MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Πανεπιστήμιο Κύπρου - Τεχνητή Νοημοσύνη

MAI612 - ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διάλεξη 12: Ομαδοποίηση

Βασίλης Βασιλειάδης, PhD

Χειμερινό Εξάμηνο 2022/23

CYENS CENTRE OF EXCELLENCE

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διάλεξη 12: Ομαδοποίηση

Μαθησιακά αποτελέσματα

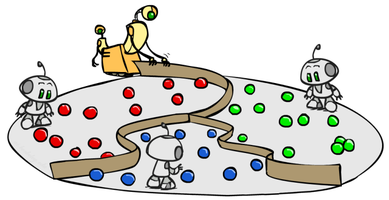
Θα μάθετε για:

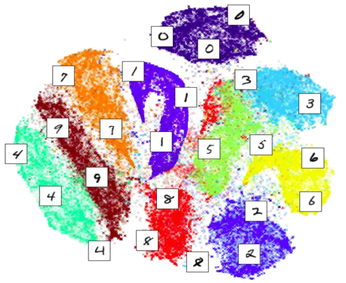
1. Το πρόβλημα του clustering στην μη εποπτευόμενη μάθηση και η διαφορά της μεταξύ της εποπτευόμενης ταξινόμησης
2. Ο αλγόριθμος k-means clustering: πώς λειτουργεί και πώς να τον ρυθμίσετε
3. Πώς η ομαδοποίηση μπορεί να βοηθήσει την εποπτευόμενη μάθηση

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μη εποπτευόμενη μάθηση

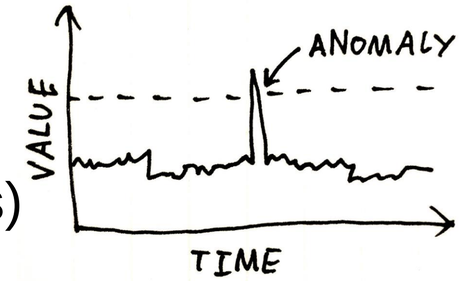
Παρουσιάζουμε μια συμβολή στο σύστημα, αλλά δεν έχουμε «δάσκαλο» για να παρέχουμε οποιοδήποτε «στόχο» αποτέλεσμα

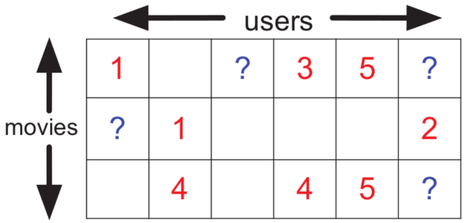




Κοινά προβλήματα:

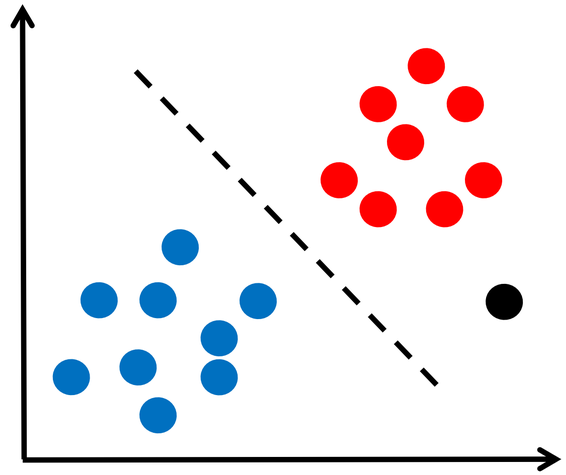
1. Clustering
2. Μείωση της διάστασης
3. Ανίχνευση ανωμαλίας
4. Ολοκλήρωση πίνακα (π.χ. recommender systems)





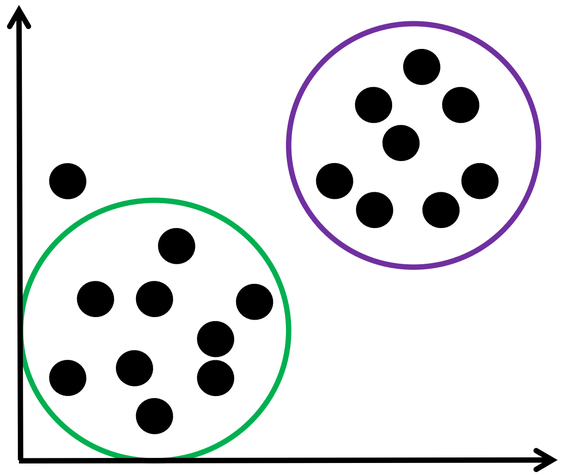
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εποπτευόμενη ταξινόμηση vs Clustering



Εποπτευόμενη δυαδική ταξινόμηση

Δεδομένα εκμάθησης: $\{(X^{(1)},y^{(1)}),\ldots (X^{(m)},y^{(m)})\}$



Clustering

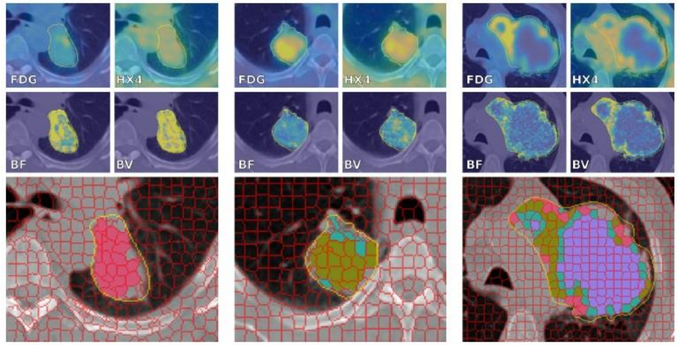
Δεδομένα εκμάθησης : $\{x^{(1)}, x^{(2)},\ldots x^{(m)}\}$

Co-financed by the European Union
Connecting Europe Facility

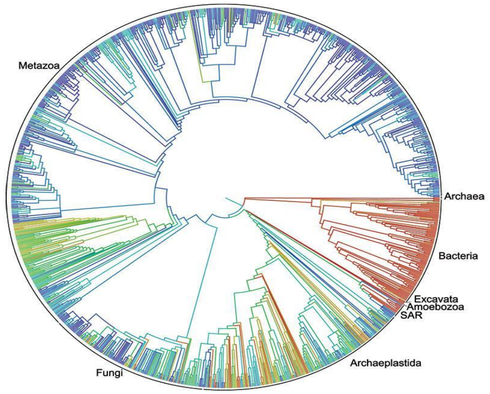
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εφαρμογές

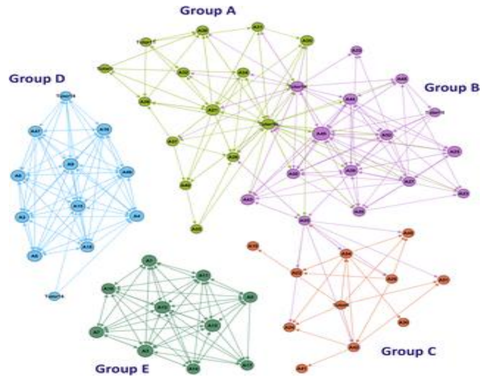
Διαχωρισμός της αγοράς



Ιατρική απεικόνιση



Οικολογία των ζώων



Ανάλυση κοινωνικών δικτύων

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος clustering

Είσοδος: Δεδομένα εκμάθησης: $\{x^{(1)}, x^{(2)}, \ldots x^{(m)}\}$, $x^{(i)} \in R^n$
Αριθμός συστάδων K

Τυχαία αρχικοποιήστε τα K Centoids συστάδων $\mu_1, \mu_2, \ldots \mu_K \in R^n$

$$min_k \left\| x^{(i)} - \mu_k \right\|^2$$

Επανέλαβε:

