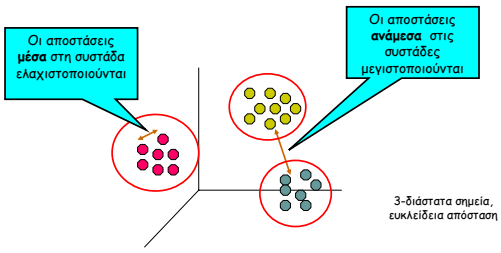
Συσταδοποίηση Ι

Μέρος των διαφανειών είναι από το **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006

Τι είναι συσταδοποίηση

Εύρεση συστάδων αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων ομάδων



Εφαρμογές

ομαδοποίηση γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία,

χαρακτηριστικά ασθενειών

μετοχών με παρόμοια διακύμανση τιμών,

ομαδοποίηση weblog για εύρεση παρόμοιων προτύπων προσπέλασης, ομαδοποίηση σχετιζόμενων αρχείων για browsing, ομαδοποίηση κειμένων κλπ

πελάτες με παρόμοια συμπεριφορά

	Discovered Clusters	Industry Group
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Συσταδοποίηση επιπέδου βροχής (precipitation) στην Αυστραλία!

Εφαρμογές

Κατανόηση – Stand-alone εφαρμογή/εργαλείο
οπτικοποίηση, συμπεράσματα για την κατανομή

Βήμα Προεπεξεργασίας

Περίληψη: Ελάττωση του μεγέθους μεγάλων συνόλων χρήση αντιπροσωπευτικών σημείων από κάθε συστάδα – πρωτότυπα (prototypes),
Συμπίεση ή
Αποδοτική κατασκευή ευρετηρίων – εύρεση κοντινότερου γείτονα κλπ

Εισαγωγή

Πότε μια συσταδοποίηση είναι καλή;

Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας

- Μεγάλη ομοιότητα εντός της συστάδας και
- Μικρή ομοιότητα ανάμεσα στις συστάδες

Η ποιότητα εξαρτάται από τη

- Μέτρηση ομοιότητας και
- Μέθοδο υλοποίησης της συσταδοποίησης

Θέματα που θα μας απασχολήσουν σήμερα

- Τι σημαίνει απόσταση/ομοιότητα;
- Είδη συσταδοποίησης
- Ένα βασικό αλγόριθμο συσταδοποίησης (k-means)

Πρώτα, ας δούμε λίγο τι μπορεί να είναι τα δεδομένα …

Δεδομένα και Γνωρίσματα

Δεδομένα και Γνωρίσματα

Συλλογή από αντικείμενα - δεδομένα.

Τα δεδομένα έχουν γνωρίσματα.
Ένα γνώρισμα (attribute) είναι μια ιδιότητα ή χαρακτηριστικό του αντικειμένου
Για παράδειγμα: χρώμα ματιών, θερμοκρασία, γεωγραφικές συντεταγμένες κλπ.

Μια συλλογή από γνωρίσματα περιγράφει ένα αντικείμενο

Διάσταση - αναφέρεται στον αριθμό των γνωρισμάτων

Αντικείμενο (δεδομένο) – πλειάδα

Συνήθως, τέτοια σύνολα πλειάδων αποθηκευμένα σε απλά αρχεία ή σε σχεσιακές βάσεις δεδομένων

Γνωρίσματα

Αντικείμενα

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

Δεδομένα και Γνωρίσματα

Κάθε γνώρισμα παίρνει τιμές από ένα πεδίο. Οι τιμές μπορεί να είναι αριθμοί η σύμβολα

Το ίδιο γνώρισμα μπορεί να παίρνει διαφορετικές τιμές, πχ όταν διαφέρει η μονάδα μέτρησης - Η θερμοκρασία μπορεί να μετριέται σε Κελσίου και Φαρενάιτ

