2/5/2017

Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων
Ακαδημαϊκό έτος 2016-17

Προπτυχιακό μάθημα: "Αναγνώριση Προτύπων"

2η Σειρά Ασκήσεων

(Ημερομηνία παράδοσης : έως Δευτέρα 22/5/2017)

Πρόβλημα : Ομαδοποίηση δεδομένων

Από την σελίδα του μαθήματος κατεβάστε τα αρχεία: 'cross.dat' και 'moonandsun.dat' που περιέχουν 2 πειραματικά σύνολα δεδομένα με 500 δεδομένα 2 διαστάσεων. Και στα δύο σύνολα, η πρώτη στήλη αναφέρεται στην πραγματική κατηγορία (true) ενώ οι υπόλοιπες 2 στήλες αντιστοιχούν στις τιμές των 2 χαρακτηριστικών των δεδομένων (συντεταγμένες δισδιάστατων σημείων). Στόχος είναι να πετύχετε την βέλτιστη ομαδοποίηση των δύο συνόλων. Κατασκευάστε τις δύο παρακάτω μεθόδους ομαδοποίησης:

- *Αλγόριθμος k-means* με Ευκλείδια απόσταση (αρχικοποίηση των K μέσων από τα δείγματα. Η τελική λύση προκύπτει από την καλύτερη λύση μεταξύ 10 επαναλήψεων της μεθόδου).
- Συνθετική Ιεραρχική Ομαδοποίηση (*Agglomerative Hierarchical Clustering*): χρησιμοποιήστε ως μέτρο την απόσταση των μέσων 2 ομάδων $D_{\text{means}}(C_i, C_j) = \|\mu_i - \mu_j\|^2$ για να κάνετε merge. Προσοχή η συγκεκριμένη μέθοδος εκτελείται **1 φορά** για οποιαδήποτε τιμή του *K* (πλήθος ομάδων) και ΔΕΝ χρειάζεται να τρέξει για διαφορετικές τιμές του K.

Και στους δύο αλγορίθμους, χρησιμοποιήστε τις τιμές *K = 2, 3 ή 4* ως πλήθος ομάδων.

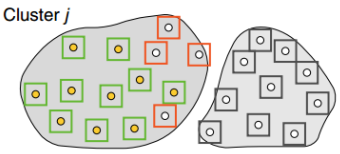
Για την αξιολόγηση του αποτελέσματος της ομαδοποίησης χρησιμοποιήσετε τα δύο παρακάτω μέτρα:

- ***Purity***: η κατηγορία κάθε ομάδας (c_j) καθορίζεται, μετά το τέλος της ομαδοποίησης, από την πλειοψηφούσα πραγματική κατηγορία (ω_k) μεταξύ των μελών της ομάδας. Τότε η ακρίβεια (*purity*) του παραπάνω καθορισμού υπολογίζεται μετρώντας το μέσο των σωστά ταξινομημένων σημείων. Δηλ.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_k \max_j |\omega_k \cap c_j|$$

  όπου N το σύνολο των δειγμάτων που έχετε στην διάθεσή σας.

- ***F-measure:***



		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Precision: $\frac{a}{a+b}$ Recall: $\frac{a}{a+c}$ F-measure: $F_\alpha = \frac{1+\alpha}{\frac{1}{precision} + \frac{\alpha}{recall}}$ $\begin{matrix}\alpha = 1 \\ \alpha \in (0;1) \\ \alpha > 1\end{matrix}$

Για κάθε cluster j, αφού καθορίσετε την πλειοψηφούσα κατηγορία ως κατηγορία cluster (όπως και στο προηγούμενο μέτρο), να βρείτε τα TP (true positive), FP (false positive) και FN (false negative) να βρείτε το F-measure, $\boldsymbol{F_a^{(j)}}$, για κάθε cluster χρησιμοποιώντας τιμή $\alpha=1$. Στο τέλος, η αξιολόγηση της μεθόδου clustering που εξετάζεται θα προκύπτει από το άθροισμα των F-measures για κάθε cluster.

$$Total\ \ F-measure = \sum_{j=1}^{K} F_1^{(j)}$$

Δώστε ένα σύντομο *report* με τον τρόπο κατασκευής των μεθόδων και τα αποτελέσματα των δοκιμών ανά περίπτωση. Να δοθεί επίσης και ο κώδικας.