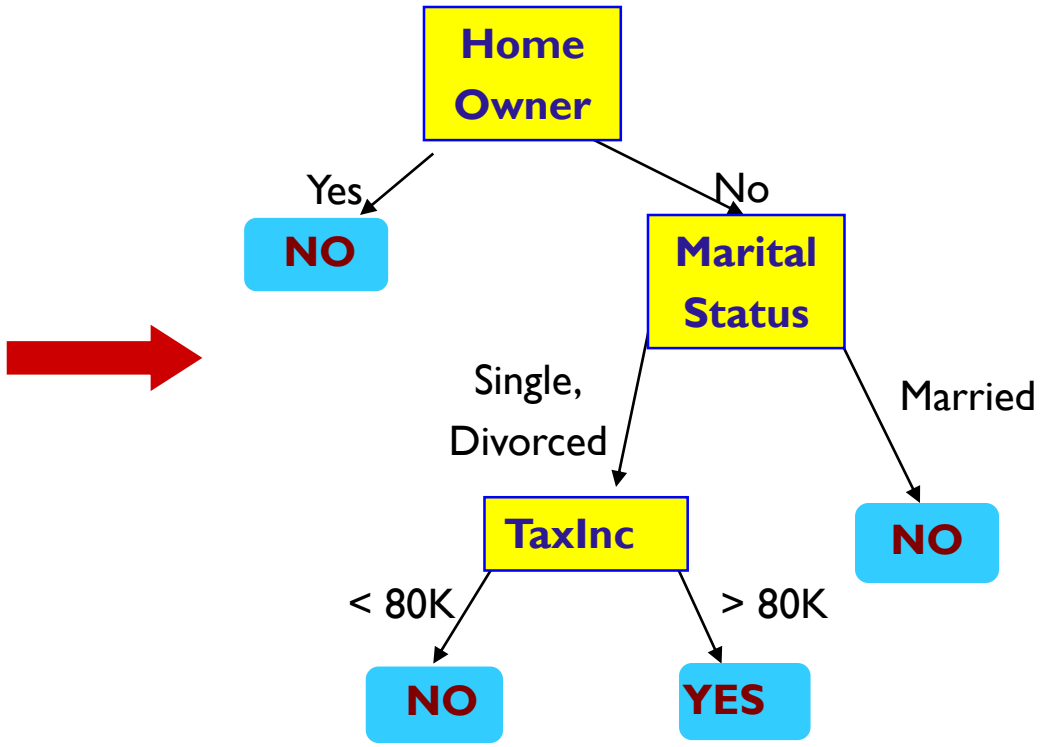
Πως μοιάζει το μοντέλο;

Παράδειγμα Μοντέλου: Δέντρο Απόφασης – Decision tree

ΚΛΑΣΗ

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

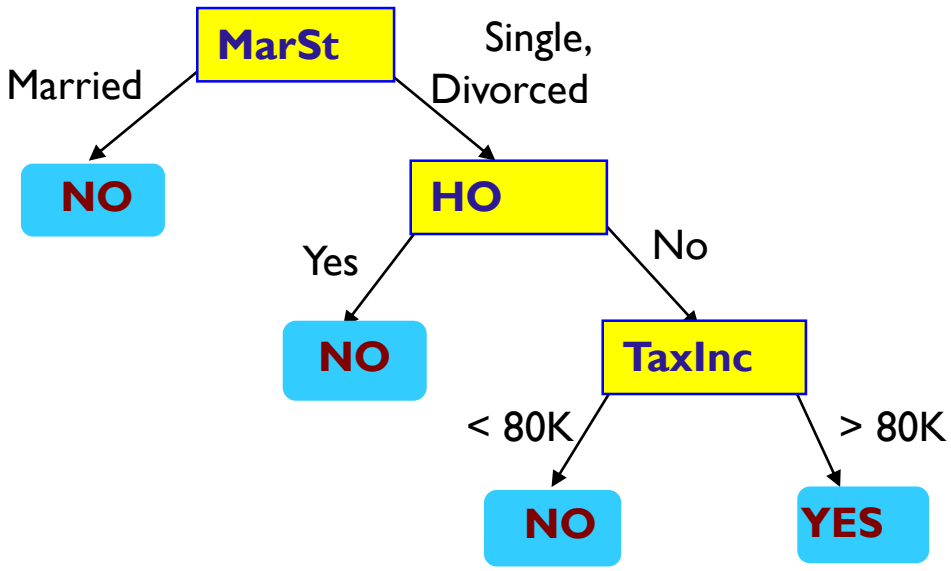
Σύνολο Εκπαίδευσης



Μοντέλο: Δέντρο Απόφασης

ΚΛΑΣΗ

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



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

Τεχνικές/Μοντέλα ταξινόμησης

- Μοντέλα
  - Δέντρα απόφασης, νευρωνικά δίκτυα, k-πιο κοντινοί γείτονες, support vector machines κλπ
- Ανάλυση σχετικότητας (relevance analysis): ποια γνωρίσματα επηρεάζουν την ταξινόμηση;

