



Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική  
Σχολή Θετικών Επιστημών  
Πανεπιστήμιο Θεσσαλίας

## ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

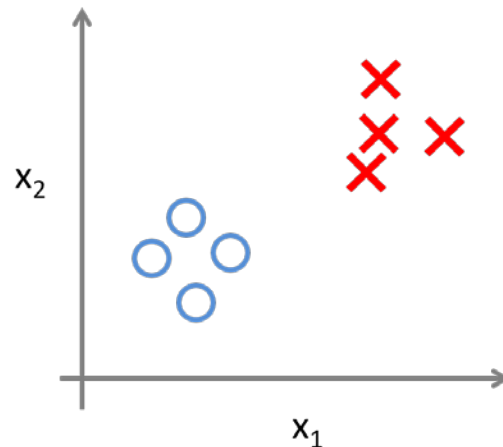
### Αξιολόγηση Κατηγοριοποίησης & Ομαδοποίησης

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD  
Adjunct Lecturer

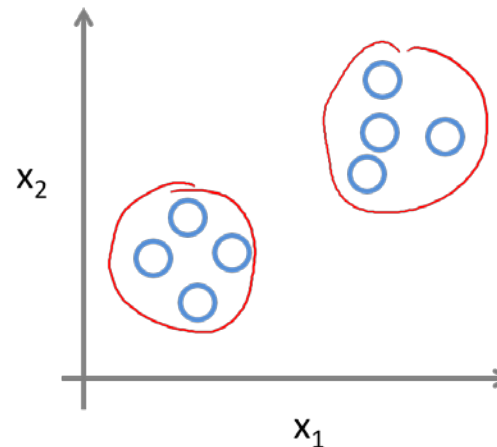
# Βασικές Κατηγορίες Αναγνώρισης Προτύπων

- Κατηγοριοποίηση:
  - Με επίβλεψη (Supervised) μάθηση
  - Γνωστός αριθμός κλάσεων (κατηγοριών)
  - Διαθέσιμα αντικείμενα για τα οποία είναι γνωστή η κλάση στην οποία ανήκουν.
- Ομαδοποίηση
  - Χωρίς επίβλεψη (Unsupervised) μάθηση
  - Άγνωστος αριθμός κλάσεων (γενικά).
  - Διαθέσιμα αντικείμενα για τα οποία δεν είναι γνωστή οποιαδήποτε πληροφορία σχετική με κλάση.

Supervised Learning



Unsupervised Learning



## Μέθοδοι Αξιολόγησης Αποτελεσμάτων

- Διάφορες τεχνικές αξιολόγησης των αποτελεσμάτων της αναγνώρισης προτύπων έχουν προταθεί με σκοπό να αποδώσουν με κάποιες μετρικές το πόσο αξιόπιστα είναι τα αποτελέσματα μας
- Υπάρχουν πολλοί τρόποι αξιολόγησης ενός αλγόριθμου, ενώ υπάρχουν επίσης πολλές μετρικές

## confusion πίνακας

- › Ένας confusion πίνακας, όπως φαίνεται και παρακάτω, περιέχει πληροφορίες σχετικά με πραγματικά και προβλέψιμα αποτελέσματα, δοσμένα από έναν ταξινομητή.

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN

- › Όπου στον άνωθεν πίνακα, ισχύουν τα εξής:
- › TP: αριθμός σωστών ταξινομήσεων των θετικών παραδειγμάτων (true positive)
- › TN: αριθμός σωστών ταξινομήσεων των αρνητικών παραδειγμάτων (true negative)
- › FP: αριθμός λανθασμένων ταξινομήσεων των θετικών παραδειγμάτων (false positive)
- › FN: αριθμός λανθασμένων ταξινομήσεων των αρνητικών παραδειγμάτων (false negative)

# confusion πίνακας

		predicted condition	
total population		prediction positive	prediction negative
true condition	condition positive	<b>True Positive (TP)</b>	<b>False Negative (FN)</b> (type II error)
	condition negative	<b>False Positive (FP)</b> (Type I error)	<b>True Negative (TN)</b>