Αντιστοίχισε κάθε σημείο εκπαίδευσης σε ένα Centoid συστάδων

για i = 1 έως m

C(i) = δείκτης (1 έως K) του κεντροειδούς συστάδας πλησιέστερα στο $x^{(i)}$

J=2

C(3)=2, c(6)=2, c(7)=2

Ενημέρωση των Centoids συστάδων

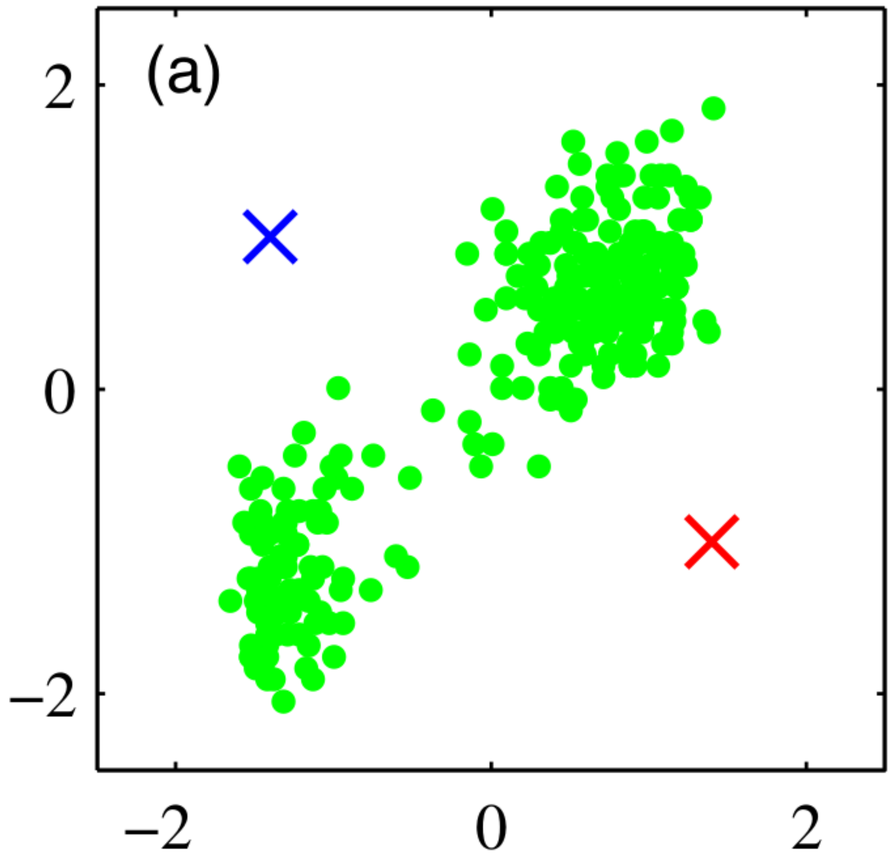
για j = 1 έως K

μ_j = μέσος όρος των σημείων που αποδίδονται στην ομάδα j

$\mu_2 = 1/3\ (x^{(3)}+x^{(6)}+x^{(7)})$

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering

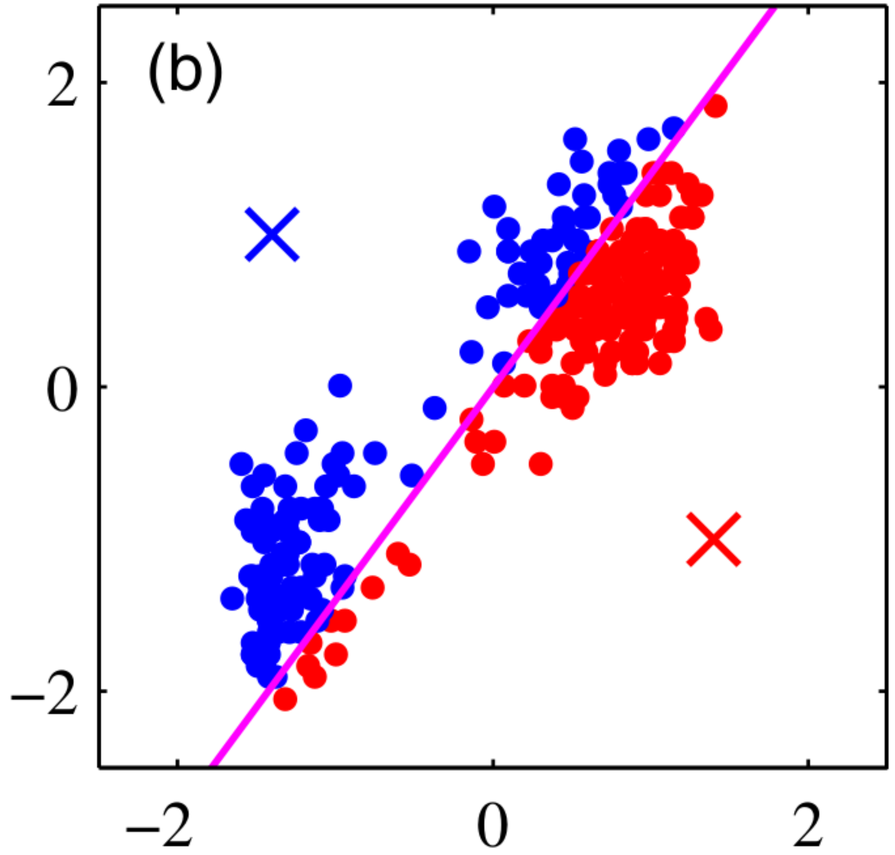


πηγή: Επίσκοπος 2006

Τυχαία αρχικοποίηση
K=2 centroids συστάδων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



