HELLENIC MEDITERRANEAN UNIVERSITY

Αναγνώριση Προτύπων 2021 - 2022

Τεχνική Αναφορά

Εισαγωγή στο θέμα επεξεργασίας των Δεδομένων:

Αρχικά ο καρκίνος του μαστού είναι ο πιο συχνή μορφή καρκίνου που επηρεάζει τους ανθρώπους. Ευτυχώς είναι μια από τους πιο εύκολους σε διάγνωση καρκίνους από οποιαδήποτε μορφή καρκίνου σε παγκόσμιο επίπεδο .Κάθε χρόνο περίπου 4 εκατομμύρια γυναίκες στης Ηνωμένες πολιτείες Αμερικής διαγιγνώσκονται με καρκίνο του μαστού και το 90% από του πάσχοντες έχουν θετικά αποτελέσματα για την καταπολέμηση του καρκίνου με άμεσες και σύγχρονες θεραπείες ανάλογα κάθε περίπτωση ασθενή προφανώς .

Πληροφορίες Χαρακτηριστικού όγκου :

1. Δείγμα κωδικού αριθμού: αριθμός ταυτότητας
2. Πάχος συστάδας: 1 - 10
3. Ομοιομορφία μεγέθους συστάδας: 1 - 10
4. Ομοιομορφία σχήματος συστάδας: 1 - 10
5. Οριακή πρόσφυση: 1 - 10
6. Μέγεθος απλών επιθηλιακών κυττάρων: 1 - 10
7. Γυμνοί Πυρήνες: 1 - 10
8. Ήπια χρωματίνη: 1 - 10
9. Κανονικοί Πυρήνες: 1 - 10
10. Μιτώσεις: 1 - 10
11. Κατηγορία: (2 για καλοήθεις, 4 για κακοήθεις)

Όργανα – Εργαλεία :

- Python 3.8
- Anacoda Navigator (Anacoda3) – Jupiter Notebook

Αρχικά χρησιμοποιούμε τα δεδομένα του Dataset και βρίσκουμε τα στατιστικά δεδομένα του όπως το Mean ,Std ,Min και στην συνέχεια δημιουργούμε boxplots για να οπτικοποιησουμε τα δεδομένα μας .

Μετρήσεις / Data: Supervised Learning (Μάθηση με επίβλεψη)

Στην συνέχεια σπάμε τα δεδομένα μας και χρησιμοποιούμε ένα supervised learning για να βγάλουμε αποτελέσματα για το Dataset.Πραγματοποιούνται οι εξής διαχωρισμοί σε training seg και test set 90%/10% , 80%/20% , 70%/30%, 60%/40%,50%/50% , 40%/60% , 30%/70%, 20%/80%, 10%/90% .

LINEAR DISCRIMINANT ANALYSIS

	True Pos	True Neg	False Pos	False Neg	Precision	Recall	F1	Support	LDA Score
90%/10%	41	2	5	22	0.9	0.9	0.9	70	0.9
80%/20%	87	4	4	45	0.96	0.94	0.94	140	0.90714
70%/30%	134	5	6	65	0.95	0.95	0.95	210	0.9476
60%/40%	177	6	8	89	0.95	0.95	0.95	280	0.95
50%/50%	222	6	11	111	0.95	0.95	0.95	350	0.9514
40%/60%	273	6	12	129	0.96	0.95	0.96	420	0.9571
30%/70%	273	6	12	129	0.96	0.96	0.96	420	0.9571
20%/80%	355	12	13	180	0.96	0.96	0.96	560	0.95553
10%/90%	402	13	21	194	0.95	0.95	0.95	630	0.946

QUADRATIC DISCRIMINANT ANALYSIS

	True Pos	True Neg	False Pos	False Neg	Precision	Recall	F1	Support	QLA Score
90%/10%	41	2	5	22	0.9	0.9	0.9	70	0.9
80%/20%	81	10	3	46	0.91	0.91	0.91	140	0.9071
70%/30%	126	13	4	67	0.92	0.92	0.92	210	0.919
60%/40%	169	14	5	92	0.94	0.93	0.93	280	0.9321
50%/50%	214	14	6	116	0.94	0.94	0.94	350	0.9428
40%/60%	255	24	5	136	0.94	0.93	0.93	420	0.9309
30%/70%	255	24	5	136	0.94	0.93	0.93	420	0.9309
20%/80%	333	34	1	192	0.95	0.94	0.94	560	0.9375
10%/90%	384	31	0	215	0.96	0.95	0.95	630	0.9507

Τα αποτελέσματα για την Linear Discriminant Analysis σχετικά με το Quadratic Discriminant Analysis είναι ότι αν παρατηρήσουμε στο split test 80% training set με το 20% test set το Linear Discriminant Analysis είναι καλυτέρα εκπαιδευμένο το μοντέλο μιας και το Score τους είναι λίγο πιο μεγαλύτερο σε σχέση με το Quadratic Discriminant Analysis.Αυτό παρατηρείται και στους επομένους διαχωρισμούς (εκτός του 90/10 & 10/90 μιας και είναι ακραίες περιπτώσεις).