$\pi$

**True positive**



**True negative**



**False positive  
(Type I error)**



**False negative  
(Type II error)**



# confusion πίνακας

		predicted condition			
total population		prediction positive	prediction negative	Prevalence = $\frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	True Positive (TP)	False Negative (FN) (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection = $\frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate = $\frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	False Positive (FP) (Type I error)	True Negative (TN)	False Positive Rate (FPR), Fall-out, Probability of False Alarm = $\frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
Accuracy = $\frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$		Positive Predictive Value (PPV), Precision = $\frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) = $\frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False Discovery Rate (FDR) = $\frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) = $\frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

## Συνολική ακρίβεια (Accuracy)

- › Η συνολική ακρίβεια του μοντέλου υπολογίζεται ως ο λόγος των σωστά προβλεπόμενων παρατηρήσεων προς όλες τις παρατηρήσεις.
- › Άρα έχουμε για το μέτρο accuracy τον τύπο

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Instances}} \quad \text{ή} \quad \text{Acc} = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}$$



## Ακρίβεια θετικής πρόβλεψης (precision):

- › Ακρίβεια θετικής πρόβλεψης ή θετική διαγνωστική αξία
- › Στην ανάκτηση πληροφορίας ο όρος precision αναφέρεται στον αριθμό των σωστά προβλεπόμενων positive παρατηρήσεων προς όλες τις παρατηρήσεις που θεωρήθηκαν σαν positive στα αποτελέσματα.
- › Για μια ασθένεια A
  - η πιθανότητα ότι το άτομο έχει το 'πρόβλημα' εάν η δοκιμασία είναι θετική

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad \text{ή} \quad \text{Pr} = \frac{Tp}{Tp + Fp}$$

## Πραγματικό ποσοστό θετικών (ευαισθησία)

- › Πραγματικό ποσοστό θετικών (True positive rate or Recall or Sensitivity):
- › Το μέτρο recall μετράει το ποσοστό από τις προβλεπόμενες positive παρατηρήσεις που είναι πραγματικές positive, δηλαδή:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad \text{ή} \quad \text{Rec} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}}$$

- › Για μια ασθένεια A
  - η αναλογία των ατόμων με το 'πρόβλημα' που η μέθοδος αναγνωρίζει ότι πάσχουν

## Πραγματικό ποσοστό αρνητικών (ειδικότητα)

- › Πραγματικό ποσοστό αρνητικών (True Negative Rate or Specificity)
- › Αυτό το μέτρο μετράει το ποσοστό των πραγματικών negative παρατηρήσεων που χαρακτηρίστηκαν ως negative:

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \text{ ή } Sp = \frac{Tn}{Tn + Fp}$$

- › Για μια ασθένεια A
  - η αναλογία των ατόμων χωρίς το 'πρόβλημα' που η μέθοδος αναγνωρίζει ότι δεν πάσχουν

## F-measure ή F-score

- › Το μέτρο F-score αποτελεί τον αρμονικό μέσο των Precision και Recall (ή και των Specificity και Recall) με τιμές ανάμεσα σε 0 και 1 (1 για την τέλεια ακρίβεια και 0 για την χειρότερη).
- › Ο τύπος του F-score είναι ο εξής:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

## F-measure ή F-score

- Για  $\beta=1$  έχουμε:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- όπου η παράμετρος  $\beta$  ορίζει τι θεωρούμε πιο σημαντικό στις μετρήσεις από τα Precision, Recall.
- Εφόσον θεωρηθούν εξίσου σημαντικά, το  $\beta$  ισούται με 1 και βρίσκουμε το F1-score .
- Σε άλλες περιπτώσεις μπορεί να είναι προτιμότερο να θεωρήσουμε άλλα F-scores, όπως τα F2-score ή F0.5-score , τα οποία δίνουν περισσότερο βάρος στα μέτρα Precision και Recall αντίστοιχα.