πηγή: Επίσκοπος 2006

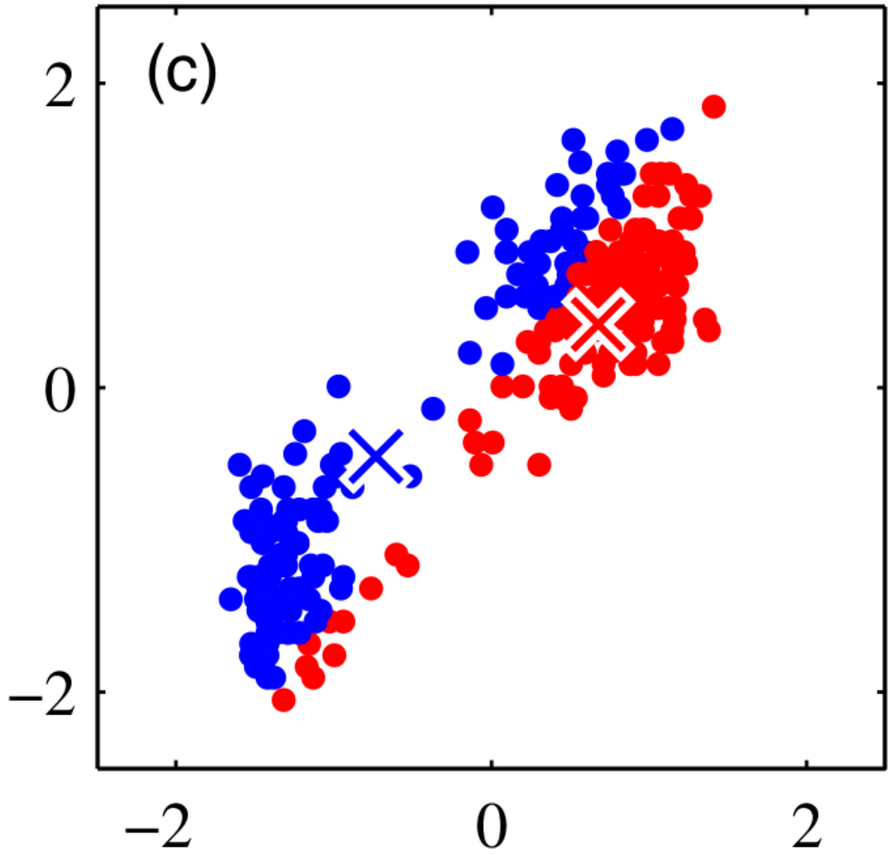
Επανάληψη: 1

Βήμα ανάθεσης

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



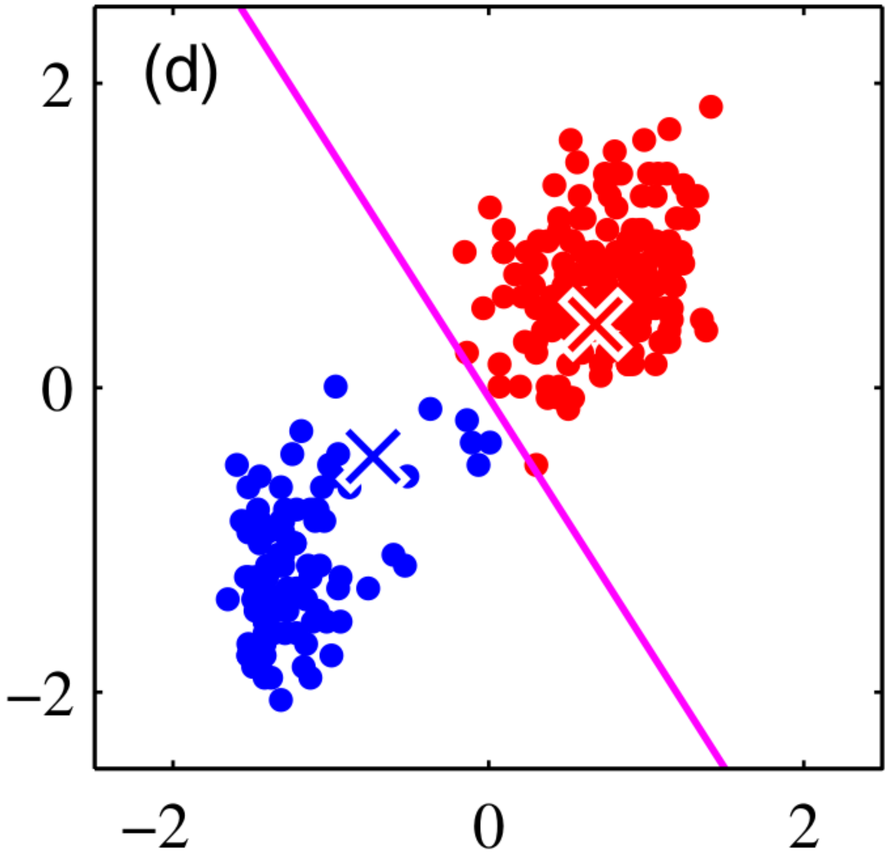
πηγή: Επίσκοπος 2006

επανάληψη: 1

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



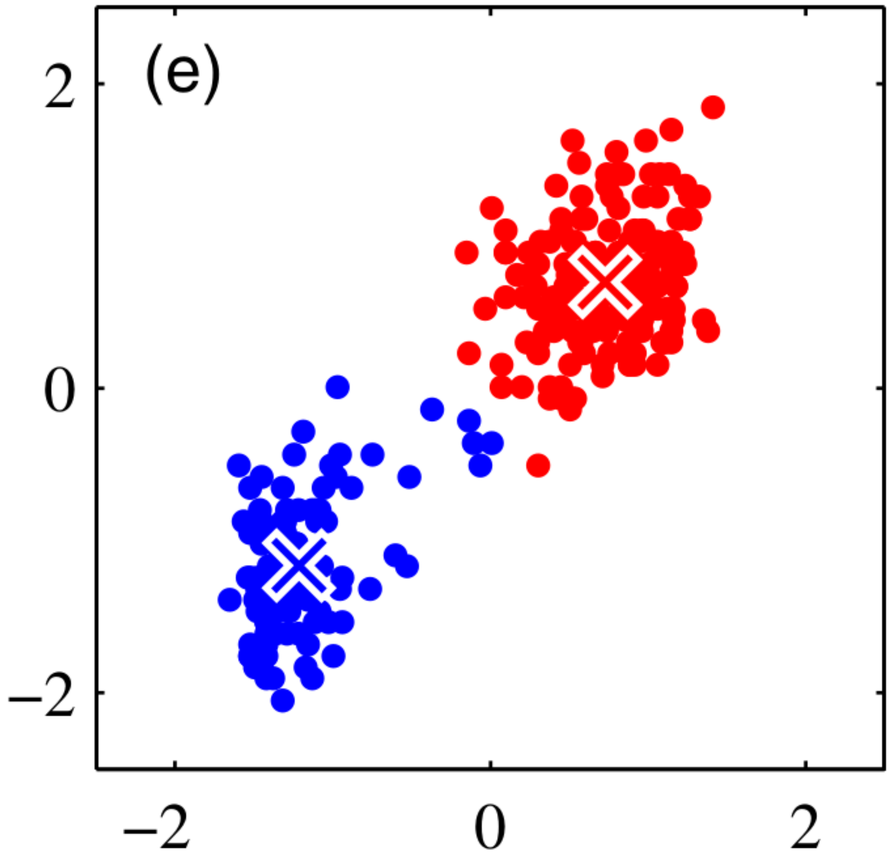
πηγή: Επίσκοπος 2006

επανάληψη: 2

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



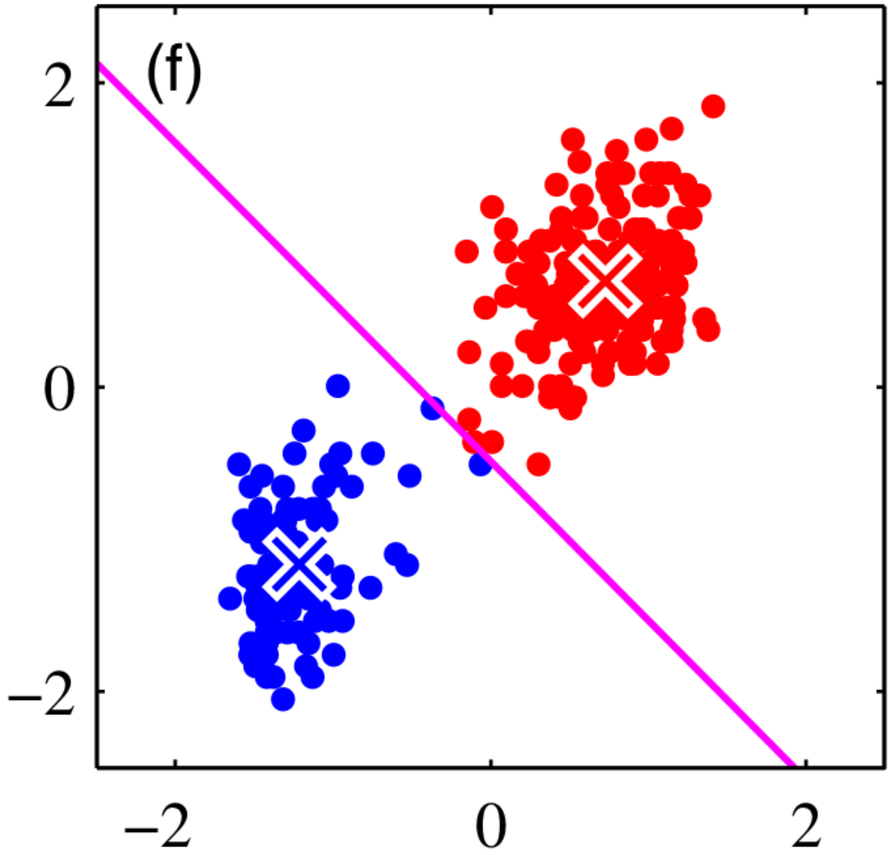
πηγή: Επίσκοπος 2006

επανάληψη: 2

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



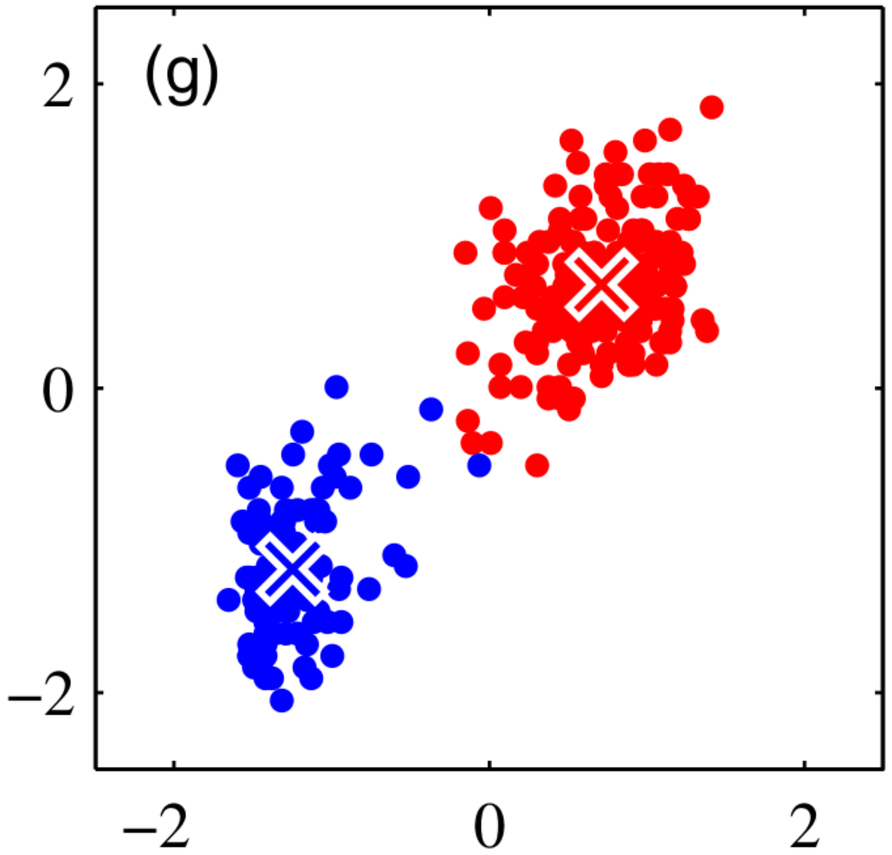
πηγή: Επίσκοπος 2006

επανάληψη: 3

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



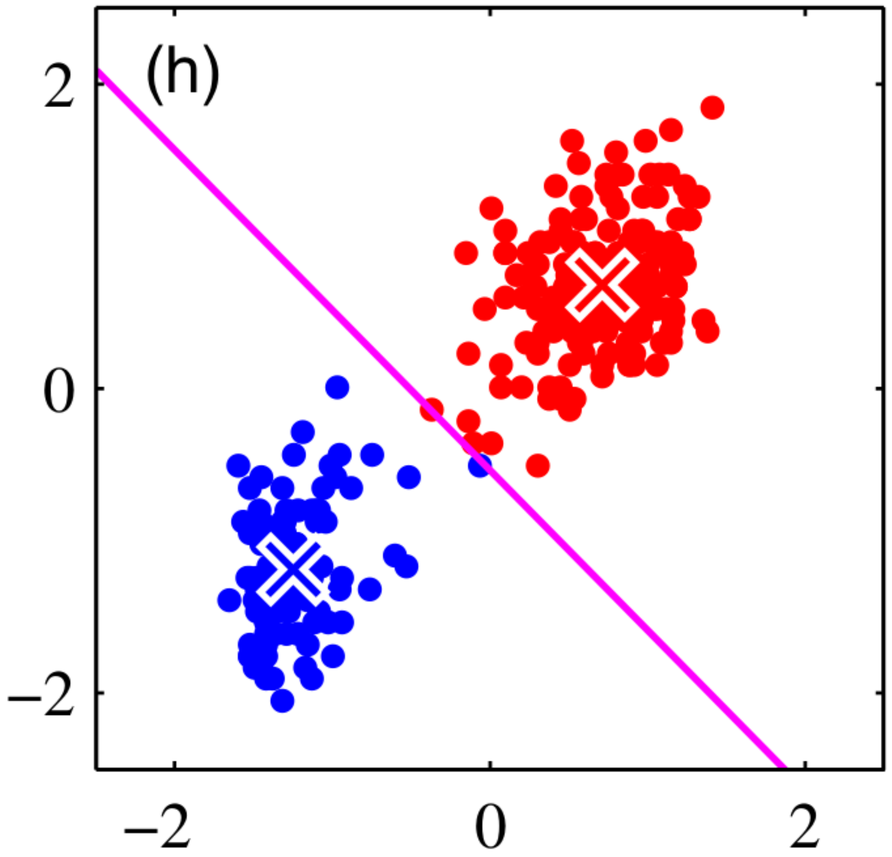
πηγή: Επίσκοπος 2006

επανάληψη: 3