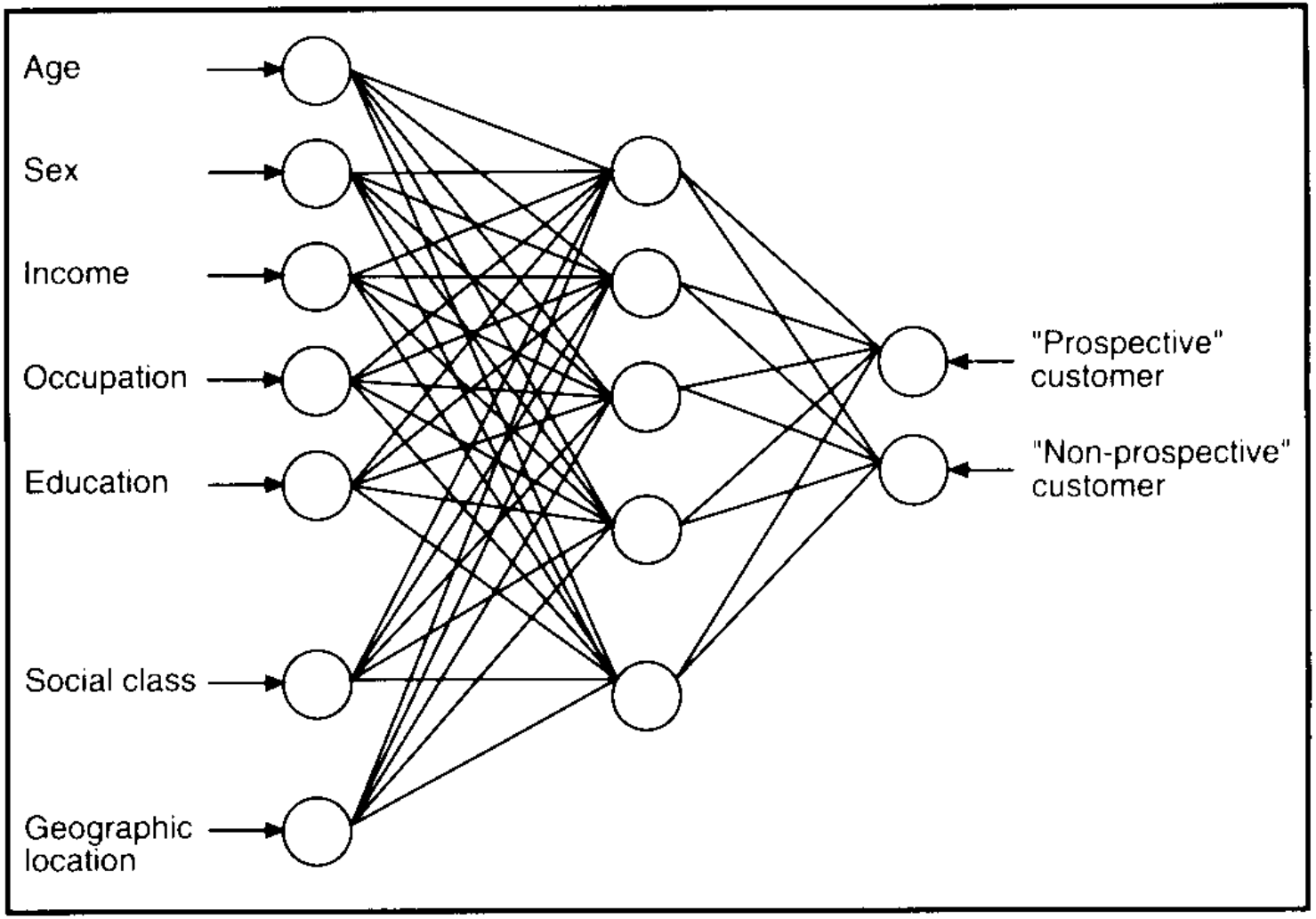
Ανάλυση σχετικότητας

- Επιλογή γνωρισμάτων
- Όταν «χτίζω» ένα δέντρο απόφασης
  - θέλω να ξεκινώ με αποφάσεις που διαχωρίζουν καλά τον πληθυσμό μου
  - θέλω να ελέγχω αρχικά τα γνωρίσματα που μου δίνουν το μεγαλύτερο κέρδος πληροφορίας – information gain
  - θέλω να αποφεύγω τις πολλές διασπάσεις στο δέντρο μου – overfitting
- Αναλύω τα γνωρίσματά μου και υπολογίζω μεγέθη όπως: πλήθος διαφορετικών τιμών, κατανομή τιμών σε ένα γνώρισμα κλπ.