# Άσκηση

- › Στον παρακάτω πίνακα είναι τα αποτελέσματα κατά την εφαρμογή του Αλγορίθμου Κατηγοριοποίησης X σε νέα πρότυπα.
- › Βρείτε τι μετρικές «Ακρίβεια θετικής πρόβλεψης (Precision)», «Πραγματικό ποσοστό θετικών» (True positive rate or Recall or Sensitivity) και σχολιάστε τις τιμές τους.

Πραγματική Κλάση	Πρόβλεψη Αλγορίθμου
Θετικό	Θετικό
Θετικό	Αρνητικό
Θετικό	Αρνητικό
Αρνητικό	Αρνητικό
Θετικό	Θετικό
Αρνητικό	Θετικό
Αρνητικό	Αρνητικό
Θετικό	Θετικό
Θετικό	Θετικό
Αρνητικό	Θετικό
Αρνητικό	Αρνητικό
Αρνητικό	Αρνητικό
Αρνητικό	Αρνητικό
Θετικό	Αρνητικό

# Ομαδοποίηση - Μέτρα Εγγύτητας

# Εγκυρότητα και αποτίμηση συσταδοποίησης

- Η αποτίμηση συσταδοποίησης επιδιώκει να αποτιμήσει την καταλληλότητα ή ποιότητα της συσταδοποίησης· η σταθερότητα συσταδοποίησης έχει ως στόχο
  - να κατανοήσει την ευαισθησία που εμφανίζει το αποτέλεσμα της συσταδοποίησης σε διάφορες αλγοριθμικές παραμέτρους, π.χ. το πλήθος των συστάδων· και
  - η τάση συσταδοποίησης αξιολογεί το κατά πόσο θα έπρεπε εξ αρχής να εφαρμοστεί η συσταδοποίηση, δηλαδή το αν τα δεδομένα εμφανίζουν οποιαδήποτε εγγενή δομή ομαδοποίησης.
- Τα μέτρα της εγκυρότητας μπορούν να χωριστούν σε τρεις κύριους τύπους:
  - Εξωτερικά: Τα εξωτερικά μέτρα εγκυρότητας χρησιμοποιούν κριτήρια που δεν είναι εγγενή για το σύνολο δεδομένων, π.χ. ετικέτες κατηγορίας.
  - Εσωτερικά: Τα εσωτερικά μέτρα εγκυρότητας στηρίζονται σε κριτήρια που προκύπτουν από τα ίδια τα δεδομένα, π.χ. μετρικές απόστασης εντός της ίδιας συστάδας ή μεταξύ διαφορετικών συστάδων.
  - Σχετικά: Τα σχετικά μέτρα εγκυρότητας επιδιώκουν να συγκρίνουν ευθέως διαφορετικές συσταδοποιήσεις, συνήθως εκείνες που προκύπτουν από διαφορετικές ρυθμίσεις των παραμέτρων του ίδιου αλγορίθμου.



# Εξωτερικά μέτρα

- Τα εξωτερικά μέτρα υποθέτουν ότι είναι γνωστή εκ των προτέρων η σωστή συσταδοποίηση (δηλαδή εκείνη που αντιστοιχεί στη δεδομένη αλήθεια), η οποία και χρησιμοποιείται για την αποτίμηση μιας δεδομένης συσταδοποίησης.
- Έστω ότι  $D = \{X_i\}_{i=1}^n$  είναι ένα σύνολο δεδομένων που αποτελείται από  $n$  σημεία σε έναν  $d$ -διάστατο χώρο, το οποίο έχει διαμεριστεί σε  $k$  συστάδες.
- Έστω ότι συμβολίζουμε με  $y_i \in \{1, 2, \dots, k\}$  τις πληροφορίες συμμετοχής στις συστάδες (ή των ετικετών των συστάδων) που αντιστοιχούν στη δεδομένη αλήθεια για κάθε σημείο.
- Η συσταδοποίηση που αντιστοιχεί στη δεδομένη αλήθεια ορίζεται ως  $T = \{T_1, T_2, \dots, T_k\}$ , όπου η συστάδα  $T_j$  αποτελείται από όλα τα σημεία με ετικέτα  $j$ , δηλαδή  $T_j = \{x_i \in D | y_i = j\}$ .
- Για λόγους σαφήνειας, θα αναφερόμαστε στη συσταδοποίηση  $T$  ως τον διαμερισμό που αντιστοιχεί στη δεδομένη αλήθεια, και σε κάθε συστάδα  $T_i$  ως διαμέριση.