Προφανώς και να φαινόταν αυτό μιας και είναι διακριτές τιμές του dataset και συνήθως γίνεται με τέτοιου είδους αλγορίθμους το Training για να μας βγάλει ένα score και να δούμε αν το μοντέλο ήταν αποδοτικό.

Μετρήσεις / Data: Unsupervised Learning (Μάθηση χωρίς επίβλεψη)

Ένα καλό χαρακτηριστικό της ιεραρχικής ομαδοποίησης είναι ότι ο αλγόριθμος βρίσκει ποιες είναι οι πιο καλύτερος τρόπος ομαδοποίησης των δεδομένων χωρίς να το κάνουμε εμείς. Στην προκείμενη περίπτωση θα δούμε την διαφορά των δυο αλγορίθμων ομαδοποίησης K-Means και των Hierarchical clustering .

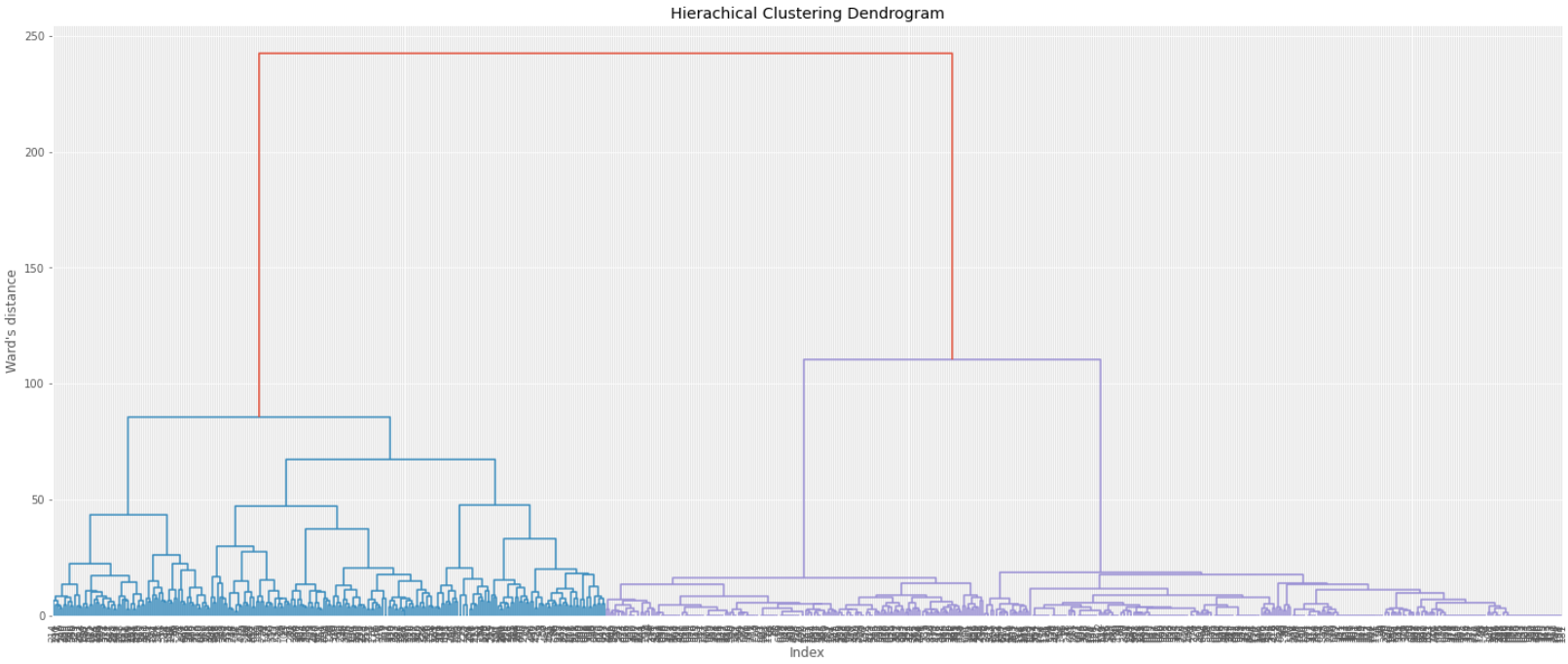
1.Hierarchical clustering (Ιεραρχική Ομαδοποίηση)