Εφαρμογή 1: Direct Marketing

- Στόχος: Θέλω να στείλω διαφημιστικά μόνο σε εκείνους τους πελάτες που είναι πιο πιθανόν να αγοράσουν ένα κινητό τηλέφωνο. Έτσι μειώνω τα ταχυδρομικά μου έξοδα.
- Προσέγγιση:
  - Γνώρισμα κλάσης: Client_type{buy, don't buy}: ξέρω ποιοι αποφάσισαν να το αγοράσουν και ποιοι όχι
  - Χρησιμοποίηση των δεδομένων από ένα παρόμοιο προϊόν που βγήκε στην αγορά πρόσφατα
  - Συλλογή ποικίλων δημογραφικών κλπ δεδομένων για τους πελάτες
  - Η πληροφορία αυτή αποτελεί τα γνωρίσματα για την εκπαίδευση ενός μοντέλου ταξινόμησης

Νευρωνικό δίκτυο

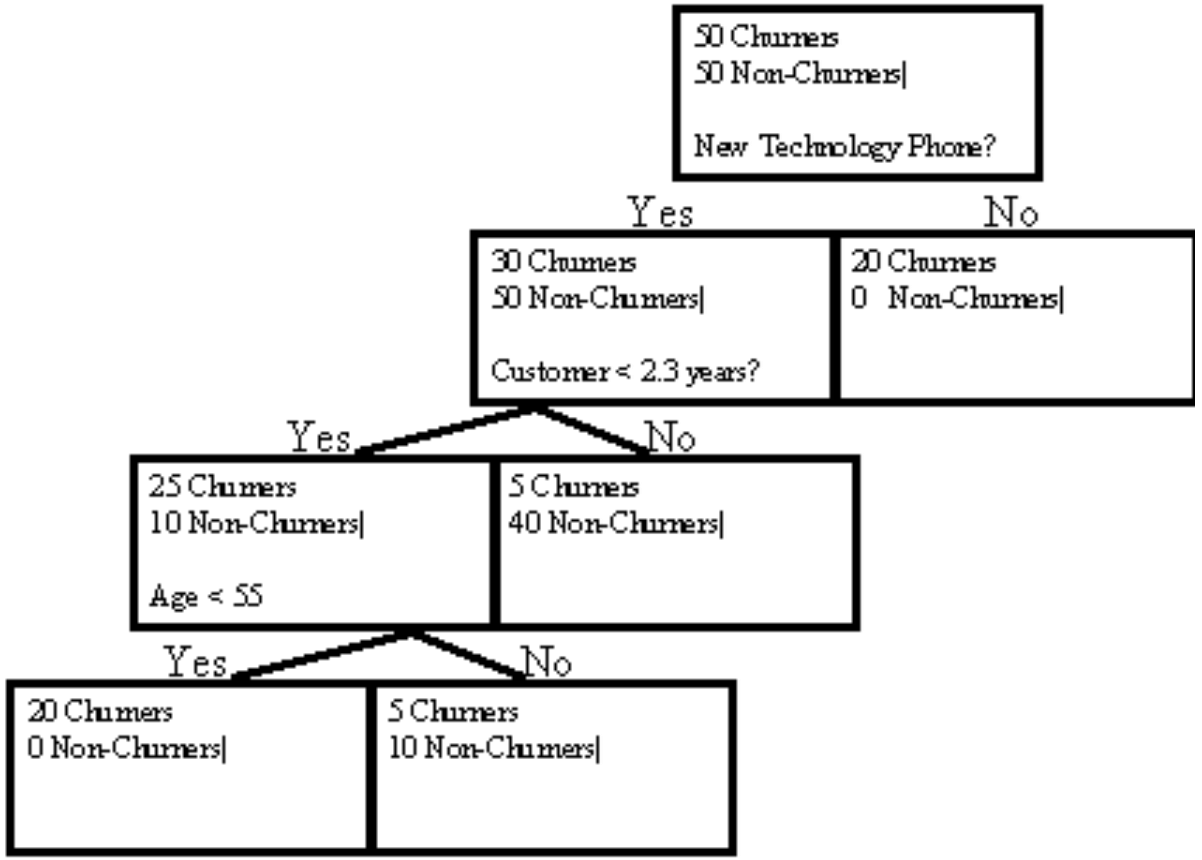


Εφαρμογή 2: Fraud Detection

- Στόχος: Θέλω να βρω ποιες συναλλαγές μιας πιστωτικής κάρτας δεν έγιναν από τον ιδιοκτήτη της
- Προσέγγιση:
  - Γνώρισμα κλάσης: transaction_type {fraud, fair}: Ξέρω κάθε προηγούμενη συναλλαγή αν ήταν απάτη ή όχι
  - Χρησιμοποίηση των δεδομένων από προηγούμενες συναλλαγές με αυτήν την κάρτα και πληροφορίες για τον κάτοχο της (τι αγοράζει, πότε, από πού, πόσο συχνά πληρώνει)
  - Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκπαίδευση ενός μοντέλου ταξινόμησης
  - Χρήση του μοντέλου για τον χαρακτηρισμό μελλοντικών συναλλαγών

Εφαρμογή 3: Customer Devotion

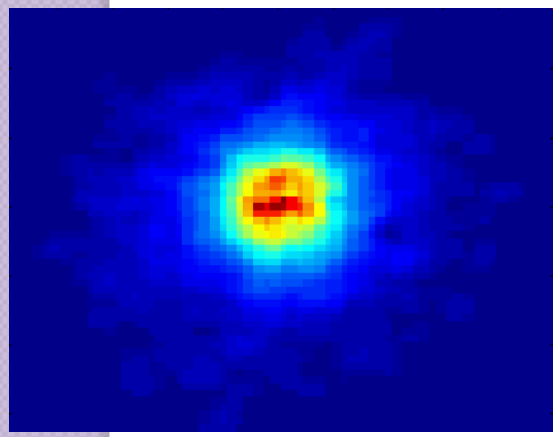
- Στόχος: Θέλω να εξετάσω αν οι πελάτες μιας εταιρίας τηλεπικοινωνιών θα προτιμήσουν μια ανταγωνιστική εταιρεία
- Προσέγγιση:
  - Γνώρισμα κλάσης: Client_type {loyal, disloyal}: Χαρακτηρισμός κάθε πελάτη ως πιστού ή όχι
  - Χρησιμοποίηση των δεδομένων από παλιές και νέες συναλλαγές πελατών (πόσο συχνά τηλεφωνούν, που πότε, οικονομική κατάσταση, οικογενειακή κατάσταση κλπ)
  - Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης


50 Churners
50 Non-Churners|
New Technology Phone?
Yes
No
30 Churners
50 Non-Churners|
Customer < 2.3 years?
20 Churners
0 Non-Churners|
Yes
No
25 Churners
10 Non-Churners|
Age < 55
5 Churners
40 Non-Churners|
Yes
No
20 Churners
0 Non-Churners|
5 Churners
10 Non-Churners|

Εφαρμογή 4: Image classification

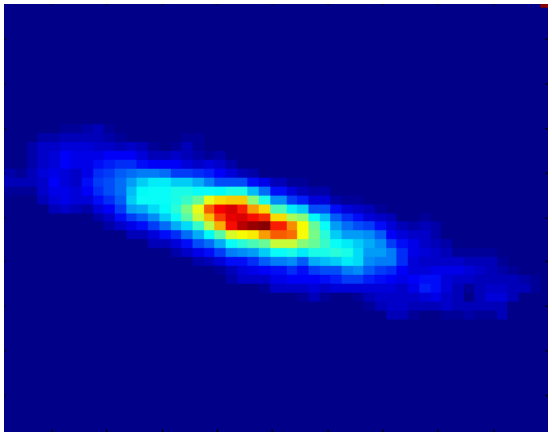
- Για ιατρικές, αστρονομικές κλπ εικόνες