Διαφορετικά γνωρίσματα τιμές από το ίδιο πεδίο τιμών

Ορολογία
- Ένα αντικείμενο λέγεται και *εγγραφή (record), σημείο (point), οντότητα entity, δείγμα (sample), περίπτωση (case),* ή *στιγμιότυπο (instance).*
- Το γνώρισμα λέγεται και *field (πεδίο), χαρακτηριστικό (characteristic) μεταβλητή (variable), feature*

Δεδομένα και Γνωρίσματα

Τι άλλα δεδομένα έχουμε;

Εγγραφές
- Πίνακες
- Κείμενα
- Συναλλαγές

Γραφήματα
- Παγκόσμιος Ιστός - World Wide Web
- Μοριακές Δομές

Διατάξεις
- Χωρικά Δεδομένα
- Χρονικά Δεδομένα
- Ακολουθίες
- Γενετικές Ακολουθίες

Δεδομένα και Γνωρίσματα

Εγγραφές

Όπως είδαμε, στη γενική περίπτωση - συλλογή από εγγραφές με προκαθορισμένο αριθμό γνωρισμάτων

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

Δεδομένα και Γνωρίσματα

Πίνακες

Αν στις εγγραφές τα γνωρίσματα παίρνουν *αριθμητικές τιμές* τότε τα αντικείμενα μπορεί να τα θεωρήσουμε και ως σημεία σε πολύ-διάστατο χώρο

Και ως n x m πίνακα, n αντικείμενα, m γνωρίσματα

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Δεδομένα και Γνωρίσματα

Έγγραφα

Κάθε έγγραφο, ένα διάνυσμα όρων
Τόσες διαστάσεις όσοι οι όροι
Τιμή σε κάθε διάσταση είναι ίσος με το πλήθος τω εμφανίσεων του αντίστοιχου όρου στο κείμενο

	team	coach	play	ball	score	game	win	lost	timeout	season

Δεδομένα και Γνωρίσματα

Συναλλαγές

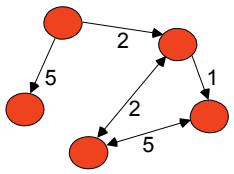
Κάθε εγγραφή περιέχει ένα σύνολο από στοιχεία

Παράδειγμα:
αγορές από κατάστημα, όπου κάθε συναλλαγή είναι το σύνολο των προϊόντων που αγοράστηκαν σε μια επίσκεψη σε ένα μαγαζί