# Εξωτερικά μέτρα

- Τα εξωτερικά μέτρα αποτίμησης προσπαθούν να αποτυπώσουν τον βαθμό στον οποίο τα σημεία από την ίδια διαμέριση εμφανίζονται στην ίδια συστάδα, καθώς και τον βαθμό στον οποίο τα σημεία από διαφορετικές διαμερίσεις ομαδοποιούνται σε διαφορετικές συστάδες.
- Όλα τα εξωτερικά μέτρα στηρίζονται στον πίνακα συνάφειας  $N$ , διαστάσεων  $r \times k$ , ο οποίος επάγεται από μια συσταδοποίηση  $C$  και τον διαμερισμό  $T$  που αντιστοιχεί στη δεδομένη αλήθεια. Ο πίνακας συνάφειας ορίζεται ως εξής:

$$N(i, j) = n_{ij} = |C_i \cap T_j|$$

- Η καταμέτρηση  $n_{ij}$  αναπαριστά το πλήθος των σημείων που ανήκουν τόσο στη συστάδα  $C_i$  όσο και στη διαμέριση  $T_j$  της δεδομένης αλήθειας.
- Έστω ότι το  $n_i = |C_i|$  είναι το πλήθος των σημείων που ανήκουν στη συστάδα  $C_i$ , και το  $m_j = |T_j|$  είναι το πλήθος των σημείων που ανήκουν στη διαμέριση  $T_j$ .
- Ο πίνακας συνάφειας μπορεί να υπολογιστεί από τον διαμερισμό  $T$  και τη συσταδοποίηση  $C$ :
  - Για κάθε σημείο  $x_i \in D$  εξετάζεται η ετικέτα  $y_i$  της διαμέρισης και η ετικέτα της συστάδας

# Μέτρα που βασίζονται στο ταίριασμα: Καθαρότητα

- Η καθαρότητα ποσοτικοποιεί τον βαθμό στον οποίο μια συστάδα  $C_i$  περιέχει οντότητες από μία μόνο διαμέριση:

$$purity_i = \frac{1}{n_i} \max_k^{j=1} \{n_{ij}\}$$

- Η καθαρότητα της συσταδοποίησης  $C$  ορίζεται ως το σταθμισμένο άθροισμα των τιμών καθαρότητας ανά συστάδα:

$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_k^{j=1} \{n_{ij}\}$$

- όπου ο λόγος  $n_i/n$  αντιπροσωπεύει το ποσοστό των σημείων που ανήκουν στη συστάδα  $C_i$ .

# Μέτρα που βασίζονται στο ταίριασμα: Μέγιστο ταίριασμα

- Το μέτρο του μέγιστου ταιριάσματος επιλέγει εκείνη την αντιστοίχιση μεταξύ συστάδων και διαμερίσεων για την οποία μεγιστοποιείται το άθροισμα του πλήθους των κοινών σημείων ( $n_{ij}$ ), με την προϋπόθεση ότι μόνο μία συστάδα μπορεί να ταιριάζει με μια καθορισμένη διαμέριση.
- Έστω ότι  $G$  είναι ένα διχοτομήσιμο γράφημα για το σύνολο κορυφών  $V = C \cup T$ , και έστω ότι το σύνολο ακμών είναι  $E = \{(C_i, T_j)\}$  με βάρος  $w(C_i, T_j) = n_{ij}$ .
- Ένα ταίριασμα  $M$  στο γράφημα  $G$  είναι υποσύνολο του  $E$ , τέτοιο ώστε οι ακμές του  $M$  να είναι μη γειτονικές ανά ζεύγη, δηλαδή να μην έχουν κοινή κορυφή.
- Το ταίριασμα μέγιστου βάρους στο γράφημα  $G$  ορίζεται ως εξής:

$$match = \arg \max_M \left\{ \frac{w(M)}{n} \right\}$$

- όπου το βάρος ενός ταιριάσματος  $M$  είναι απλώς το άθροισμα των βαρών όλων των ακμών του  $M$ , και δίνεται από τη σχέση

$$w(M) = \sum_{e \in M} w(e)$$

# Μέτρα που βασίζονται στο ταιρίασμα: F-μέτρο

- Για μια συστάδα  $C_i$ , έστω ότι συμβολίζουμε με  $j_i$  τη διαμέριση που περιέχει το μέγιστο πλήθος σημείων από τη  $C_i$ , δηλαδή

$$j_i = \max_{j=1}^k \{n_{ij}\}$$

- Η ακρίβεια μιας συστάδας  $C_i$  είναι ίδια με την καθαρότητά της:

$$prec_i = \frac{1}{n_i} \max_k \{n_{ij}\} = \frac{n_{ij_i}}{n_i}$$

- Η ανάκληση της συστάδας  $C_i$  ορίζεται ως

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

- όπου  $m_{j_i} = |T_{j_i}|$ .

# Μέτρα που βασίζονται στο ταιρίασμα: F-μέτρο

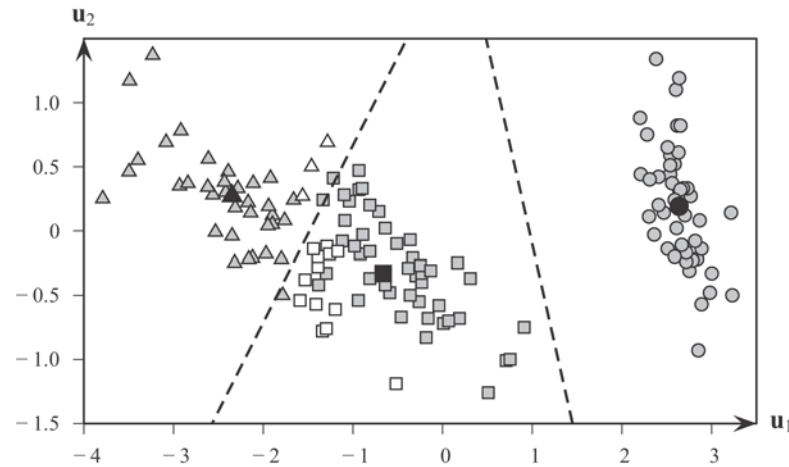
- Το F-μέτρο είναι ο αρμονικός μέσος των τιμών ακρίβειας και ανάκλησης για κάθε συστάδα  $C_i$

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2 n_{ij_i}}{n_i + m_{j_i}}$$

- Το F-μέτρο για τη συσταδοποίηση C είναι ο μέσος των τιμών του F-μέτρου ανά συστάδα

$$F = \frac{1}{r} \sum_{i=1}^r F_i$$

# Αλγόριθμος K-means



(α) Αλγόριθμος K μέσων: καλή

– Πίνακας συνάφειας:

	iris-setosa	iris-versicolor	iris-virginica	
	$T_1$	$T_2$	$T_3$	$n_i$
$C_1$ (τετράγωνα)	0	47	14	61
$C_2$ (κύκλοι)	50	0	0	50
$C_3$ (τρίγωνα)	0	3	36	39
$m_j$	50	50	50	$n = 150$

purity = ???, match = ???, F = ???