Αρχικό



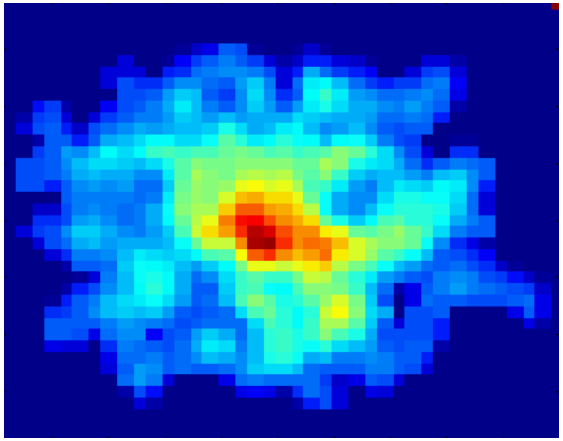
Κλάση:
- **Στάδιο δημιουργίας**

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**

Συνοψίζοντας

- Βρίσκω τα γνωρίσματα που με ενδιαφέρουν
- Τα μετατρέπω ώστε να ταιριάζουν στον αλγόριθμο
- Ορίζω το γνώρισμα κλάσης και τις διαφορετικές τιμές που μπορεί να πάρει
- Εκπαιδεύω το μοντέλο μου χρησιμοποιώντας μέρος των δεδομένων (ορισμένες μόνο συναλλαγές)
- Ελέγχω την αποτελεσματικότητα του μοντέλου μου χρησιμοποιώντας τα υπόλοιπα δεδομένα μου

Παλινδρόμηση

Ανάλυση παλινδρόμισης

- Regression analysis: στατιστική εκμάθηση μιας συνάρτησης που απεικονίζει ένα στοιχείο σε μια πραγματική τιμή
- Χρησιμοποιείται για αριθμητικές προβλέψεις
- Αναλύει τα υπάρχοντα δεδομένα για να καθορίσει τη συνάρτηση (συχνά γραμμική)
- Αξιοποιεί τη συνάρτηση για να προβλέψει αποτελέσματα για τα νέα δεδομένα

Παράδειγμα regression analysis

- Επιλέγω πέντε φοιτητές που παρακολουθούν το μάθημα της «Εξόρυξης Δεδομένων» και τους κάνω ένα τεστ στις βάσεις δεδομένων
- Θέλω να απαντήσω τα ακόλουθα:
  - Ποια συνάρτηση γραμμικής παλινδρόμησης προβλέπει καλύτερα την επίδοση στο μάθημα της εξόρυξης με βάση την επίδοση στο τεστ στις ΒΔ;
  - Αν κάποιος γράψει 80% στο τεστ, τι βαθμό να περιμένω στο μάθημα;
  - Πόσο καλά ταιριάζει στα δεδομένα η συνάρτηση παλινδρόμησης;

Δεδομένα

Τεστ

Μάθημα

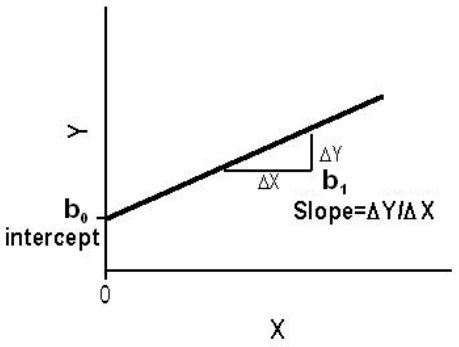
	Stud	x_i	y_i	$(x_i - x_\mu)$	$(y_i - y_\mu)$	$(x_i - x_\mu)^2$	$(y_i - y_\mu)^2$	$(x_i - x_\mu)(y_i - y_\mu)$
	1	95	85	17	8	289	64	136
	2	85	95	7	18	49	324	126
	3	80	70	2	-7	4	49	-14
	4	70	65	-8	-12	64	144	96
	5	60	70	-18	-7	324	49	126
Sum		390	385			730	630	470
Mean		78	77					

Γραμμική λύση: $y=b_0+b_1x$

Λύνουμε ως προς b_0 και b_1:

$b_1 = \Sigma [(x_i - x_\mu)(y_i - y_\mu)] / \Sigma [(x_i - x_\mu)^2] = 470/730 = 0.644$

$b_0 = y_\mu - b_1 * x_\mu = 77 - (0.644)(78) = 26.768$



Αποτελέσματα

- Εκτιμώμενος βαθμός για επίδοση 80% στο τεστ
  - $\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80 = 26.768 + 51.52 = 78.288$
  - Η τιμή του x πρέπει να είναι στα όρια που χρησιμοποιήθηκαν για την εκπαίδευση
  - Διαφορετικά μπορεί να έχουμε αλλοιώσεις (**extrapolation**)
- Υπολογισμός καταλληλότητας της συνάρτησης
  - Coefficient of Determination

$$R^2 = \left(\frac{1}{N} \frac{\sum_{i=1}^{N} (x_i - x_\mu)(y_i - y_\mu)}{\sigma_x \sigma_y} \right)^2$$