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

Δεδομένα και Γνωρίσματα

Γραφήματα



Γενικό γράφημα

```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Μια HTML σελίδα

0
7
1

Δεδομένα και Γνωρίσματα

Ακολουθίες Συναλλαγών

Γεγονότα

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

Στοιχείο της ακολουθίας

Δεδομένα και Γνωρίσματα

Διατάξεις

Genomic sequence data

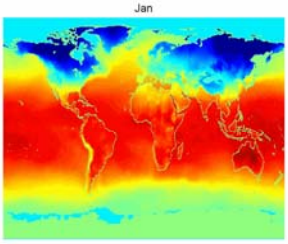
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```


Δεδομένα και Γνωρίσματα

Διατάξεις

Χωρο-χρονικά Δεδομένα

Μέση Μηνιαία Θερμοκρασία



Δεδομένα και Γνωρίσματα

- Η συσταδοποίηση είναι δυνατή σε όλες αυτές τις κατηγορίες δεδομένων
- Σημασία έχει η έννοια της ομοιότητας
- Διαφορετικές μέθοδοι/είδη συσταδοποίσης πιο κατάλληλες για καθεμία κατηγορία

Κάποια θέματα που μας απασχολούν

- Πολλές διαστάσεις (curse of dimensionality)
- Αραιά δεδομένα (sparsity) – μας ενδιαφέρουν μόνο ορισμένες τιμές
- Ακρίβεια τιμών (resolution)

Είδη Γνωρισμάτων

Ας δούμε τα βασικά είδη γνωρισμάτων για δεδομένα τύπου εγγραφών

Το είδος των γνωρισμάτων εξαρτάται από τι ιδιότητες έχει:

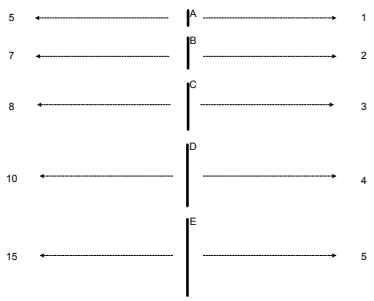
Κατηγορικά
- Διακριτότητα, Distinctness: = ≠
- Διάταξη - Order: < >

Αριθμητικά
- Προσθετική- Addition: + -
- Πολλαπλασιαστική - Multiplication: * /

Είδη Γνωρισμάτων

	Τύπος Δεδομένου	Περιγραφή	Παραδείγματα
Ποιοτικά ή Κατηγορικά	Nominal	Οι τιμές είναι απλώς διαφορετικά ονόματα (αναγνωριστικά) με αρκετή πληροφορία ώστε να γίνει διάκριση ανάμεσα τους (=, ≠) (και οι δυαδικές μεταβλητές 0 – 1)	ταχυδρομικός κωδικός, αριθμός ταυτότητας, χρώμα ματιών, φύλο
	Διάταξη - Ordinal	Οι τιμές περιέχουν πληροφορία διάταξης (<, >)	Ποιότητα υλικού (καλή, πιο καλή, άριστη), βαθμοί: {καλών, λίαν καλώς, άριστα}, αριθμοί στις διευθύνσεις
Ποσοτικά ή Αριθμητικά	Διαστήματος - Interval	Έχει σημασία η διαφορά μεταξύ δύο τιμών, υπάρχει μονάδα μέτρησης (+, -)	Θερμοκρασία σε Celsius ή Fahrenheit
	Ratio	Έχει σημασία και ο λόγος μεταξύ δύο τιμών (*, /)	Νομισματικές ποσότητες, ηλικία, θερμοκρασία σε Kelvin, ηλικία, μήκος

Δεδομένα και Γνωρίσματα



Διαφορετικοί τρόποι να μετρήσουμε το μήκος

Ο τρόπος στα αριστερά μόνο διάταξη – ο τρόπος στα δεξιά και προσθετικός

Είδη Γνωρισμάτων

	Μετασχηματισμοί	Παράδειγμα
Nominal	Οποιοσδήποτε ένα-προς-ένα απεικόνιση (πχ permutation)	Πχ δεν έχει διαφορά αν ξαναδώσουμε από την αρχή αριθμούς ταυτότητας ή διαβατηρίου
Ordinal	Αλλαγή τιμών που να διατηρεί την διάταξη πχ *νέα_τιμή = f(παλιά_τιμή)* όπου *f* μονότονη συνάρτηση.	Ένα γνώρισμα για διαβάθμιση μπορεί να είναι {C, B, A} ή {1, 2, 3} ή { 0.5, 1, 10}.
Interval	*νέα_τιμή =a * παλιά_τιμή + b* όπου *a* και *b* σταθερές	Εξαρτάται από που είναι τι μηδέν και το μέγεθος της μονάδας (πχ μεταξύ Fahrenheit και Celsius)