# Αλγόριθμος K-means

	iris-setosa	iris-versicolor	iris-virginica	
	$T_1$	$T_2$	$T_3$	$n_i$
$C_1$ (τετράγωνο)	30	0	0	30
$C_2$ (κύκλοι)	20	4	0	24
$C_3$ (τρίγωνο)	0	46	50	96
$m_j$	50	50	50	$n = 150$

F1

precision(c3) = 50/96=

Recal(c3) = 50/50 =

$$\text{purity} = (30 + 20 + 50)/150, \text{ match} = (30+4+50)/150 \text{ F} = (\text{Fc1}+\text{Fc2}+\text{Fc3})/3$$



## Μέτρα ανά ζεύγη

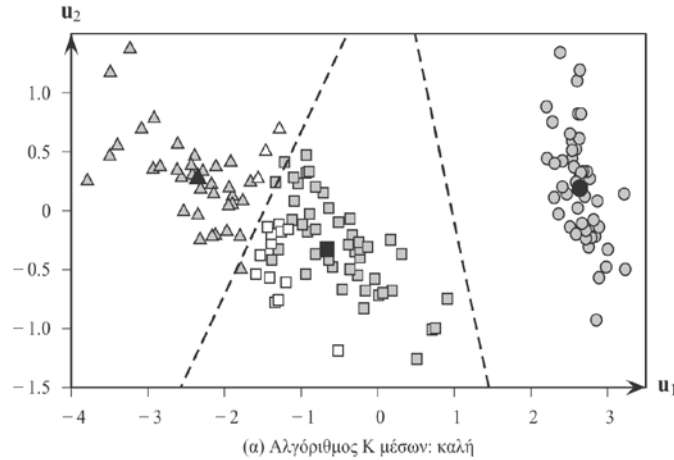
- Δίνεται η συσταδοποίηση  $C$  και ο διαμερισμός  $T$  που αντιστοιχεί στη δεδομένη αλήθεια· έστω ότι  $x_i, x_j \in D$  είναι δύο οποιαδήποτε σημεία, με  $i \neq j$ .
- Αν τόσο το  $x_i$  όσο και το  $x_j$  ανήκουν στην ίδια συστάδα, περιγράφουμε αυτή την κατάσταση με τον όρο θετικό συμβάν
- Αν δεν ανήκουν στην ίδια συστάδα, χρησιμοποιούμε τον όρο αρνητικό συμβάν.
- Ανάλογα με το αν οι ετικέτες των συστάδων συμφωνούν με τις ετικέτες των διαμερίσεων, υπάρχουν τέσσερα ενδεχόμενα που πρέπει να ληφθούν υπόψη:

## Μέτρα ανά ζεύγη

- Αληθώς θετικά (TP): Τα σημεία  $x_i$  και  $x_j$  ανήκουν στην ίδια διαμέριση του διαμερισμού  $T$ , αλλά και στην ίδια συστάδα της συσταδοποίησης  $C$ .
- Ψευδώς αρνητικά (FN): Τα σημεία  $x_i$  και  $x_j$  ανήκουν στην ίδια διαμέριση του  $T$ , αλλά όχι και στην ίδια συστάδα της  $C$ .
- Ψευδώς θετικά (FP): Τα σημεία  $x_i$  και  $x_j$  δεν ανήκουν στην ίδια διαμέριση του  $T$ , αλλά ανήκουν στην ίδια συστάδα της  $C$ .
- Αληθώς αρνητικά (TN): Τα σημεία  $x_i$  και  $x_j$  δεν ανήκουν ούτε στην ίδια διαμέριση του  $T$ , ούτε στην ίδια συστάδα της  $C$ .
- Επειδή υπάρχουν  $N = \binom{n}{2} = \frac{n(n-1)}{2}$  ζεύγη σημείων, προκύπτει η ακόλουθη ισότητα

$$N = TP + FN + FP + TN$$

# Αλγόριθμος K μέσων:



Πίνακας συνάφειας:

	iris-setosa	iris-versicolor	iris-virginica
$C_1$	$T_1$ 0	$T_2$ 47	$T_3$ 14
$C_2$	50	0	0
$C_3$	0	3	36

- TP = ???, FN = ???, FP = ???, TN = ???

