2^{η} ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Ε17155,ΤΣΟΥΦΗΣ ΘΕΟΔΩΡΟΣ

Email : tsoufis.thodoris@gmail.com

ΕΞΑΜΗΝΟ $4^{ο}$

2η ΕΡΓΑΣΙΑ

ΘΕΟΔΩΡΟΣ ΤΣΟΥΦΗΣ
ΤΜΗΜΑ : ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

tsoufis.thodoris@gmail.com

ΠΕΡΙΛΗΨΗ

Σκοπος της εργασιας ειναι ,εφαρμοζοντας στα δεδομενα μας καποιον αλογοριθμο συσταδοποιησης , να δουμε εαν οι εγγραφες ομαδοποιουνται σε καλα διαχωρισμενες συσταδες . Να μπορεσουμε δηλαδη να ανακαλυψουμε ομαδες σημειων (εγγραφων) τα οποια ειναι ομοια , ή αλλιως εχουν μικρη αποσταση μεταξυ τους.

ΕΙΣΑΓΩΓΗ

Μας δινεται ενα συνολο δεδομενων στο οποιο καθε εγγραφη αποτελει μια αιτηση η οποια εχει ενα Ranking για την εισαγωγη σε καποιο Nursery School.Το πεδιο Ranking της καθε εγγραφης μπορει να παρει 5 τιμες .Επομενως αυτο που θελω να διακρινω στο τελος της αναλυσης μου ,ειναι 5 διαφορετικες και σχετικα ευκρινης συσταδες .Αυτο που κανω αρχικα ειναι να μετατρεψω τα δεδομενα μου , κυριως γιατι αυτα ειναι κατηγορικα και ειναι δυσκολο να δουλεψω μαζι τους .Ας παρουμε για παραδειγμα το χαρακτηριστικο has_nurs το οποιο μπορει να εχει values : proper, less_proper, improper, critical, very_crit . Αυτο που κανω ειναι να αντιστοιχησω εναν ακεραιο σε καθε value ετσι ωστε το «χειροτερο» value να εχει την μικροτερη τιμη και το «καλυτερο» την μεγαλυτερη . Δηλαδη θα εχω very_crit=1 , critical=2,improper=3,less_proper=4,proper=5.Με τον ιδιο τροπο αντικαθιστω ολα τα values ολων των χαρακτηριστικων καθε ενος αντικειμενου του συνολου δεδομενων πριν ξεκινησω την αναλυση του.

ΠΕΡΙΓΡΑΦΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Ενα αντικειμενο του συνολου δεδομενων αποτελειται απο 9 χαρακτηριστικα.Το πρωτο εχει να κανει με την φυση των επαγγελματων των γονεων .Το δευτερο ειναι το has nursery , το τριτο εχει να κανει με τη δομη της οικογενειας , το τεταρτο με τον αριθμο των παιδιων στην οικογενεια , το πεμπτο με την κατασταση κατοικιας της οικογενειας,το εκτο με την οικονομικη της κατασταση , το εβδομο με την κοινωνικη κατασταση της , το ογδοο με την κατασταση υγειας και το ενατο ειναι το Ranking που δοθηκε στο συγκεκριμενο αντικειμενο του συνολου δεδομενων.
Σημαντικο επισης ειναι πως τα χαρακτηριστικα ομαδοποιουνται.Τα πρωτα 2 αποτελουν την ομαδα EMPLOY, τα επομενα 2 την ομαδα STRUCTURE, ,τα επομενα 2 την ομαδα FINANCE και τα επομενα 2 την ομαδα SOC_HEALTH .

ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

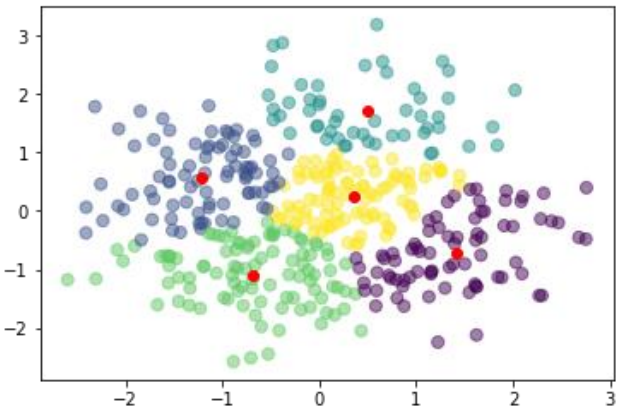
Για τους σκοπους της εργασιας , ο αλγοριθμος που χρησιμοποιειται ειναι ο K-MEANS .Ο αλγοριθμος αυτος, αναλογα με τον αριθμο συσταδων που θελουμε να οπτικοποιησουμε , υπολογιζει τα κεντρα για καθε συσταδα (πχ αν θελουμε 5 συσταδες , θα βρει 5 κεντρικα σημεια).Επειτα αναθετει καθε σημειο του συνολου δεδομενων στο κοντινοτερο κεντρο . Η διαδικασια αναθεσης των σημειων στα κοντινοτερα κεντρα , γινεται επαναληπτικα εως οτου επιτευχθεί κάποιο κριτήριο σύγκλισης.Ο K-MEANS χρησιμοποιειται εδω κυριως διοτι ξερουμε εξ’ αρχης τον

αναμενομενο αριθμο συσταδων(5 συμφωνα με το dataset ,οσες δηλαδη και οι διαφορετικες τιμες που μπορει να παρει το χαρακτηριστικο Ranking).

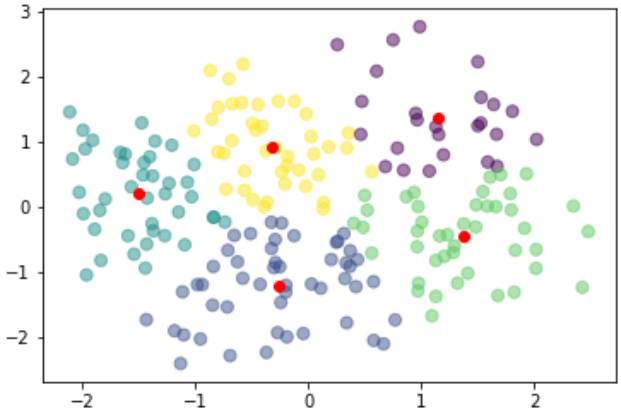
ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Το πρωτο βημα της μελετης ειναι αυτο που περιγραφηκε στην εισαγωγη (μετατροπη των δεδομενων) .Στη συνεχεια αυτο που κανουμε ειναι να παρουμε ενα δειγμα 400 απο τις 12960 εγγραφες , διοτι υποθετουμε πως με 400 εγγραφες οι συσταδες θα ειναι πιο ευκρινης .Επειτα αφαιρουμε το χαρακτηριστικο Ranking απο καθε εγγραφη .Τωρα τα δεδομενα μας αποτελουνται απο 8 χαρακτηριστικα .Τις τιμες των χαρακτηριτικων αυτων , τις τυποποιουμε συμφωνα με τον τυπο : $z = \frac{x-u}{s}$, οπου u ειναι ο μεσος ορος για ενα συγκεκριμενο χαρακτηριστικο ολων των εγγραφων ,s η τυπικη αποκλιση για ενα συγκεκριμενο χαρακτηριστικο ολων των εγγραφων, x η αρχικη τιμη του χαρακτηριστικου και z η νεα τιμη που θα παρει το χαρακτηριστικο .Αυτο το κανουμε ετσι ωστε οι διαφορετικες συσταδες να αναγνωριζονται πιο ευκολα.Υστερα χρησιμοποιουμε PCA ετσι ωστε να μειωσουμε τον αριθμο των τυποποιημενων χαρακτηριστικων(διαστασεων) ανα εγγραφη, απο 8 σε 2 .Αμεσως μετα εφαρμοζουμε τον K-MEANS στο προεπεξεργασμενο data set μας και παιρνουμε τα ακολουθα αποτελεσματα :