Ratio	*νέα_τιμή = a * παλιά_τιμή,* όπου *a* σταθερά	Μήκος σε μέτρα ή πόδια

Απόσταση και Ομοιότητα

Κριτήρια Ομοιότητας -Απόσταση

Ομοιότητα
Μια αριθμητική μέτρηση για το πόσο όμοια είναι δυο αντικείμενα
Μεγαλύτερη όσο πιο όμοια είναι τα αντικείμενα μεταξύ τους
Συχνά τιμές στο [0, 1]

Μη Ομοιότητα (dissimilarity)
Μια αριθμητική μέτρηση για το πόσο διαφορετικά είναι δυο αντικείμενα
Μικρότερη όσο ποιο όμοια είναι τα αντικείμενα μεταξύ τους
Η ελάχιστη τιμή είναι συνήθως 0 (όταν τα ίδια), αλλά το πάνω όρο διαφέρει

Γειτονικότητα (Proximity) αναφέρεται είτε στην ομοιότητα είτε στην μη ομοιότητα

Κριτήρια Ομοιότητας -Απόσταση

Η ομοιότητα-μη ομοιότητα μεταξύ δύο αντικειμένων μετριέται συνήθως βάση μιας συνάρτησης απόστασης ανάμεσα στα αντικείμενα

Εξαρτάται από το είδος των δεδομένων, δηλαδή από το είδος των γνωρισμάτων τους

Κριτήρια Ομοιότητας

Συναρτήσεις απόστασης (distance functions)

Συχνές ιδιότητες:

1. $d(i, j) \geq 0$
2. $d(i, i) = 0$ (ανακλαστική)
3. $d(i, j) = d(j, i)$ (συμμετρική)
4. $d(i, j) \leq d(i, h) + d(h, j)$ (τριγωνική ανισότητα)

Όταν ισχύουν και οι 4, η συνάρτηση απόστασης ονομάζεται και **μετρική απόστασης (distance metric)**

Κριτήρια Ομοιότητας

Γνωστές ιδιότητες για την ομοιότητα:

$s(p, q) = 1$ (ή μέγιστη ομοιότητα) μόνο αν $p = q$.

$s(p, q) = s(q, p)$ για κάθε p και q. (Συμμετρία)

Μετασχηματισμοί

Αν θέλουμε να έχουμε τιμές ομοιότητας (απόστασης) στο [0, 1]
Αν οι τιμές μας στο [min_s, max_s] ([min_d, max_d])

Απλός μετασχηματισμός (x – min_s) / (max_s – min_s)

Τι γίνεται αν $[0, \infty]$
Χρήση μη γραμμικού μετασχηματισμού πχ d/1+d
Αλλά χάνεται πληροφορία μεγέθους (scale distortion)
Πχ 0, 0.5, 2, 10, 100, 1000
0, 0.33, 0.67, 0.90, 0.99, 0.999

Μετασχηματισμοί

Από ομοιότητα -> απόσταση
Γενικά μπορούμε να χρησιμοποιήσουμε οποιαδήποτε μονότονα φθίνουσα συνάρτηση για τον μετασχηματισμό

$d = 1 - s, -s$
$s = 1 - d$

$s = 1/1+d, e^{-d}, 1 - (d - min\_d) / (max\_d - min\_d)$

Πχ απόσταση	0,	1,	10,	100
	0	-1	-10	-100
	1	0.5	0.09	0.01
	1	0.37	0.00	0.00
	1	0.99	0.00	0.00

Πίνακας Απόστασης

Συχνά, αφού υπολογιστεί η απόσταση, χρησιμοποιούμε όχι τα αρχικά αντικείμενα αλλά έναν πίνακα με τις αποστάσεις

Γνωρίσματα (διαστάσεις)

Πίνακας δεδομένων
(two modes-οι γραμμές και οι στήλες αφορούν διαφορετικές οντότητες)
n-σημεία διάστασης p

Σημεία (δεδομένα)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Πίνακας Απόστασης
(ανομοιότητας) – Πίνακας Γειτονικότητας – Contingency Matrix

n x n πίνακας

One mode

Αν συμμετρική,

d(i, j) = d(j, i)

Σημεία

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Σημεία

Κριτήρια Ομοιότητας -Απόσταση

Πως ορίζεται η απόσταση ανάμεσα σε πολύ-διάστατα δεδομένα;

Εξαρτάται από τις τιμές των γνωρισμάτων

Ας δούμε πρώτα 1 μεταβλητή