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



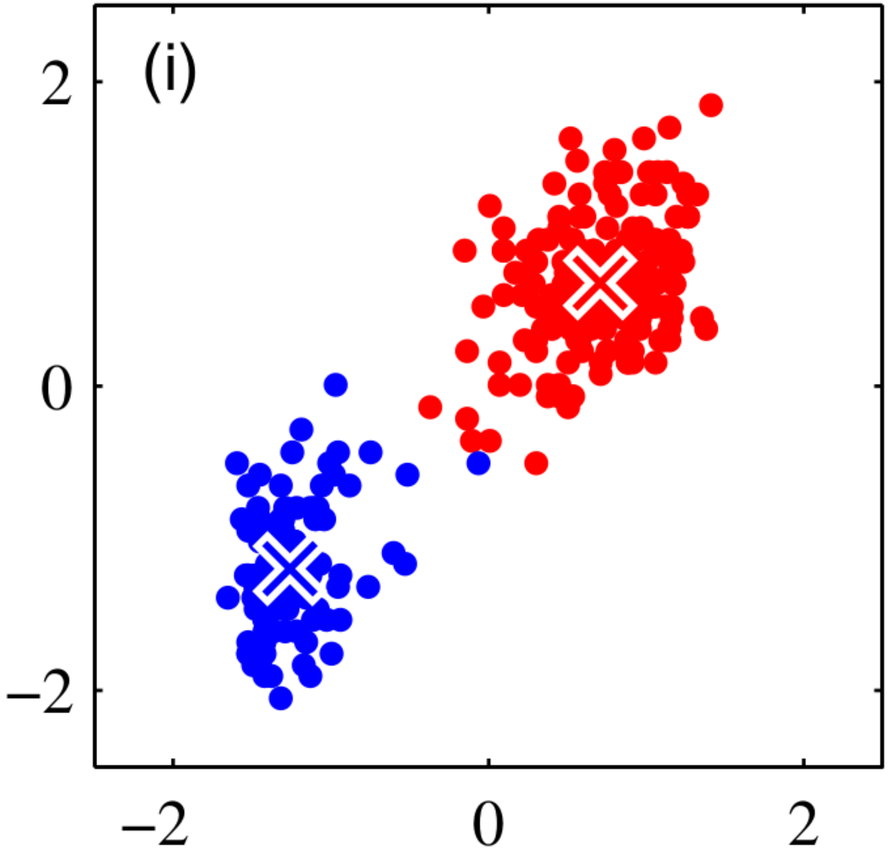
πηγή: Επίσκοπος 2006

επανάληψη: 4

Βήμα ανάθεσης

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-means clustering



πηγή: Επίσκοπος 2006

επανάληψη: 4

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Τυχαία αρχικοποίηση

Αντί για τυχαία αρχικά centroids,
τυχαία επιλογή περιπτώσεων εκπαίδευσης K

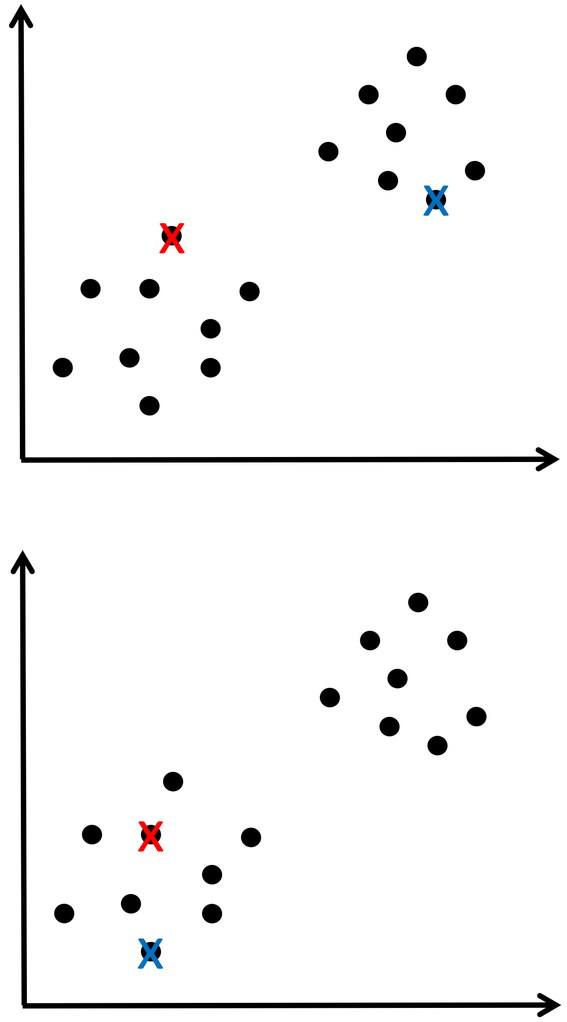
Σύνολο $\mu_1, \mu_2, \ldots \mu_K$ ίσο με αυτά τα σημεία K

Πρέπει να έχουμε K < m

K=2

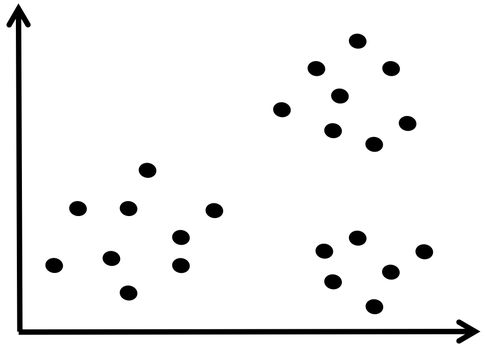
$\mu_1 = x^{(i)}$

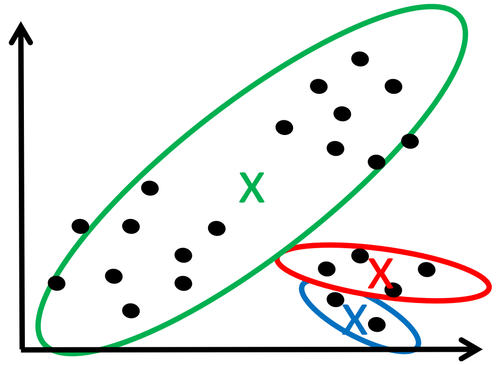
$\mu_2 = x^{(j)}$



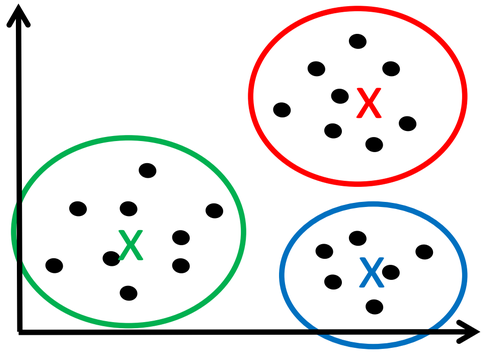
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Τυχαία αρχικοποίηση