Πρώτα θα δούμε μέσο της Ιεραρχικής Ομαδοποίησης πως μπορούμε να ομαδοποιήσουμε τα δεδομένα .Μέσα στην Ιεραρχική ομαδοποίηση θα χρησιμοποιήσουμε τον αλγόριθμο Ward που υπολογίζει :

$$\Delta(A,B) = \sum_{i\in A\cup B} \|\vec{x}_i - \vec{m}_{A\cup B}\|^2 - \sum_{i\in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i\in B} \|\vec{x}_i - \vec{m}_B\|^2 \quad (2)$$

$$= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2 \quad (3)$$

Δυο αποστάσεις μεταξύ δυο ομάδων A και B που το άθροισμα του τετράγωνου τους θα αυξηθεί όταν θα γίνει η ομαδοποίηση .

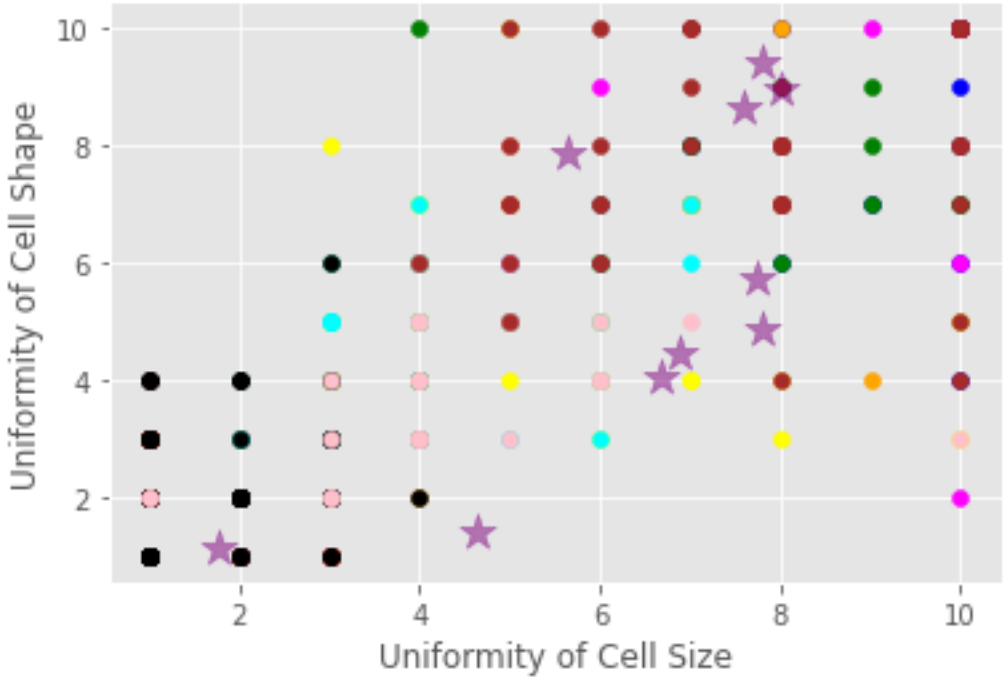


Από ότι παρατηρούμε η μέθοδος Ιεραρχικής ομαδοποίησης με τον αλγόριθμο ward χρησιμοποιούνε 2 βασικές ομαδοποιήσεις που είναι η βέλτιστες στον dataset μας όπως θα δούμε και με την μέθοδο K-Means.

2.K-Means Clustering

Αρχικά θα υπολογίσουμε την παρτιτούρα Silhouette για να δούμε ποιος είναι ο βέλτιστος αριθμός ομάδων που μπορούμε να πάρουμε και να μας ωφελήσει στο να κάνουμε την ομαδοποίηση των δεδομένων.

Παρατηρούμε ότι ο αλγόριθμος K-Means μετρά από της δυο ομαδοποιήσεις δεν έχει και τόσο νόημα να βάλουμε παραπάνω από 2 ομάδες και αυτό αποδίνεται από την ξαφνική καμπυλότητα του γράφου στο σχήμα μας.