- N – πλήθος παρατηρήσεων, σ_x και σ_y - η τυπική απόκλιση των x και y:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x_\mu)^2}{N}}$$

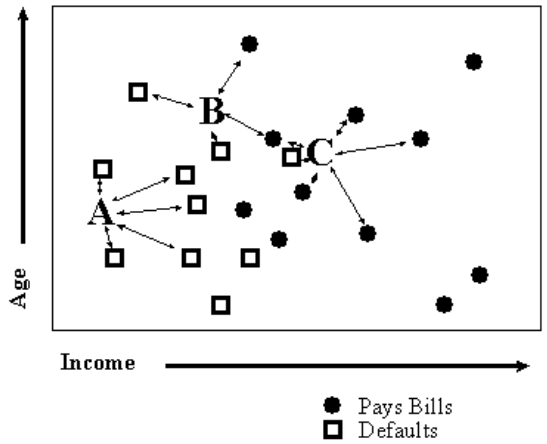
Συσταδοποίηση

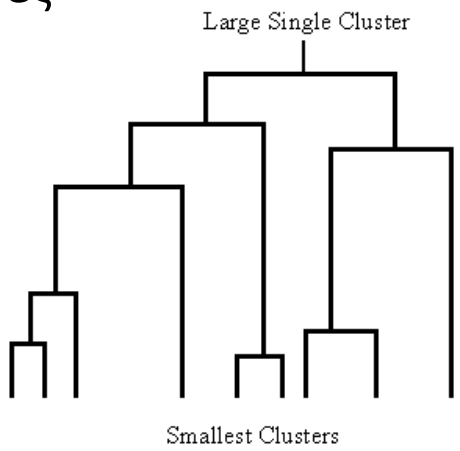
Ορισμός

- Μου δίνεται
  - Ένα σύνολο από αντικείμενα (σημεία) που καθένα έχει κάποια γνωρίσματα
  - Ένας τρόπος μέτρησης ομοιότητας μεταξύ αντικειμένων
- Εύρεση συστάδων (clusters, ομάδων) τέτοιων ώστε:
  - Τα σημεία της ίδιας συστάδας να είναι όσο το δυνατόν πιο όμοια μεταξύ τους
  - Τα σημεία σε διαφορετικές συστάδες να είναι όσο το δυνατόν λιγότερα όμοια μεταξύ τους
- Σε αντίθεση με την ταξινόμηση, οι συστάδες δεν είναι γνωστές από πριν
- Αλγόριθμοι: k-Means (πρέπει να ορίσουμε πόσα clusters θέλουμε), Ιεραρχικοί (δημιουργώ μια ιεραρχία από clusters και υπο-clusters αυτών)

Παράδειγμα

- Κριτήρια/απαιτήσεις
  - Οι αποστάσεις μέσα στη συστάδα (intra cluster similarity) ελαχιστοποιούνται
  - Οι αποστάσεις ανάμεσα στις συστάδες (intra cluster dissimilarity) μεγιστοποιούνται
- Ο ορισμός ενός μέτρου απόστασης/ομοιότητας μεταξύ των στοιχειωδών δεδομένων μου είναι το σημαντικότερο πρόβλημα στη συσταδοποίηση
- Το δεύτερο σημαντικότερο πρόβλημα είναι να χαρακτηρίσω (να περιγράψω) τις παραγόμενες συστάδες