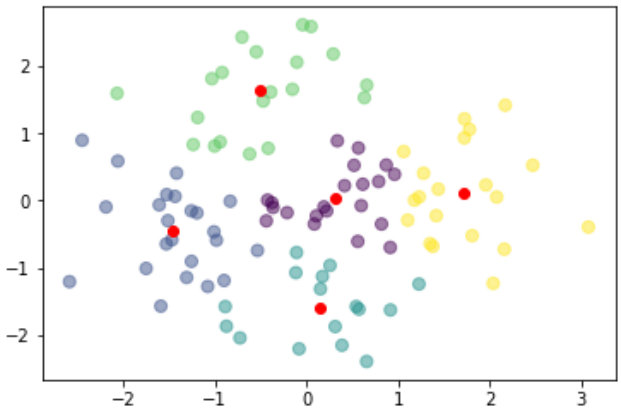
-Για αριθμο συσταδων ισο με 5 και 400 δειγματα απο το dataset :



-Για αριθμο συσταδων ισο με 5 και 200 δειγματα απο το dataset :



-Για αριθμο συσταδων ισο με 5 και 100 δειγματα απο το dataset :



Τα κοκκινα σημεια ειναι τα κεντρα των συσταδων .Οπως φαινεται και οπως αναμενωταν , μπορουμε να διακρινουμε 5 διαφορετικες και σχετικα ευκρινης συσταδες , οσα ειναι και τα διαφορετικα Ranking.

Τελος αυτο που κανω ειναι να μετρησω το Silhouette coefficient για διαφορετικα πληθη κεντρικων σημειων , συγκεκριμενα για πληθη κεντρων ισα με 2,3,4,5.Τα αποτελεσματα που παιρνω ειναι τα εξης :

-$1^{η}$ *περιπτωση* :

```
For n_clusters = 2 The average
silhouette_score is : 0.34
For n_clusters = 3 The average
silhouette_score is : 0.39
For n_clusters = 4 The average
silhouette_score is : 0.33
For n_clusters = 5 The average
silhouette_score is : 0.33
```

-$2^{η}$ *περιπτωση* :

```
For n_clusters = 2 The average
silhouette_score is : 0.32
For n_clusters = 3 The average
silhouette_score is : 0.35
For n_clusters = 4 The average
silhouette_score is : 0.35
For n_clusters = 5 The average
silhouette_score is : 0.36
```

-$3^{η}$ *περιπτωση* :

```
For n_clusters = 2 The average
silhouette_score is : 0.36
For n_clusters = 3 The average
silhouette_score is : 0.40
For n_clusters = 4 The average
silhouette_score is : 0.38
For n_clusters = 5 The average
silhouette_score is : 0.36
```

ΣΥΜΠΕΡΑΣΜΑΤΑ

Αυτο που συμπεραινω μετα απο την εφαρμογη του K-MEANS στο συνολο δεδομενων μου , ειναι πως προκειται για εναν αρκετα καλο αλγοριθμο συσταδοποιησης ο οποιος βρισκει ομαδες(συσταδες) εγγραφων με σχετικη επιτυχια, οπως φαινεται και απο το silhouette score .Επισης πολυ σημαντικο συμπερασμα που εβγαλα απο την εργασια αυτη ειναι η μεγαλη σημαντικοτητα που εχει η προεπεξεργασια των δεδομενων μου , πριν κανω την αναλυση μου.

ΠΗΓΕΣ :

-Websites : https://pandas.pydata.org/
https://scikit-learn.org/stable/

-Βιβλια:

- Εξόρυξη και Ανάλυση Δεδομένων: Βασικές Έννοιες και Αλγόριθμοι
Mohammed J. Zaki, Wagner Meira Jr. , Εκδόσεις Κλειδάριθμος