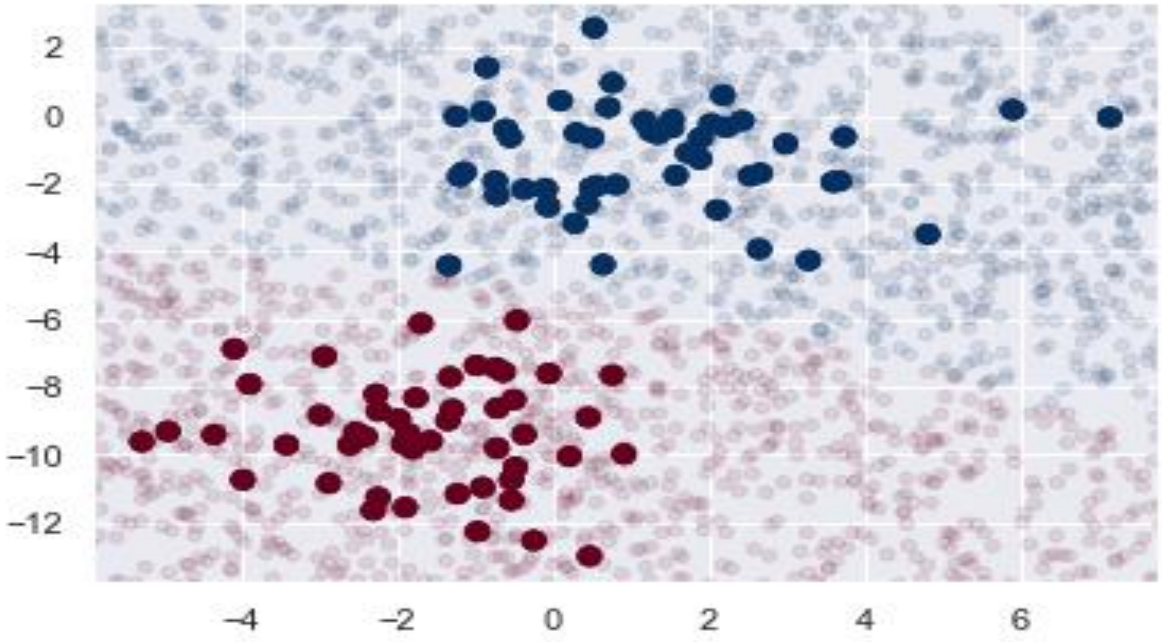


Στο σχήμα βλέπουμε την ομαδοποίηση των δεδομένων σχετικά με το σχήμα του όγκου και το μέγεθος του.

Εδώ βλέπουμε της διάφορες ομαδοποιήσεις που μπορούν να προκύψουν όταν χρησιμοποιήσουμε τον αλγόριθμο και βάλουμε εμείς 10 ομαδοποιήσεις. Δεν είναι απαραίτητό αλλά το χρησιμοποιήσαμε από περιέργεια πως θα είναι να ομαδοποιηθεί σε 10 ομάδες (Τα αστεράκια είναι το avg των ομάδων).

Τα αποτελέσματα είναι ότι και οι δυο αλγόριθμοι προβλέπουν ορθά το ελάχιστο αριθμό ομαδοποίησης που είναι απαραίτητος (στην περίπτωση μας είναι δυο).

1.Θεωριας Πιθανοτήτων του Bayes



Παρατηρούμε πάλι ότι από το γράφημα ότι δημιουργούνται δυο ομαδοποιήσεις μέσα στα δεδομένα μας .

Συμπεράσματα

Τα αποτελέσματα από το supervised και το unsupervised learning είναι ότι το supervised είναι κυρίως χρησιμοποιείται για το regression και να γίνει η ομαδοποιήσει μέσο μιας γραμμικών συναρτήσεων και μη ανάλογα τον αλγόριθμο. Ενώ η Ιεραρχική χρησιμοποιεί τα διαφορά διανύσματα εξαρτάται τον αλγόριθμο για να δημιουργηθούν τα αποτελέσματα . Το συμπέρασμα μου είναι ότι είναι πιθανώς το unsupervised learning είναι το κατάλληλο εργαλείο για την ομαδοποίηση των δεδομένων μιας και είναι πιο ακριβής στα αποτελέσματα του . Αυτό εξαρτάται και από την φύση των δεδομένων και του προβλήματος που ήμαστε σε θέση να αντιμετωπίσουμε.