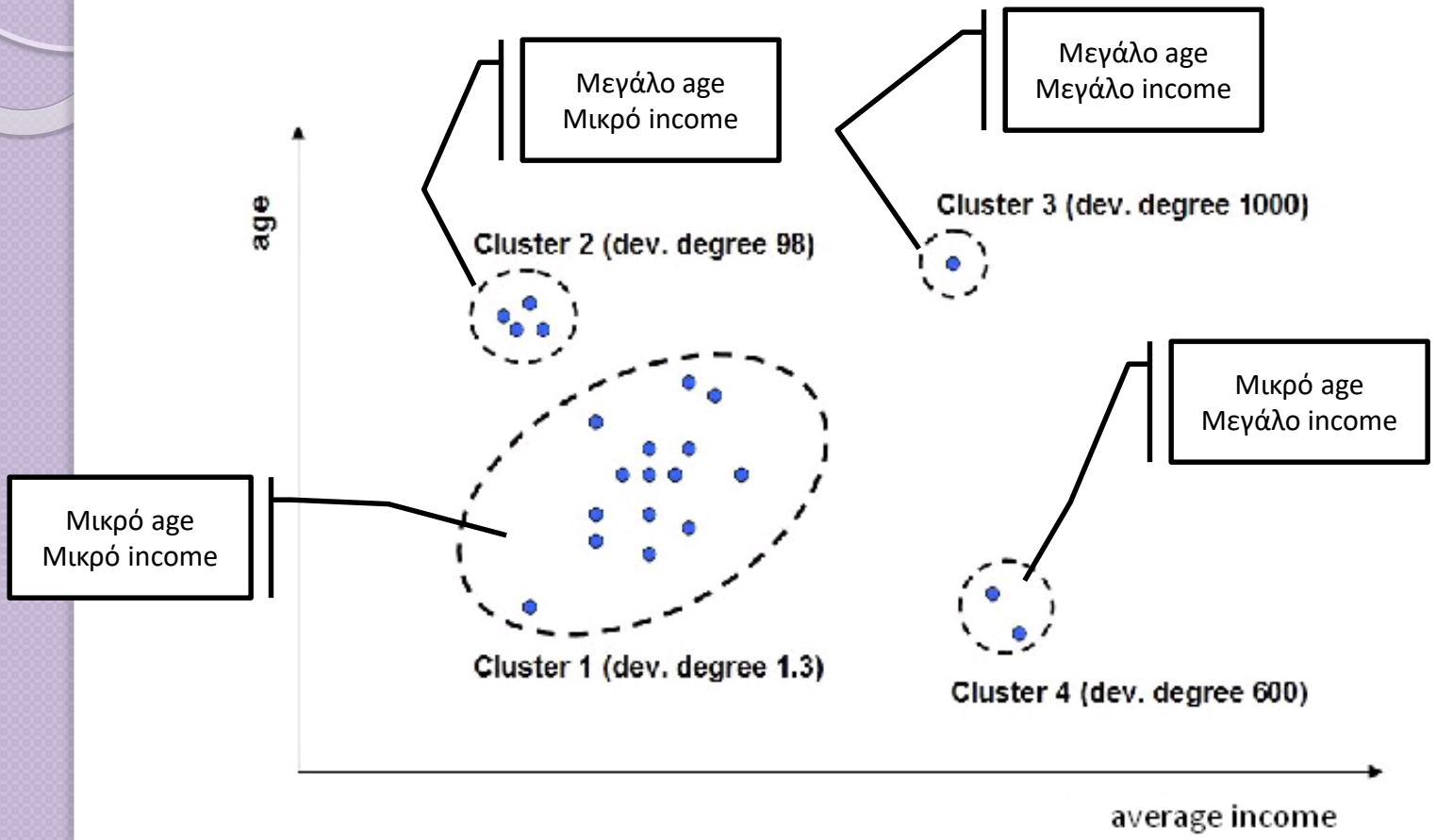




Εφαρμογή 1: Market Segmentation

- Στόχος: Θέλω να χωρίσω τους καταναλωτές σε ομάδες και στη συνέχεια να εφαρμόσω μια συγκεκριμένη πολιτική μάρκετινγκ για κάθε ομάδα
- Πρέπει επίσης να χαρακτηρίσω τις παραγόμενες ομάδες
- 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)

Παράδειγμα


Μεγάλο age
Μικρό income
Μεγάλο age
Μεγάλο income
Cluster 2 (dev. degree 98)
Cluster 3 (dev. degree 1000)
Μικρό age
Μεγάλο income
Μικρό age
Μικρό income
Cluster 1 (dev. degree 1.3)
Cluster 4 (dev. degree 600)
age
average income

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 των όρων που εμφανίζονται συχνά σε κάθε έγγραφο
  - Μέτρηση ομοιότητας με βάση τη συχνότητα των διαφορετικών όρων. Χρήση μέτρου 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

Παράδειγμα

Αντικείμενα: 3204 Άρθρα των Los Angeles Times
Μέτρηση Ομοιότητας: Πόσες κοινές λέξεις έχουν δύο κείμενα

cosine $\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$

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

Κανόνες συσχέτισης

Ορισμός

- Δεδομένα: Ένα σύνολο 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
- Βασίζονται στην εύρεση συχνών συνόλων στοιχείων
- Αλγόριθμοι: apriori, FP-Growth

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}

Εφαρμογή 1: Προώθηση πωλήσεων

- Έστω ότι ο κανόνας που ανακαλύφθηκε είναι:

  {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 π.χ. ένα μαγαζί σταματήσει να τα πουλάει.
- Bagels στα αριστερά and Potato chips στα δεξιά => Ποια προϊόντα πρέπει να πουληθούν μαζί με Bagels για την προώθηση των Potato chips!

