Συσταδοποίηση

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Συσταδοποίηση

Συσταδοποίηση

Συσταδοποίηση

Ο βασικός στόχος αυτού του μαθήματος είναι η εξοικείωση με θέματα που αφορούν την τρίτη σημαντική εργασία της εξόρυξης δεδομένων, δηλαδή την ανάλυση των συστάδων.

Πιο συγκεκριμένα, παρουσιάζεται μία σειρά από βασικούς ορισμούς αναφορικά με την ανάλυση συστάδων και τη συσταδοποίηση, και εξετάζονται με λεπτομέρεια τρεις κατηγορίες τεχνικών συσταδοποίησης: η διαμεριστική συσταδοποίηση, η ιεραρχική συσταδοποίηση, και συσταδοποίηση που βασίζεται στην πυκνότητα. Στη συνέχεια γίνεται αναφορά σε συγκεκριμένους αλγορίθμους συσταδοποίησης, όπως ο αλγόριθμος k-means, ο συσσωρευτικός ιεραρχικός αλγόριθμος και ο αλγόριθμος DBSCAN.

Παρουσιάζονται, επίσης, διαφορετικές τεχνικές εφαρμογής της ιεραρχικής συσταδοποίησης, όπως είναι η τεχνική του απλού συνδέσμου (ή της ελάχιστης απόστασης), η τεχνική του πλήρους συνδέσμου (ή της μέγιστης απόστασης), η τεχνική του μέσου όρου ομάδας και η μέθοδος Ward.

Συσταδοποίηση

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Στην επιβλεπόμενη μάθηση (Supervised Learning) μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής.

Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο να μπορεί να κατηγοριοποιήσει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις.

Αντίθετα, στη μη επιβλεπομένη μάθηση (Unsupervised Learning) μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέρουσα δομή των δεδομένων.

Για παράδειγμα, η συσταδοποίηση είναι μια από τις τεχνικές μη επιβλεπόμενης μάθησης. Δοθέντων κάποιων δεδομένων χωρίς κλάσεις, οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα δεδομένα σε συστάδες, έτσι ώστε εγγραφές, οι οποίες ανήκουν στην ίδια συστάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.

Συσταδοποίηση

Έννοια της Συστάδας

Στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις ή ετικέτες και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες.

Οι συστάδες που δημιουργούνται θέλουμε να διαχωρίζουν ορθά τα δεδομένα.

Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ' ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Συσταδοποίηση

Αλγόριθμος k-means

Ο αλγόριθμος k-means ξεκινάει με k τυχαία σημεία, τα οποία ονομάζονται κεντροειδή της συστάδας και δηλώνουν το κέντρο βάρους της συστάδας.

Το k υποδηλώνει σε πόσες συστάδες θέλουμε ο αλγόριθμος να δημιουργήσει.

Ο αλγόριθμος εκτελεί επαναληπτικά δύο βήματα. Το πρώτο βήμα αφορά την ανάθεση σε κάποια συστάδα, ενώ το δεύτερο βήμα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροειδούς κάθε συστάδας.

Πιο αναλυτικά, όσον αφορά στο πρώτο βήμα, δηλαδή την ανάθεση σε κάποια συστάδα, ο αλγόριθμος εξετάζει κάθε δείγμα σε σχέση με τα κεντροειδή των συστάδων.

Με χρήση κάποιου μέτρου απόστασης, αναθέτει το εξεταζόμενο δείγμα στη συστάδα, της οποίας το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα.

Στο δεύτερο βήμα, παίρνοντας τον μέσο όρο των δειγμάτων κάθε συστάδας, επανυπολογίζονται τα κεντροειδή της κάθε συστάδας, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στην πρόσφατα διαμορφωμένη συστάδα.

Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρις ότου τα κεντροειδή των συστάδων να μετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου.

Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου.