Κριτήρια Ομοιότητας –Απόσταση

Έστω p και *q* γνωρίσματα δυο αντικειμένων

Attribute Type	Dissimilarity (Απόσταση (μη ομοιότητα))	Similarity (Ομοιότητα)
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$ s = 1 - d
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ n διαφορετικές τιμές (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d,\ s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Ομοιότητα και απόσταση για απλές μεταβλητές

Κριτήρια Ομοιότητας -Απόσταση

Τι γίνεται αν έχουμε παραπάνω από ένα γνωρίσματα;

Γενικά για αριθμητικά δεδομένα (αλλά μπορεί να επεκταθεί)

Ορισμός Απόστασης

Έστω δυο μεταβλητές i και j με n γνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

Ο πιο συνηθισμένος τρόπος – Ευκλείδεια απόσταση:

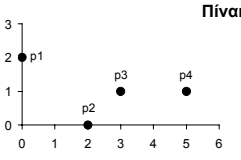
$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{in} - x_{jn}|^2)}$$

Είναι μετρική απόσταση

Ορισμός Απόστασης

Παράδειγμα

Πίνακας Δεδομένων



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Πίνακας Απόστασης

Ορισμός Απόστασης

Έστω δυο μεταβλητές i και j με n γνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

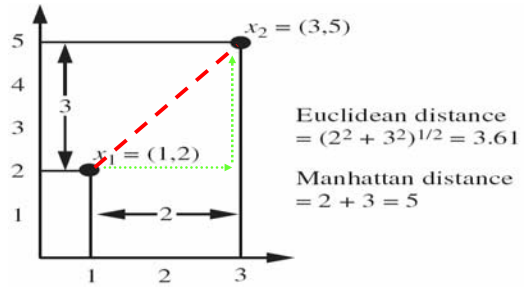
Manhattan ή city-block

$$L_1(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$$

Είναι μετρική απόστασης

Ορισμός Απόστασης

Παράδειγμα



Ορισμός Απόστασης

Έστω δυο μεταβλητές i και j με n γνωρίσματα x_{ik} και x_{jk} $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

Minkowski:

$$L_p(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p \right)^{1/p}$$

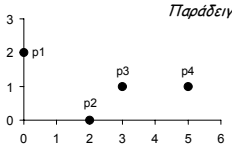
Είναι μετρική απόστασης

Ορισμός Απόστασης

- $r = 1$. City block (Manhattan, taxicab, L_1 norm).
  - Hamming distance, όταν δυαδικά διανύσματα = αριθμός bits που διαφέρουν
- $r = 2$. Ευκλείδεια απόσταση
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) απόσταση.
  - Η μέγιστη απόσταση μεταξύ οποιουδήποτε γνωρίσματος (διάστασης) των δυο διανυσμάτων

Ορισμός Απόστασης

Παράδειγμα



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Πίνακες Απόστασης

Ορισμός Απόστασης

Συχνά,

Βάρη πχ για Ευκλείδεια απόσταση:

$$d(i,j) = \sqrt{(w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + \ldots + w_n |x_{in} - x_{jn}|^2)}$$

Ορισμός Απόστασης

Μεταβλητές Διαστήματος

Standardization

Μέση απόλυτη απόκλιση (mean absolute deviation)

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

όπου

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

z-score

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Ορισμός Απόστασης

Δυαδικές Μεταβλητές

d(i, j) = 1 αν i = j και 0 αλλιώς

Συχνά δεδομένα με μόνο δυαδικά γνωρίσματα (δυαδικά διανύσματα)

Συμμετρικές (τιμές 0 και 1 έχουν την ίδια σημασία)
Invariant ομοιότητα