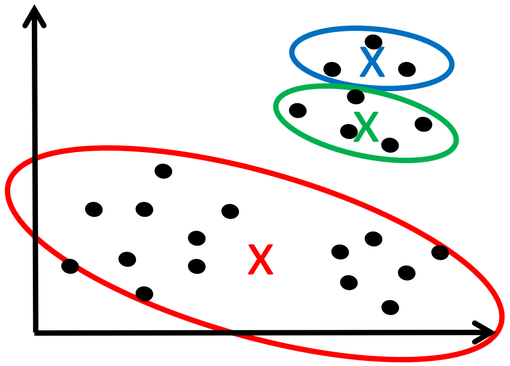




Τοπικό βέλτιστο

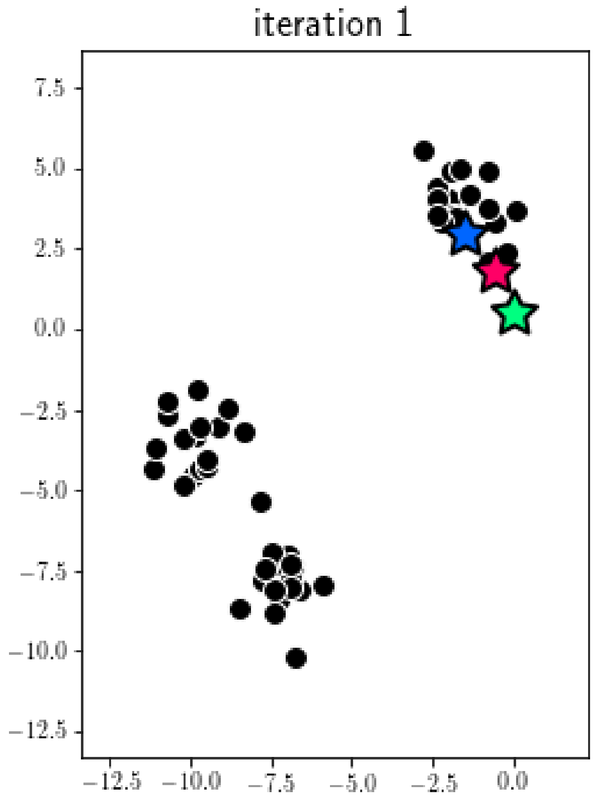


Γενικό βέλτιστο



This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

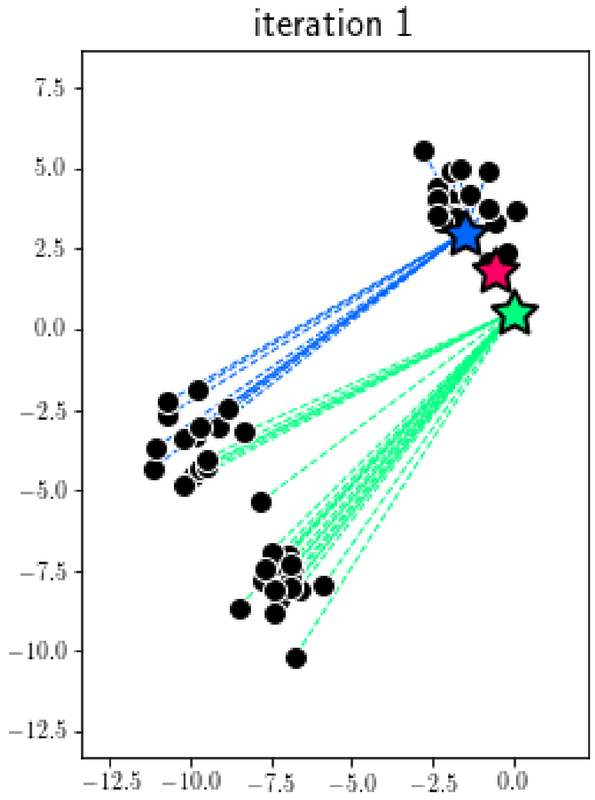
Κενές συστάδες



ΠΗΓΗ

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

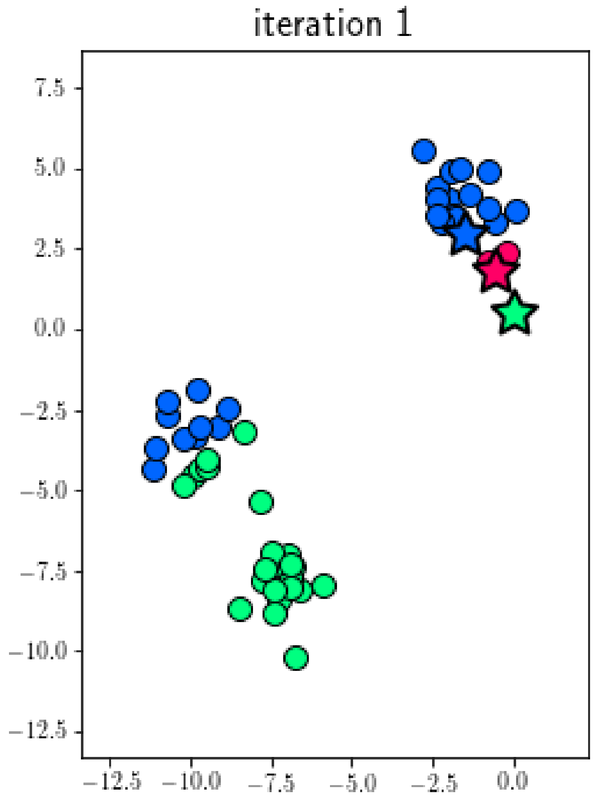


πηγη

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

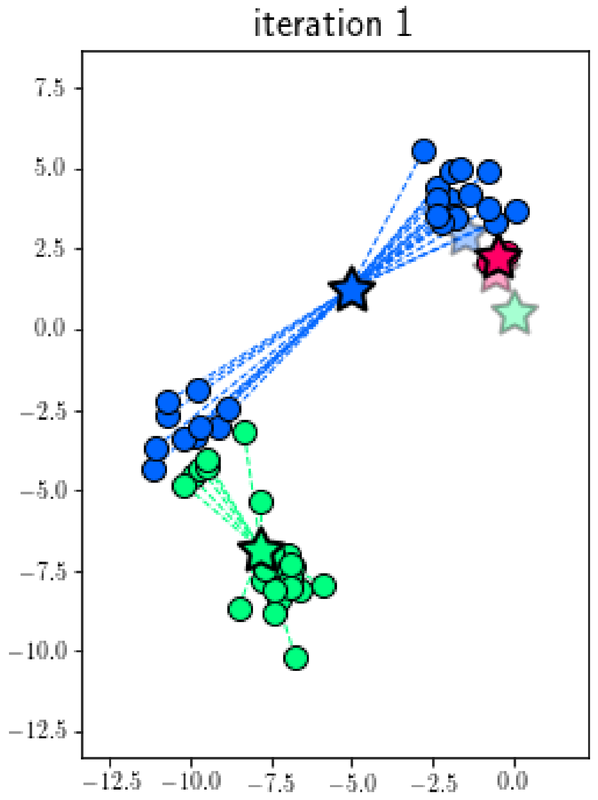


πηγη

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

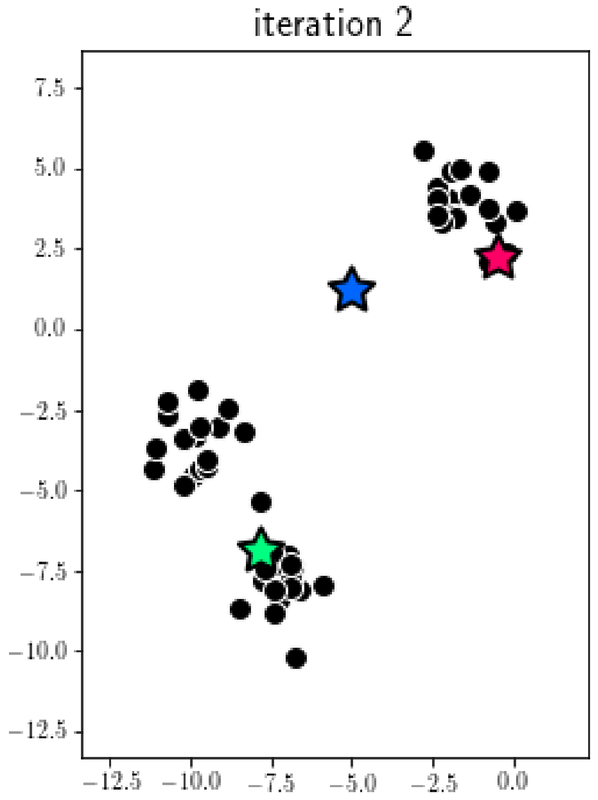


πηγη

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

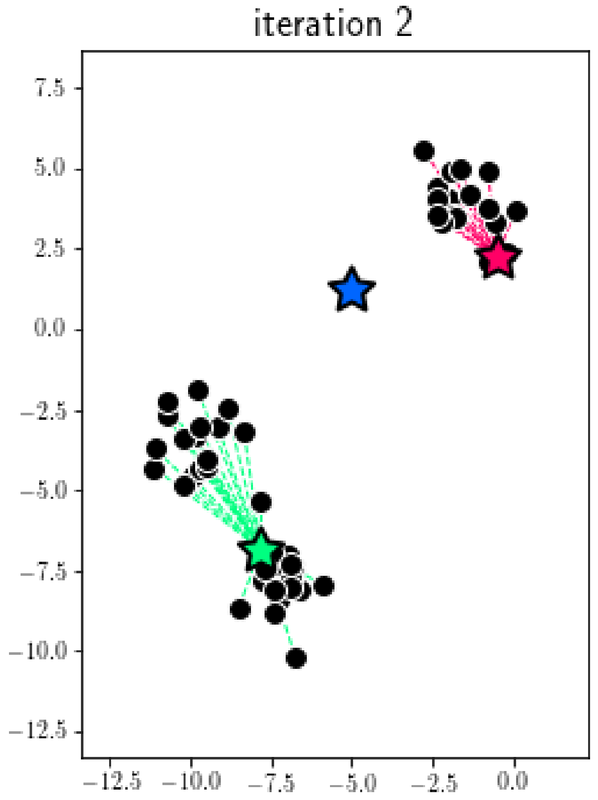
Κενές συστάδες



ΠΗΓΗ

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες



πηγη

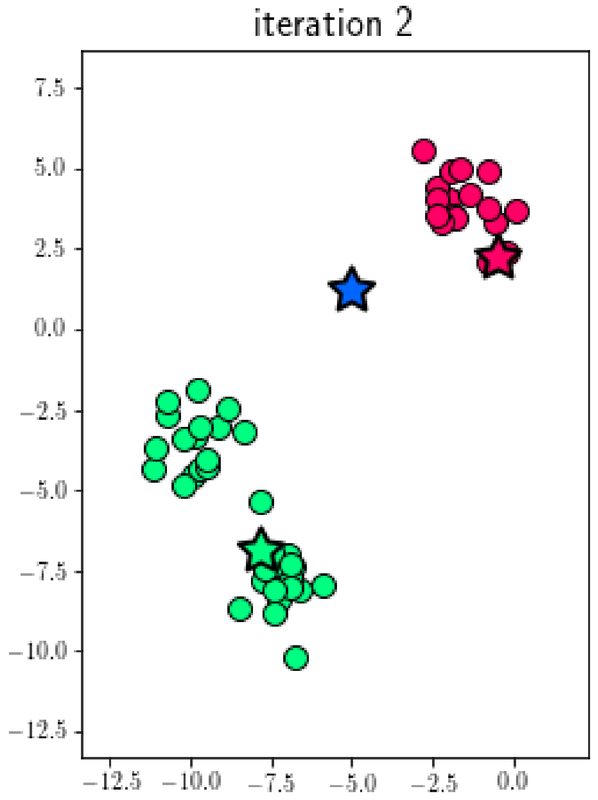
Βήμα ανάθεσης

Κανένα σημείο δεν είναι κοντά στο μπλε κεντροειδές

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες



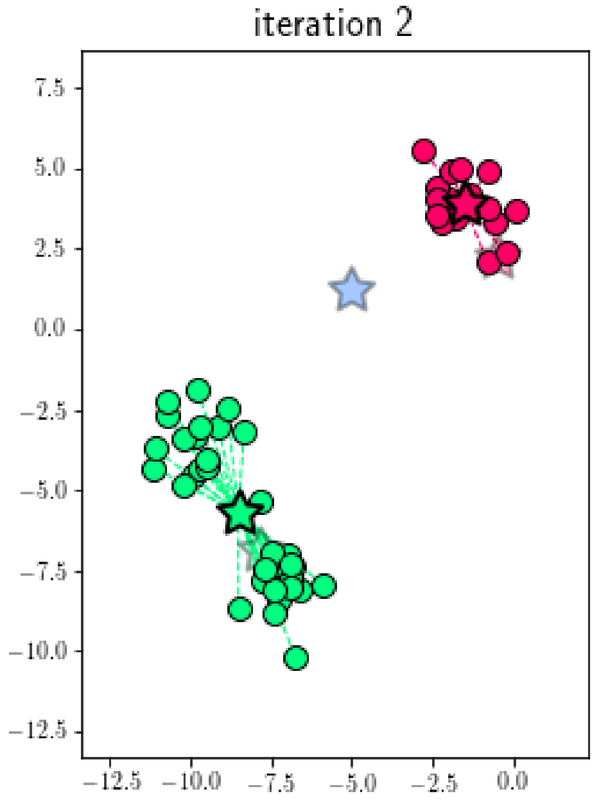
πηγη

Βήμα ανάθεσης

Το μπλε κεντροειδές καταλήγει χωρίς σημεία που του έχουν αποδοθεί

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

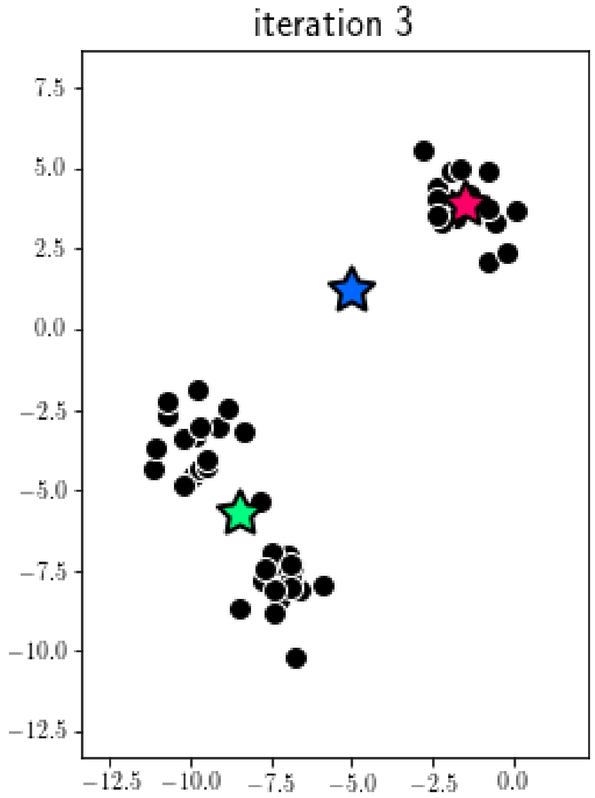


πηγη