Ψευδοκώδικας

```
Αρχικοποίησε τυχαία τα k κεντροειδή των συστάδων μ1,μ2,…,μk.
Επανέλαβε{
  Εξέτασε κάθε δείγμα και ανέθεσε το στη συστάδα με το
  πλησιέστερο κεντροειδές (min |x^(i) - μ_k|^2 )
  Επανυπολόγισε τα κεντροειδή υπολογίζοντας το μέσο όρο των
  δειγμάτων της συστάδας
}
```

Συσταδοποίηση

Τυχαία Αρχικοποίηση Κεντροειδών

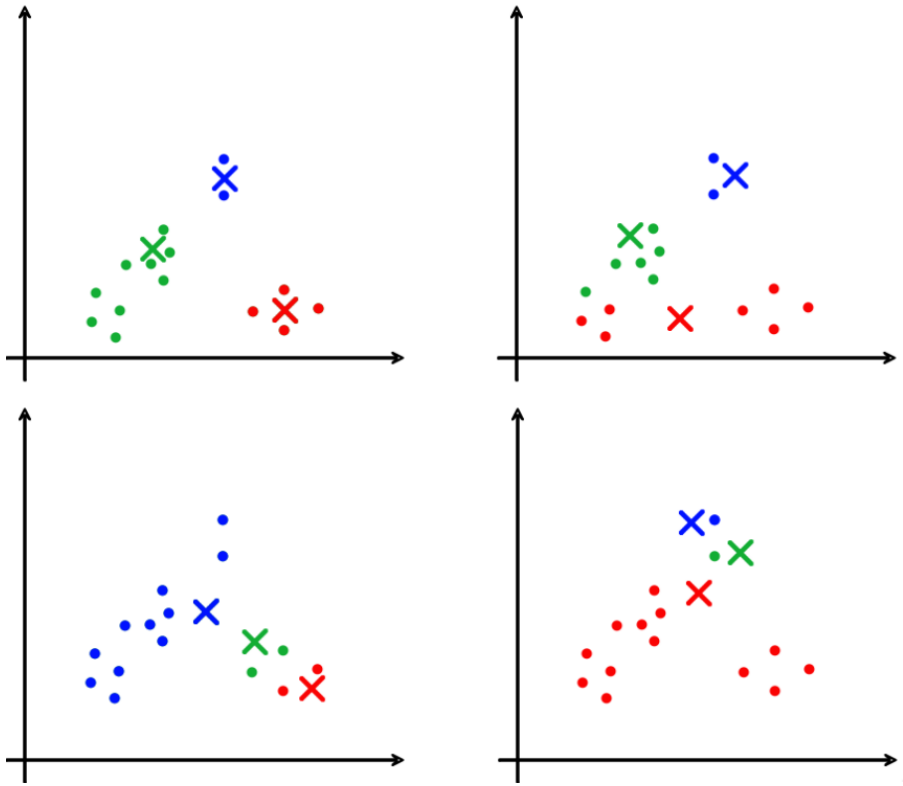
Το πρώτο βήμα του αλγορίθμου k-means είναι η τυχαία αρχικοποίηση των k κεντροειδών των συστάδων.

Παρόλο που το συγκεκριμένο βήμα φαίνεται απλό και ασήμαντο, αρκετές φορές μια «κακή» αρχικοποίηση μπορεί να οδηγήσει σε κακής ποιότητας συστάδες στην πορεία.

Στην παρακάτω εικόνα βλέπουμε ένα παράδειγμα τεσσάρων τυχαίων αρχικοποιήσεων των κεντροειδών, ενώ με χρώμα υποδεικνύεται το πώς τελικά καταλήγουν να είναι οι συστάδες που δημιουργεί ο αλγόριθμος.

Συσταδοποίηση

Τυχαία Αρχικοποίηση Κεντροειδών



Πάνω αριστερά έχουμε την καλύτερη περίπτωση.

Ακολουθεί μια λιγότερο ποιοτικά καλή συσταδοποίηση πάνω δεξιά.

Στις δυο τελευταίες περιπτώσεις είναι προφανές ότι η αρχικοποίηση επηρεάζει αρνητικά τη διαδικασία συσταδοποίησης.

Συσταδοποίηση

Επιλογή του Αριθμού Συστάδων

Ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του k, δηλαδή του αριθμού των συστάδων.

Ο αριθμός των συστάδων δίνεται ως είσοδος από τον χρήστη και η επιλογή του σωστού αριθμού επαφίεται στη δική του γνώση και εμπειρία.

Να υπενθυμίσουμε ότι κατά τη συσταδοποίηση δεν δίνεται το επιπλέον χαρακτηριστικό κλάσης των δειγμάτων.