Μη συμμετρικές (η συμφωνία στο 1 πιο σημαντική – πχ όταν το 1 σηματοδοτεί την ύπαρξη κάποιας ασθένειας)
Non-invariant (Jaccard)

Ορισμός Απόστασης

Μεταξύ δύο αντικειμένων i και j με δυαδικά γνωρίσματα

M_{01} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 0 και το j έχει 1
M_{10} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 1 και το j έχει 0
M_{00} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 0 και το j έχει 0
M_{11} = ο αριθμός των γνωρισμάτων που το i έχει τιμή 1 και το j έχει 1

ΟΜΟΙΟΤΗΤΑ

Απλό ταίριασμα – συμμετρικές μεταβλητές

SMC = αριθμός ταιριασμάτων / αριθμός γνωρισμάτων
= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = αριθμός 11 ταιριασμάτων / αριθμό μη μηδενικών γνωρισμάτων
= $(M_{11}) / (M_{01} + M_{10} + M_{11})$

Συντελεστής Jaccard – Jaccard Coefficient – μη συμμετρικές μεταβλητές (διαφορετική σημασία στην τιμή 1 και στην τιμή 0)

Ορισμός Απόστασης

Παράδειγμα

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 0 1 0 0 1

M_{01} = 2
M_{10} = 1
M_{00} = 7
M_{11} = 0

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = $(M_{11}) / (M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

J = #1(p BAND q) / #1(p BOR Q)

Ορισμός Απόστασης

Contigency πίνακας για δυαδικά δεδομένα

Αντικείμενο *i* \ Αντικείμενο *j*	1	0
1	M_{11}	M_{10}
0	M_{01}	M_{00}

Μέτρηση απόστασης για συμμετρικές δυαδικές μεταβλητές
1 – συμμετρική-ομοιότητα

$$d(i, j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

Μέτρηση απόστασης για μη συμμετρικές δυαδικές μεταβλητές

$$d(i, j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

Jaccard coefficient

$$sim_{Jaccard}(i, j) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

Ορισμός Απόστασης

Παράδειγμα

τα γνωρίσματα μη συμμετρικά
Έστω Y-P να αντιστοιχούν στο 1 και το N στο 0

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Ορισμός Απόστασης

Κατηγορικές Μεταβλητές χωρίς Διάταξη (nominal)

Γενίκευση των δυαδικών μεταβλητών (γνωρισμάτων) όπου μπορούν να πάρουν παραπάνω από 2 τιμές, πχ κόκκινο, πράσινο, κίτρινο

1η Μέθοδος: Απλό ταίριασμα
m: # ταιριάσματα, *p*: συνολικός # μεταβλητών

$$d(i, j) = \frac{p - m}{p}$$

2η Μέθοδος: Χρήση πολλών δυαδικών μεταβλητών
Μία για κάθε μία από τις μ τιμές

Ορισμός Απόστασης

Ομοιότητα συνημίτονου (cosine similarity)

- Αν d_1 and d_2 είναι διανύσματα κειμένου
  $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,
  όπου • εσωτερικό γινόμενο $||d||$ το μήκος του d.

- Παράδειγμα:

  d_1 = 3 2 0 5 0 0 0 2 0 0
  d_2 = 1 0 0 0 0 0 0 1 0 2

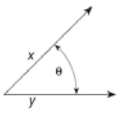
Θέλουμε μια απόσταση που να αγνοεί τα 0 (όπως η Jaccard) αλλά να δουλεύει και για μη δυαδικά δεδομένα

Επίσης, αγνοεί το *μήκος* των διανυσμάτων

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481