Βήμα ενημέρωσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

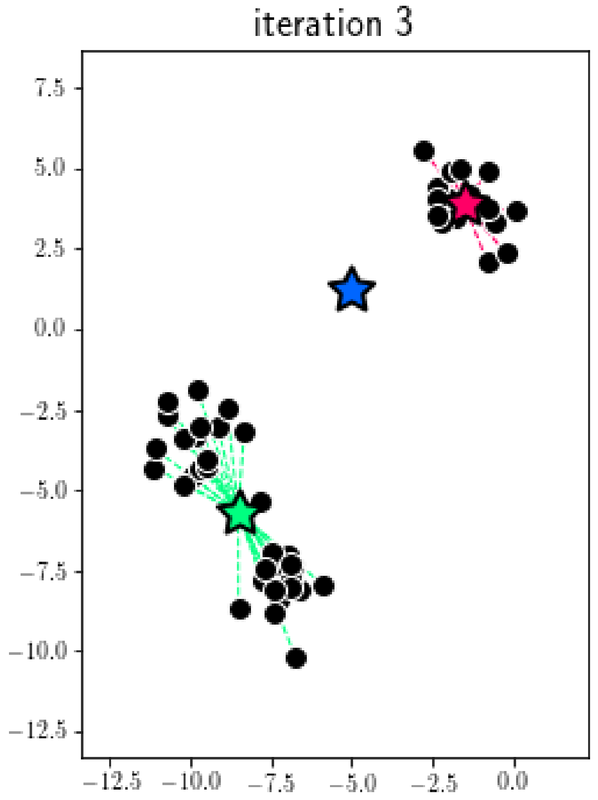
Κενές συστάδες



ΠΗΓΗ

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

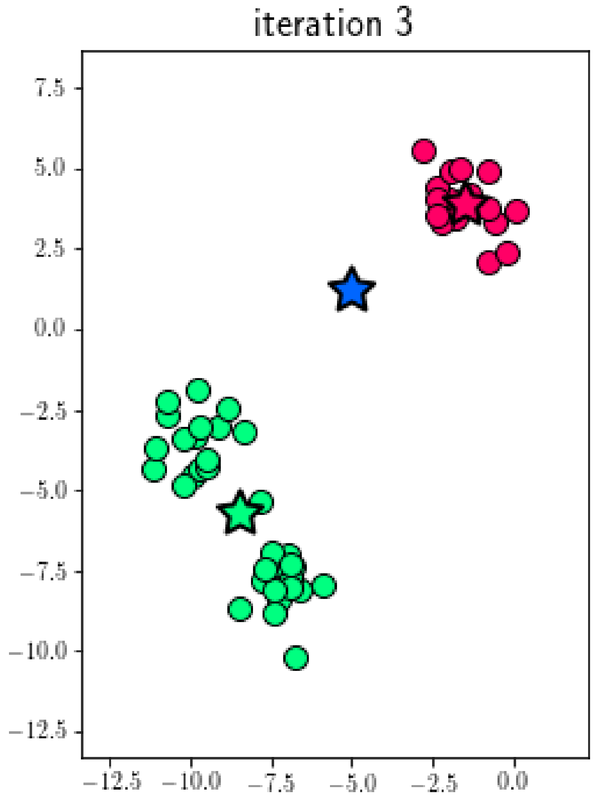


πηγη

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες

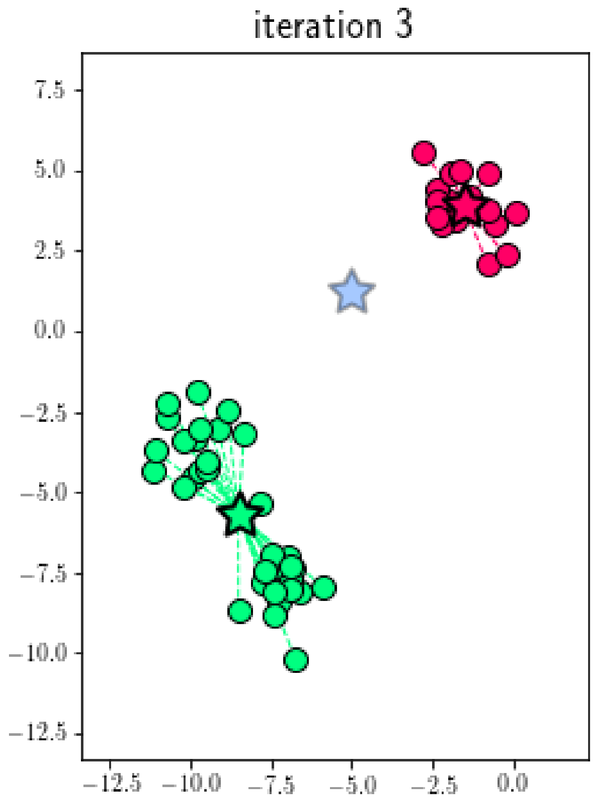


πηγη

Βήμα ανάθεσης

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κενές συστάδες



ΠΗΓΗ

Βήμα ενημέρωσης

Το μπλε centroid δεν θα λάβει ποτέ σημεία.

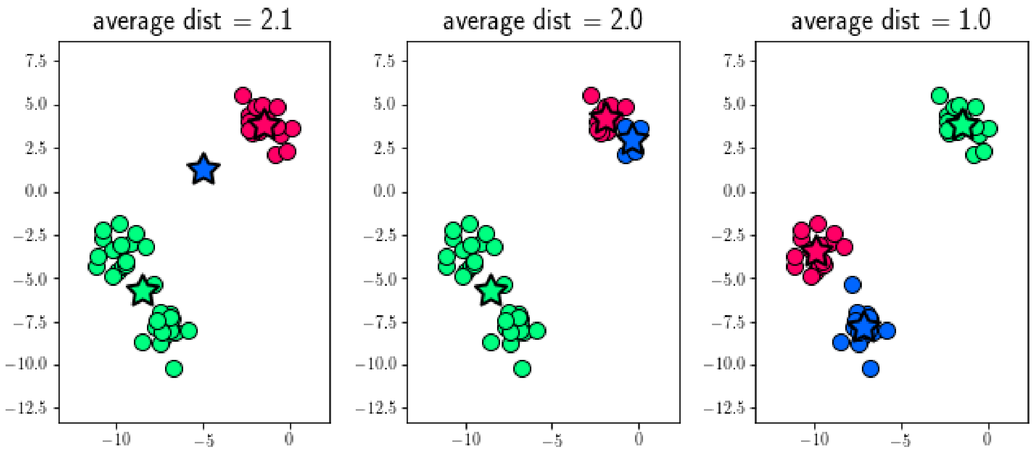
Πιθανή λύση: κατά την ανακάλυψη κενών συστάδων, μειώστε τον K κατά 1 ή επανεκκινήστε από μια διαφορετική τυχαία αρχικοποίηση

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αποφεύγετε το τοπικό βέλτιστο

Εκτελέστε k-means πολλές φορές:

για i = 1 έως N
  Τυχαία αρχικοποιήσε τα k-means
Τρέξε k-means. Θ = $c(1),...,c(m),\mu_1,...,\mu_K$
Υπολογίστε τη συνάρτηση κόστους: L(**θ**) = $L(c(1),...,c(m),\mu_1,...,\mu_K)$



Ομαδοποίηση με το χαμηλότερο κόστος L(**θ**)

Ποια είναι η συνάρτηση κόστους;

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Συνάρτηση κόστους

Συμβολισμοί

μ_j = centoid συστάδας j ($\mu_j \in R^n$)