## Μέτρα ανά ζεύγη: Συντελεστής Jaccard, στατιστικό Rand, μέτρο FM

- Συντελεστής Jaccard: Μετρά το ποσοστό των αληθώς θετικών ζευγών (σημείων), αφού όμως πρώτα αγνοήσει τα αληθώς αρνητικά ζεύγη.

$$Jaccard = \frac{TP}{TP + FN + FP}$$

- Στατιστικό Rand: Μετρά το ποσοστό των αληθώς θετικών και αληθώς αρνητικών ζευγών για όλα τα ζεύγη σημείων.

$$Rand = \frac{TP + TN}{N}$$

- Μέτρο των Fowlkes-Mallows: Ορίζουμε τη συνολική ακρίβεια ανά ζεύγη και ανάκληση ανά ζεύγη για μια συσταδοποίηση C, όπως φαίνεται παρακάτω

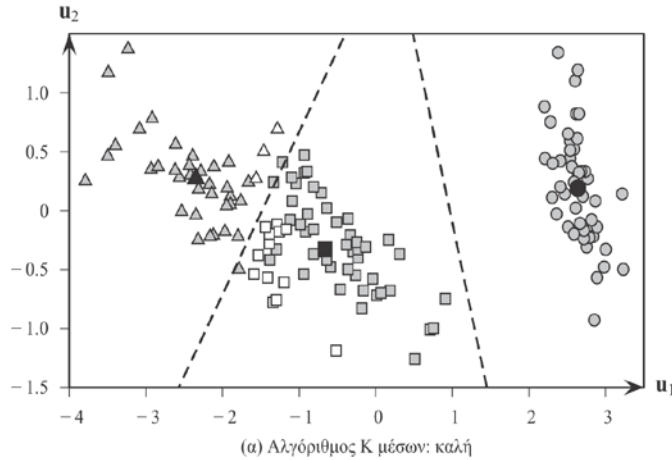
$$prec = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- Το μέτρο των Fowlkes-Mallows (FM) ορίζεται ως ο γεωμετρικός μέσος της ακρίβειας ανά ζεύγη και της ανάκλησης ανά ζεύγη

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

# Αλγόριθμος K μέσων:

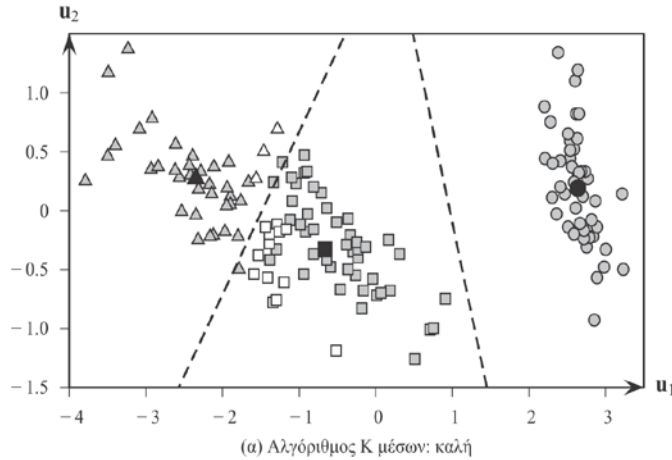


Πίνακας συνάφειας:

	iris-setosa	iris-versicolor	iris-virginica
$C_1$	$T_1$ 0	$T_2$ 47	$T_3$ 14
$C_2$	50	0	0
$C_3$	0	3	36

- TP = 3030, FN = 645, FP = 766, TN = 6734
- Jaccard = ???
- Rand = ???
- FM = ???

## Αλγόριθμος K μέσων:



Πίνακας συνάφειας:

	iris-setosa	iris-versicolor	iris-virginica
$C_1$	$T_1$ 0	$T_2$ 47	$T_3$ 14
$C_2$	50	0	0
$C_3$	0	3	36

- $TP = 3030, FN = 645, FP = 766, TN = 6734$
- $Jaccard = 0,68$
- $Rand = 0,87$
- $FM = 0,81$