$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245
$\cos(d_1, d_2)$ = .3150

Ορισμός Απόστασης

Ομοιότητα συνημίτονου (cosine similarity)



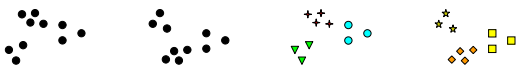
Γεωμετρική ερμηνεία

Ομοιότητα 1, όταν η γωνία 0 – που σημαίνει ότι τα x και y ίδια (αν εξαιρέσεις το μήκος τους)

Ομοιότητα 0, όταν η γωνία 90 (κανένα κοινό όρο)

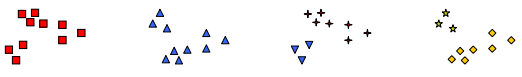
Είδη Συσταδοποίησης

Ασάφεια



Πόσες Ομάδες?

6 ομάδες



2 ομάδες

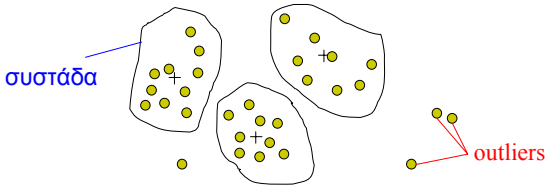
4 ομάδες

Γενικές Απαιτήσεις

- Scalability – στον αριθμό σημείων και διαστάσεων
- Να υποστηρίζει διαφορετικούς τύπους δεδομένων
- Να υποστηρίζει συστάδες με διαφορετικά σχήματα (συνήθως, «σφαίρες»)
- Να είναι εύκολο να δώσουμε τιμές στις παραμέτρους εισόδου (αριθμό συστάδων, μέγεθος κλπ)
- Να μην εξαρτάται από τη *σειρά* επεξεργασίας των σημείων εισόδου

Γενικές Απαιτήσεις

Αντιμετώπιση θορύβου και outliers



Outlier (ακραίο σημείο) τιμές που είναι εξαιρέσεις ως προς τα συνηθισμένες ή αναμενόμενες τιμές

Είδη συσταδοποίησης

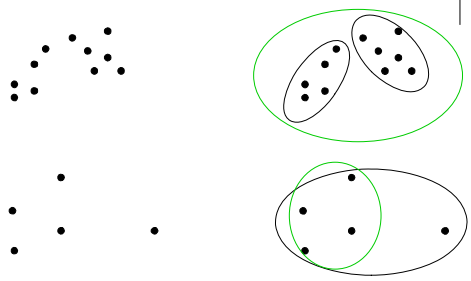
Μια συσταδοποίηση είναι ένα σύνολο από συστάδες

Βασική διάκριση ανάμεσα στο *ιεραρχικό* (*hierarchical*) και *διαχωριστικό* (*partitional*) σύνολο από ομάδες

Διαχωριστική Συσταδοποίηση (Partitional Clustering)
Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα - non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

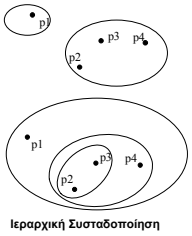
Ιεραρχική Συσταδοποίηση (Hierarchical clustering
Ένα σύνολο από *εμφωλευμένες* (nested) ομάδες
Επιτρέπουμε σε μια συστάδα να έχει υποσυστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Διαχωριστική και Ιεραρχική Συσταδοποίηση



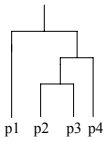
Αρχικά Σημεία

Διαχωριστική και Ιεραρχική Συσταδοποίηση



Ιεραρχική Συσταδοποίηση

Διαχωριστική Συσταδοποίηση



Παραδοσιακό Δένδρο-γράμμα (Dendrogram)

- Φύλλα: απλά σημεία ή απλές συστάδες
- Ως ακολουθία διαχωριστικών
- Να «κόψουμε» το δέντρο

Άλλες διακρίσεις μεταξύ συνόλων συστάδων

Επικαλυπτόμενο ή όχι
Ένα σημείο ανήκει σε περισσότερες από μια συστάδες(πχ οριακά σημεία)

Ασαφή συσταδοποίηση
Στην ασαφή συσταδοποίηση ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του 0 και του 1
Συχνά τα βάρη για κάθε σημείο έχουν άθροισμα 1
Η πιθανοτική συσταδοποίηση έχει παρόμοια χαρακτηριστικά

Μερική - Πλήρης
Σε ορισμένες περιπτώσεις θέλουμε να ομαδοποιήσουμε μόνο κάποια από τα δεδομένα (άλλα θόρυβος, ή μη ενδιαφέρουσα πληροφορία)

Ετερογενή - Ομογενή
Συστάδες με πολύ διαφορετικά μεγέθη, σχήματα και πυκνότητες (densities)

Αλγόριθμοι Συσταδοποίησης

Θα δούμε ανάμεσα σε άλλους τους:

- K-means και παραλλαγές
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την Πυκνότητα (DBSCAN)
- BIRCH (δεδομένα στο δίσκο!)