C(i) = δείκτης της συστάδας (1 έως K) στον οποίο αποδίδεται $x^{(i)}$

$\mu_{c(i)}$ = centroid συστάδας στο οποίο αποδίδεται $x^{(i)}$

Στόχος $\underset{\theta}{minimize}\ L(\boldsymbol{\theta})$

Μέση απόσταση ενδο-συστάδων

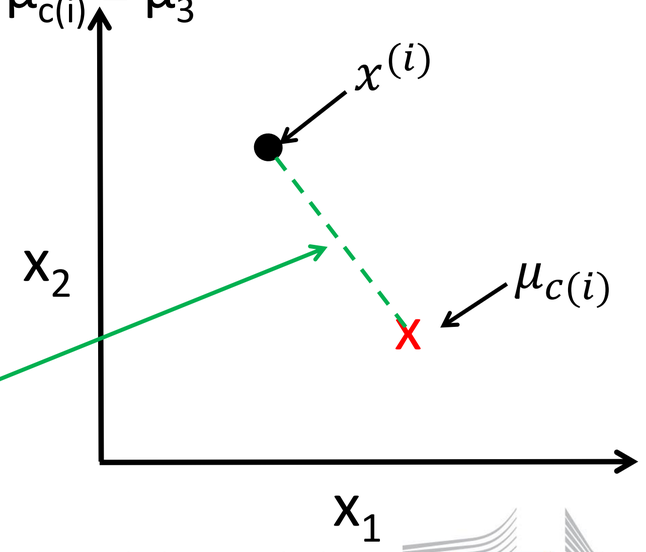
$$L(\boldsymbol{\theta}) = L(c(1),\ldots,c(m),\mu_1,\ldots,\mu_K) = \frac{1}{m}\sum_{i=1}^{m} || x^{(i)} - \mu_{c(i)} ||^2$$

Παράδειγμα

$x^{(i)}$ που έχουν ανατεθεί στην ομάδα 3

C(i) = 3

$\mu_{c(i)} = \mu_3$

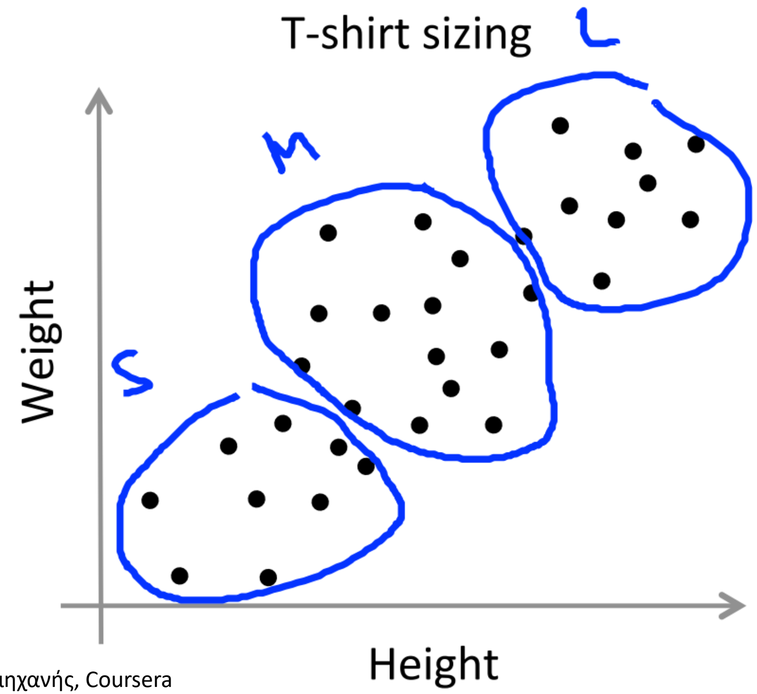


This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πώς να επιλέξετε το K

Γνώση τομέα

Μέγεθος T-shirt: K=3

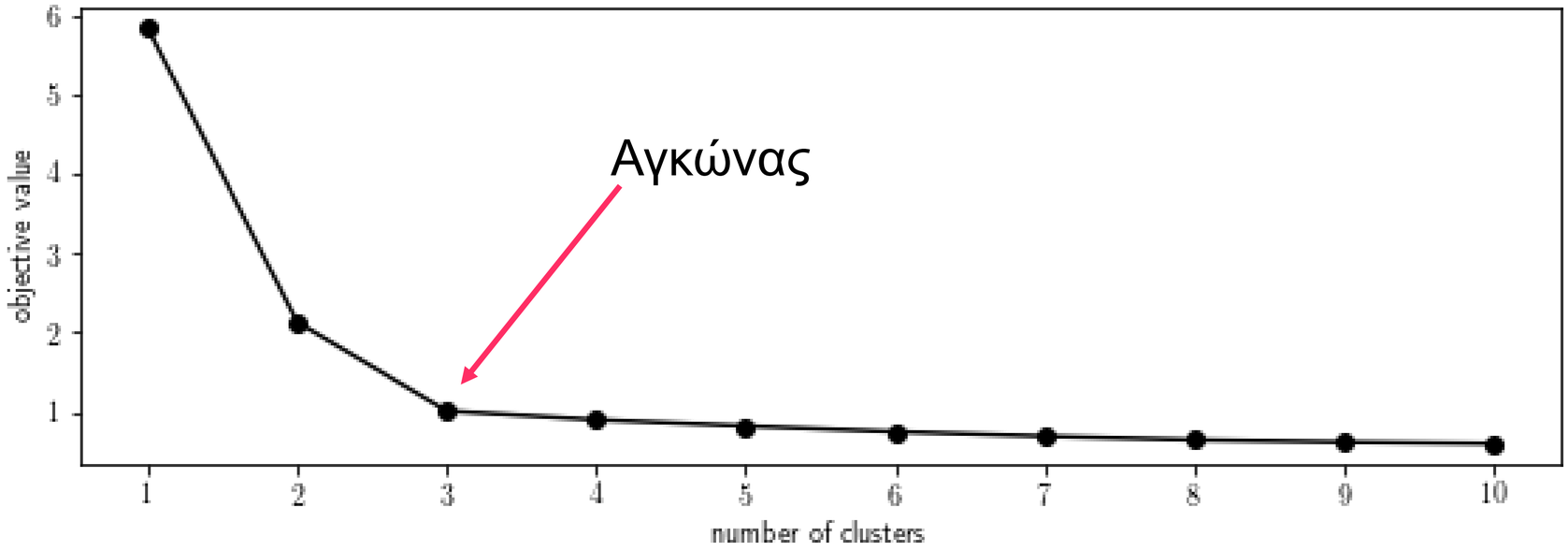


πηγή: Andrew Ng - Μάθηση μηχανής, Coursera

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πώς να επιλέξετε το K

Μέθοδος «Αγκώνας»



K = 3

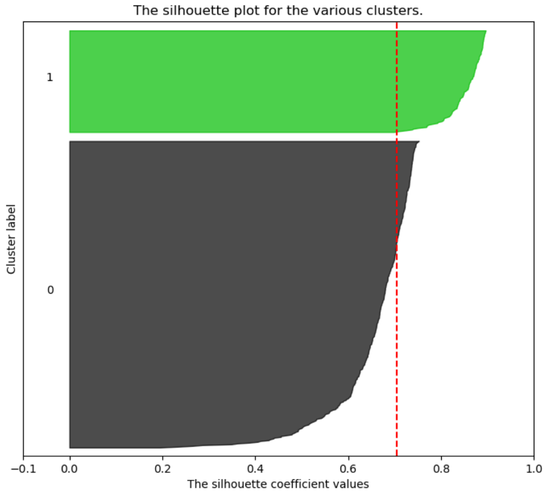
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πώς να επιλέξετε το K

- Ομοιόμορφο πάχος
- Χωρίς αρνητικές τιμές
- Όλα πέρα από το avg

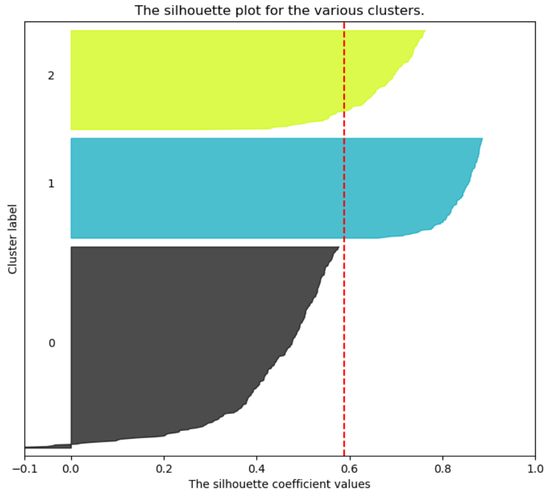
Βαθμολογία σιλουέτας

K=2



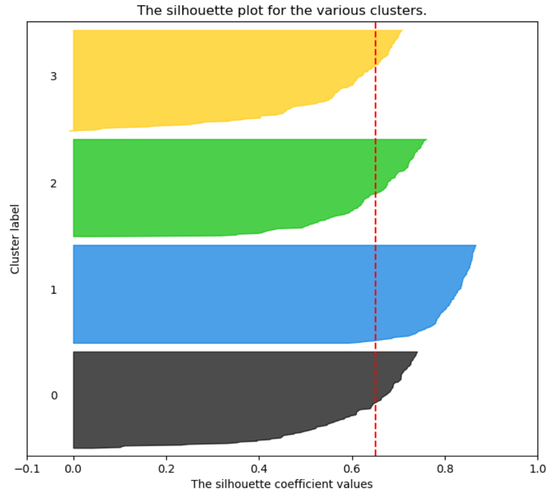
0.705

K=3



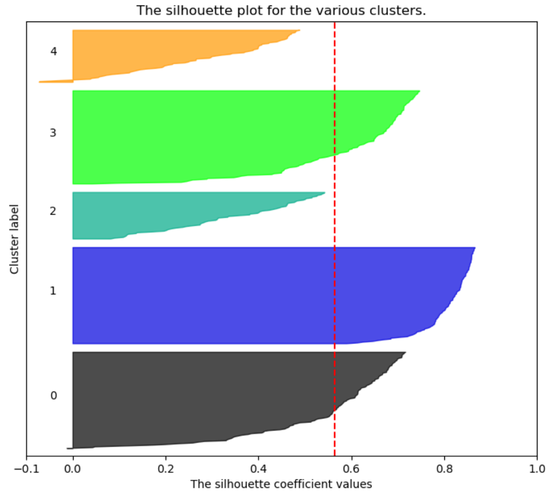
0.588

K=4



0.651

K=5



0.564

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Αλγόριθμος K-medoids

- Παρόμοια με τον k-means, με την προσπάθεια να ελαχιστοποιηθεί η απόσταση μεταξύ των σημείων που θα είναι στη συστάδα και τα σημεία που ορίζονται ως το κέντρο της συστάδας

