Γιώργος Μπισιρόπουλος ΑΜ:2771

Αναγνώριση προτύπων - 2η Σειρά ασκήσεων

Files:

- kMeans.py
- spectrClustr.py
- agglomHierCl.py
- m_cl_loader: περιέχει βοηθητικές συναρτήσεις τις οποίες υλοποίησα εγώ, για το διάβασμα των δεδομένων από τα csv αρχεία
- m_cl_metrics : περιέχει βοηθητικές συναρτήσεις τις οποίες υλοποίησα εγώ, για τον υπολογισμό των: purity , f1 measure.

Για τα files:

Λόγω χρόνου για την υλοποίηση των υπολοίπων μεθόδων χρησιμοποίησα imports από την βιβλιοθήκη sklearn .

Επιπλέον χρειάστηκε να κάνω preproccessing τα δεδομένα κάνοντας χρήση της μεθόδου StandardScaler για να βελτιώσω την απόδοση.

Comparison:

Spambase Dataset				
Method	K	Purity	F1 score	Notes
K-means	2	0.605955	0.69300029	
K-means	4	0.770049	0.91428961	
K-means	8	0.851554	0.86647837	
Spectral Clust	2	-	-	Graph is not fully connected,
Spectral Clust	4	-	-	Graph is not fully connected,
Spectral Clust	8	-	-	Graph is not fully connected, spectral embedding
Agglom Hier	2	0.60595522	0.6927902	
Agglom Hier	4	0.81134535	-	

Agglom Hier	8	0.81134535	0.90382	

Occupancy Dataset				
Method	K	Purity	F1 score	Notes
K-means	2	0.9474395	-	
K-means	4	0.932334	0.952089	
K-means	8	0.973474	0.5679426	
Spectral Clust	2	0.787670	0.799062	
Spectral Clust	4	0.957263	-	
Spectral Clust	8	0.9501412	0.8697619	
Agglom Hier	2	0.9252118	-	
Agglom Hier	4	0.9252118	0.937870	
Agglom Hier	8	0.9252118	0.5460705	

Spectral Clustering ιδιοτιμές
K2: [1.97813837 6.12097755]
K4: [1.37261035 1.47745334 1.63381208 7.76031166]
K8: [1.105481 1.14173902 1.18004878 1.19516963 1.43075301 1.5047578 1.9259757 7.11671264]

<u>Επιλογή βέλτιστου αριθμού ομάδων με βάση τις τιμές των ιδιοτιμών</u>

- Ταξινομούμε τις ιδιοτιμές σε αύξουσα σειρά
- Υπολογίζουμε τις διαφορές διαδοχικών ιδιοτιμών
- Θέλουμε K ώστε όλες οι διαφορές Δλκ μέχρι τότε να είναι μικρές αλλά η επόμενη διαφορά να είναι μεγάλη

Γενικά comments:
Για K =4 είχαμε καλά αποτελέσματα βάση Purity και F1 score οπότε είχαμε σχετικά ομοιογενή clusters και είναι μια καλή τιμή για αριθμό cluster.

Από άποψη χρόνου σχεδόν όλες οι μέθοδοι χρειάστηκαν περίπου το ίδιο εκτός από την Spectral Clustering για το spam dataset, που λόγω των δεδομένων εισόδου εμφάνιζε το warning (Graph is not fully connected) και ακόμα και αρκετά λεπτά μετά (20+), δεν έβγαζε αποτέλεσμα.