Συνεπώς, η διαδικασία επιλογής του αριθμού συστάδων, ενδεχομένως, να απαιτήσει την εξερεύνηση και μελέτη των δεδομένων, για παράδειγμα, μέσα από οπτικοποιήσεις, προκειμένου να καταλήξουμε στον σωστό αριθμό συστάδων.

Συσταδοποίηση

Επιλογή του Αριθμού Συστάδων

Ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του k, δηλαδή του αριθμού των συστάδων.

Ο αριθμός των συστάδων δίνεται ως είσοδος από τον χρήστη και η επιλογή του σωστού αριθμού επαφίεται στη δική του γνώση και εμπειρία.

Να υπενθυμίσουμε ότι κατά τη συσταδοποίηση δεν δίνεται το επιπλέον χαρακτηριστικό κλάσης των δειγμάτων.

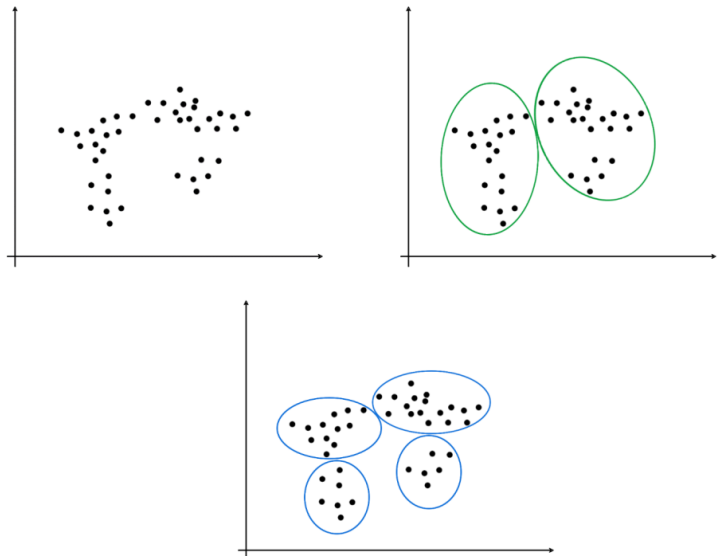
Συνεπώς, η διαδικασία επιλογής του αριθμού συστάδων, ενδεχομένως, να απαιτήσει την εξερεύνηση και μελέτη των δεδομένων, για παράδειγμα, μέσα από οπτικοποιήσεις, προκειμένου να καταλήξουμε στον σωστό αριθμό συστάδων.

Συσταδοποίηση

Επιλογή του Αριθμού Συστάδων

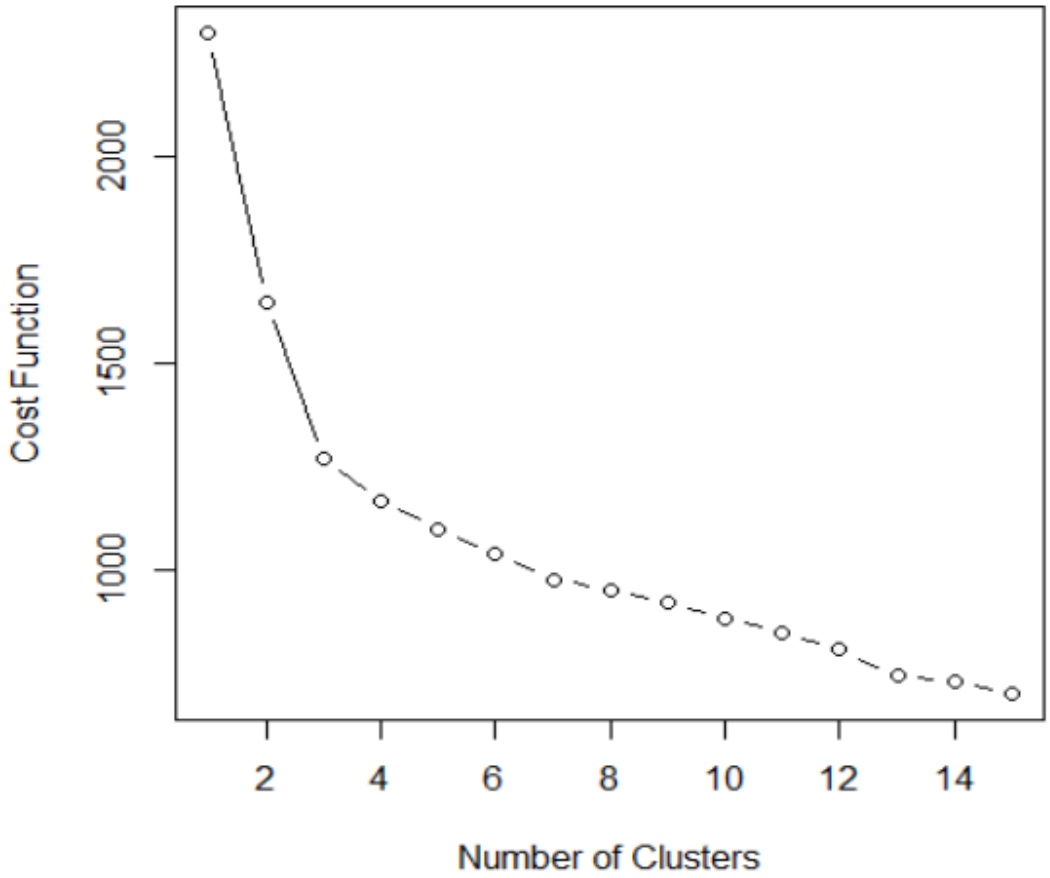
Αρκετές φορές τα ίδια τα δεδομένα είναι διφορούμενα. Για παράδειγμα, η οπτικοποίηση των δεδομένων, όπως φαίνονται στην Εικόνα, δείχνει ότι τα δεδομένα δεν είναι εύκολα διαχωρίσιμα (πάνω αριστερά).