Ακολουθιακές εξαρτήσεις

- 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

Συνοψίζοντας

- 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

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)

Εκτίμηση ενδιαφέροντος

- Χαρακτηρισμό του «ενδιαφέροντος» ενός προτύπου:
  - Εύκολα κατανοητό
  - Να ισχύει σε δεδομένα ελέγχου ή σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας
  - Πιθανά χρήσιμο
  - Νέα πληροφορία
- Υπάρχουν υποκειμενικά (αναμενόμενα και μη αναμενόμενα) και αντικειμενικά κριτήρια – Κάποιες τιμές κατωφλίου
- Πληρότητα (όλα τα ενδιαφέροντα πρότυπα)
- Βελτιστοποίηση (μόνο τα ενδιαφέροντα πρότυπα)

Μάθηση ως αναζήτηση

Μάθηση ως αναζήτησης

- Inductive learning: Βρίσκω ένα σύνολο κανόνων που ταιριάζουν στα δεδομένα
- Παράδειγμα: Τα σύνολα κανόνων χρησιμοποιούνται ως γλώσσα περιγραφής
  - Τεράστιος αλλά πεπερασμένος χώρος αναζήτησης
- Απλή λύση:
  - Απαρίθμηση του χώρου κανόνων
  - Διαγραφή όσων περιγραφών (descriptions) δεν ταιριάζουν στα δεδομένα
  - Τα σύνολα κανόνων (περιγραφές) που απομένουν ταιριάζουν στα δεδομένα

Παράδειγμα

- Ο χώρος αναζήτησης για το πρόβλημα του καιρού
  - 4 x 4 x 3 x 3 x 2 = 288 δυνατοί συνδυασμοί τιμών
  - Με 14 κανόνες ➔ 2.7×10^{34} πιθανά σύνολα κανόνων
- Λύση: Απαλοιφή κανόνων
- Πρακτικά προβλήματα
  - Να επικρατήσουν πολλαπλές περιγραφές
  - Να μη μείνει καμία περιγραφή
    - Αδυναμία περιγραφής του προβλήματος
    - Ύπαρξη θορύβου

Επιρροή - Bias

Επιρροή - Bias

- Σε μια διαδικασία εκπαίδευσης πρέπει αποφασίσουμε:
  - Τη γλώσσα/μοντέλο περιγραφής του χώρου
  - Τη σειρά αναζήτησης
  - Πώς θα αποφύγουμε το overfitting
- Αν δεν τα αποφασίσουμε αυτά έχουμε:
  - Επιρροή γλώσσας: Language bias
  - Επιρροή αναζήτησης: Search bias
  - Επιρροή υπερεκπαίδευσης: Overfitting-avoidance bias

Language bias

- Είναι γενικό το μοντέλο που χρησιμοποιούμε ή περιορίζει τη μάθηση;
- Μια Universal language μπορεί να εκφράζει κάθε παράδειγμα
- Αν το μοντέλο περιέχει τη διάζευξη τότε είναι γενικό (π.χ. rule sets)
- Η γνώση πεδίου μπορεί να εξαιρέσει ορισμένες περιγραφές πριν την έναρξη της αναζήτησης

Search bias

- Ευριστικό αναζήτησης
  - “Άπληστη” αναζήτηση: επιλέγω το καλύτερο βήμα κάθε φορά
  - “Beam search”: διατηρώ πάντα εναλλακτικές
  - …
- Κατεύθυνση αναζήτησης
  - *Γενικό-προς-ειδικό*
    - π.χ. εξειδίκευση κανόνων με προσθήκη επιπλέον συνθηκών
  - *Ειδικό-προς-γενικό*
    - π.χ. αρχικά γενικεύω κάθε στιγμιότυπο σε κανόνα

Overfitting-avoidance bias

- Θεωρείται ένα είδος search bias
- Τροποποιώ το κριτήριο αξιολόγησης
  - π.χ. ανάμεσα στην απλότητα των κανόνων και στο πλήθος σφαλμάτων
- Τροποποιώ την στρατηγική αναζήτησης
  - π.χ. με κλάδεμα/pruning κανόνων (απλοποίηση της περιγραφής)
    - Pre-pruning: σταματώ σε απλές περιγραφές και δεν αναζητώ πιο σύνθετες
    - Post-pruning: παράγω σύνθετες περιγραφές και στη συνέχεια τις απλοποιώ