Σε πόσες συστάδες θα πρέπει να διαχωριστούν; Σε δύο (πάνω δεξιά) ή σε τέσσερις (κάτω) συστάδες.



Συσταδοποίηση

Επιλογή του Αριθμού Συστάδων



Σχήμα: Ο κανόνας του αγκώνα.

Συσταδοποίηση

Υλοποίηση k-Means στην R

Στο ακόλουθο παράδειγμα θα παρουσιάσουμε την υλοποίηση του αλγορίθμου k-means στην R και τις σχετικές συναρτήσεις (βλέπε Κώδικας σε προηγούμενη διαφάνεια).

Το σύνολο δεδομένων iris περιέχει 50 μετρήσεις για καθένα από τα τρία διαφορετικά είδη λουλουδιών: setosa, versicolor και virginica (συνολικά 150 δείγματα).

Οι μετρήσεις αφορούν το μήκος και το πλάτος (σε cm) των πετάλων και των σέπαλων των λουλουδιών κάθε είδους.

Στόχος του παραδείγματος είναι να ελέγξουμε την ποιότητα της συσταδοποίησης, αφού αφαιρέσουμε το χαρακτηριστικό Species, το οποίο δηλώνει το είδος στο οποίο ανήκει το λουλούδι.

Συσταδοποίηση

Υλοποίηση k-Means στην R

Στο ακόλουθο παράδειγμα θα παρουσιάσουμε την υλοποίηση του αλγορίθμου k-means στην R και τις σχετικές συναρτήσεις (βλέπε Κώδικας σε προηγούμενη διαφάνεια).

Το σύνολο δεδομένων iris περιέχει 50 μετρήσεις για καθένα από τα τρία διαφορετικά είδη λουλουδιών: setosa, versicolor και virginica (συνολικά 150 δείγματα).

Οι μετρήσεις αφορούν το μήκος και το πλάτος (σε cm) των πετάλων και των σέπαλων των λουλουδιών κάθε είδους.

Στόχος του παραδείγματος είναι να ελέγξουμε την ποιότητα της συσταδοποίησης, αφού αφαιρέσουμε το χαρακτηριστικό Species, το οποίο δηλώνει το είδος στο οποίο ανήκει το λουλούδι.

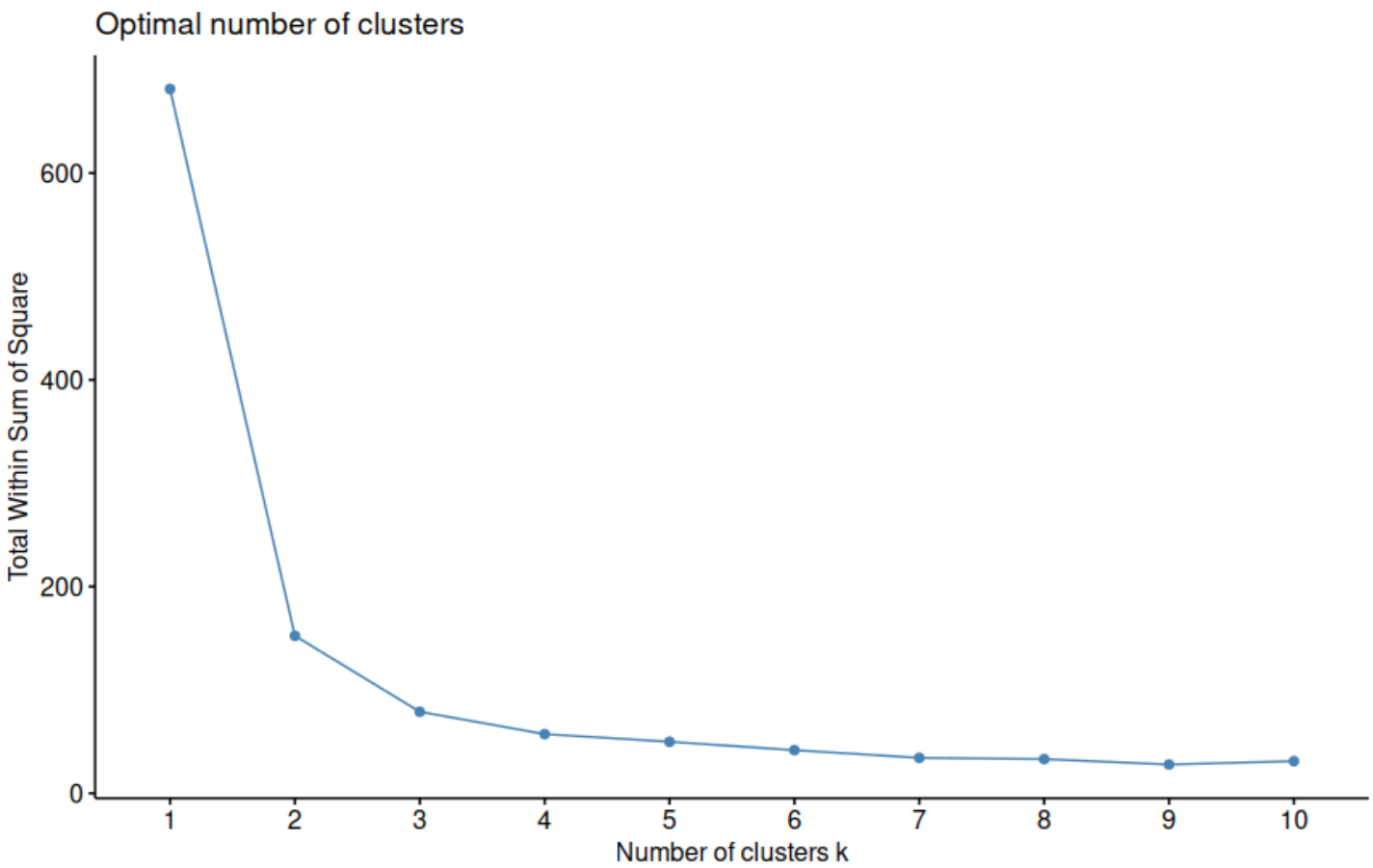
Συσταδοποίηση

Υλοποίηση k-Means στην R

```
iris_new <- iris
iris_new{$}Species <- NULL

library(cluster)
library(factoextra)
#create plot of number of clusters vs total within sum of squares
fviz_nbclust(iris_new, kmeans, method = "wss")

kc <- kmeans(iris_new, 3)
table(iris{\$}Species, kc{\$}cluster)

1  2  3
setosa      50  0  0
versicolor   0 48  2
virginica    0 14 36

plot(iris_new[c("Sepal.Length", "Sepal.Width")], col=kc$cluster)
points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col=1:3, pch=8, cex=2)
```

Συσταδοποίηση

Υλοποίηση k-Means στην R



Συσταδοποίηση

Υλοποίηση k-Means στην R