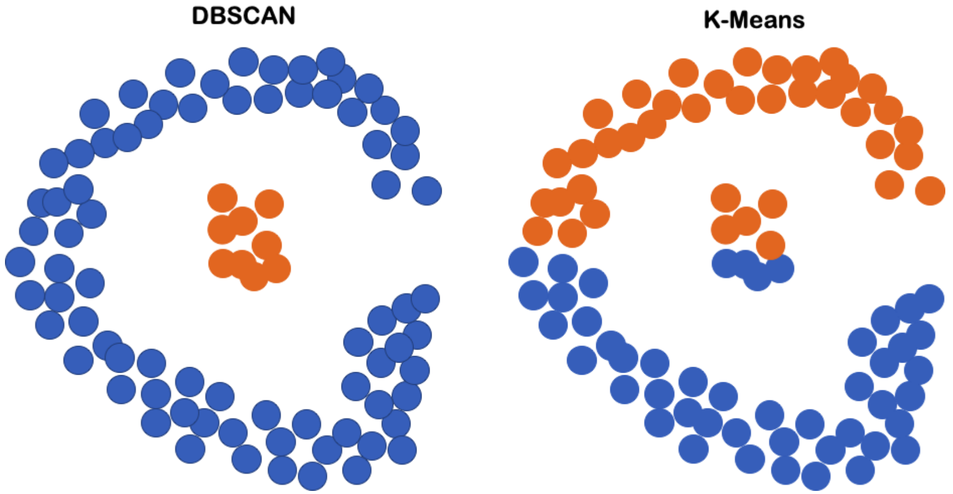
- Η διαφορά:
  - τα K-medoids επιλέγουν τα πραγματικά σημεία δεδομένων ως κέντρα
  - Ελαχιστοποιεί το άθροισμα των κατά ζεύγη ανομοιοτήτων αντί του αθροίσματος των τετραγώνων Ευκλείδειων αποστάσεων

- Αυτό το καθιστά πιο εύρωστο στο θόρυβο και τις ακραίες τιμές από ότι τον k-means

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

DBSCAN

- Αλγόριθμος clustering με βάση την πυκνότητα
- Δεν χρειάζεται να προσδιοριστεί ο αριθμός των συνεργατικών σχηματισμών
- Το βρίσκει αυτόματα με βάση την πυκνότητα των σημείων
- Ανάγκη καθορισμού της μέγιστης απόστασης μεταξύ δύο δειγμάτων για να θεωρηθεί το ένα ως εσωτερικό της γειτονιάς του άλλου



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Ιεραρχικό Clustering

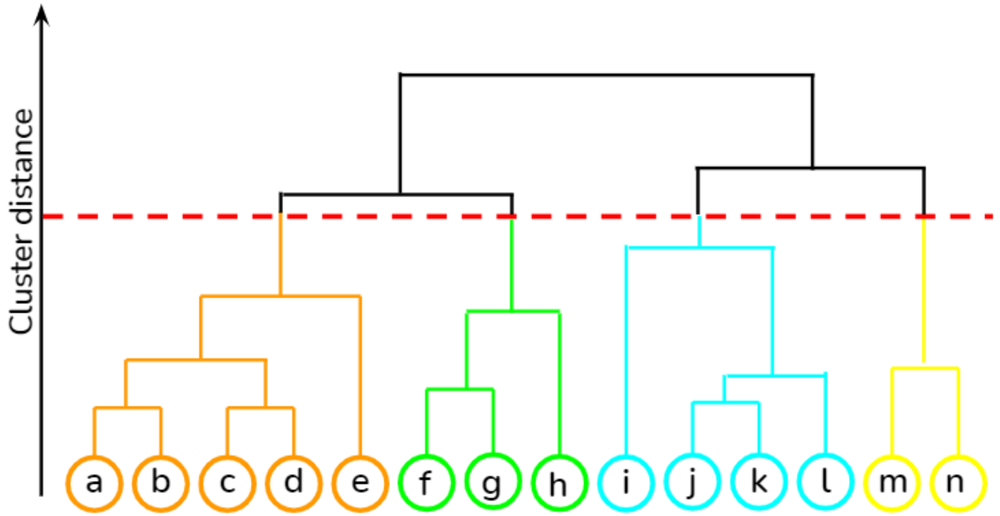
Χρησιμοποιείται όταν θέλουμε να οικοδομήσουμε μια ιεραρχία των συστάδων, και δεν έχουμε γνώση σχετικά με τον αριθμό των συστάδων

Δύο κατηγορίες:

- Agglomerative clustering: προσέγγιση από τη βάση προς την κορυφή
  - Κάθε σημείο ξεκινά από το δικό του σύμπλεγμα και τα ζεύγη των συστάδων συγχωνεύονται καθώς ανεβαίνουμε την ιεραρχία.

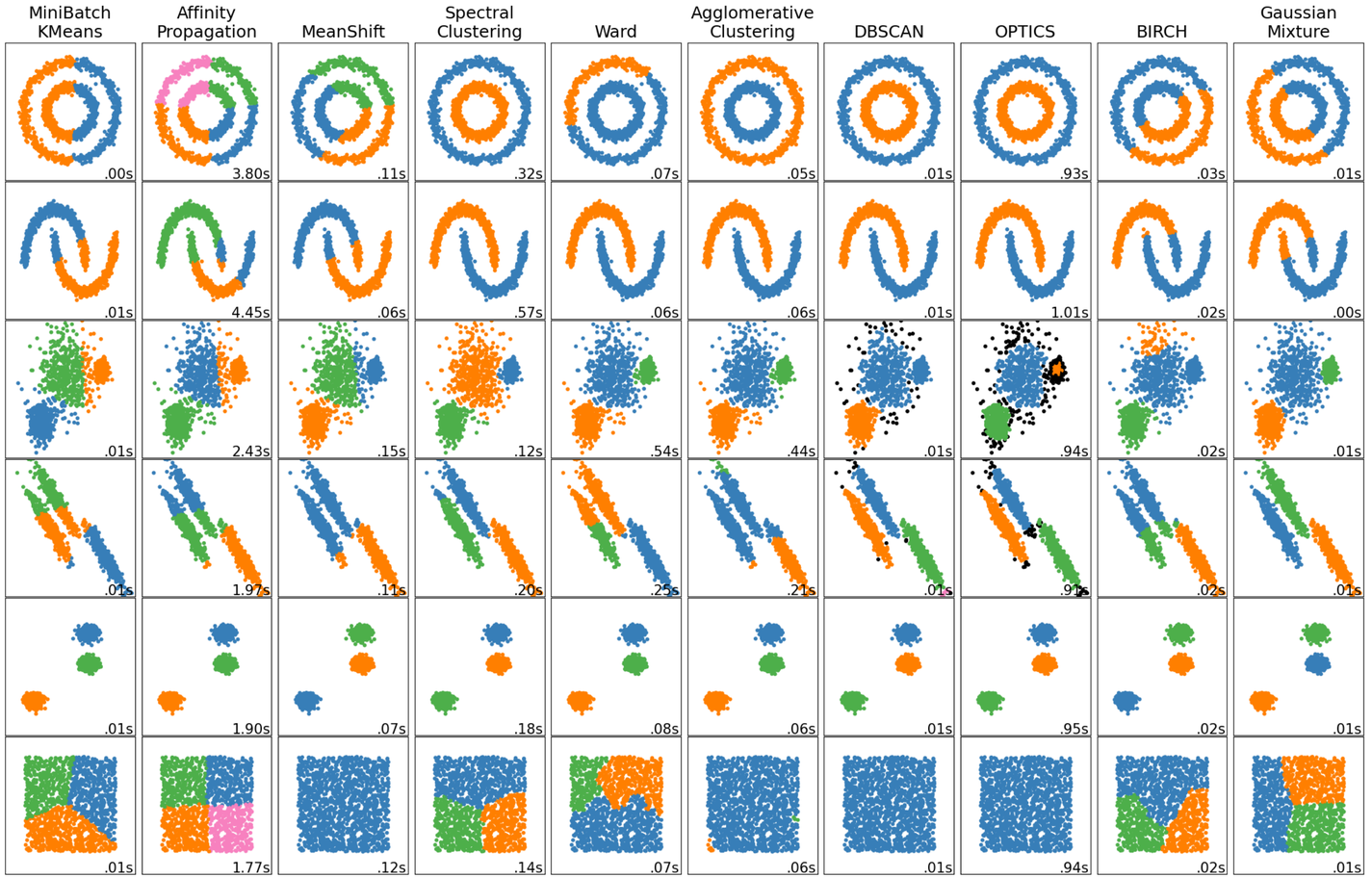
- Divisive clustering : προσέγγιση από την κορυφή προς τη βάση
  - Όλα τα σημεία ξεκινούν σε ένα σύμπλεγμα και οι διαιρέσεις εκτελούνται αναδρομικά καθώς κινούμαστε προς τα κάτω της ιεραρχίας.

Παραγωγή: μια δομή σε σχήμα δέντρου γνωστή ως δενδρόγραμμα



This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Clustering στο scikit-learn



K-means στο scikit-learn χρησιμοποιεί μια καλύτερη προεπιλεγμένη στρατηγική αρχικοποίησης (k-means++)

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

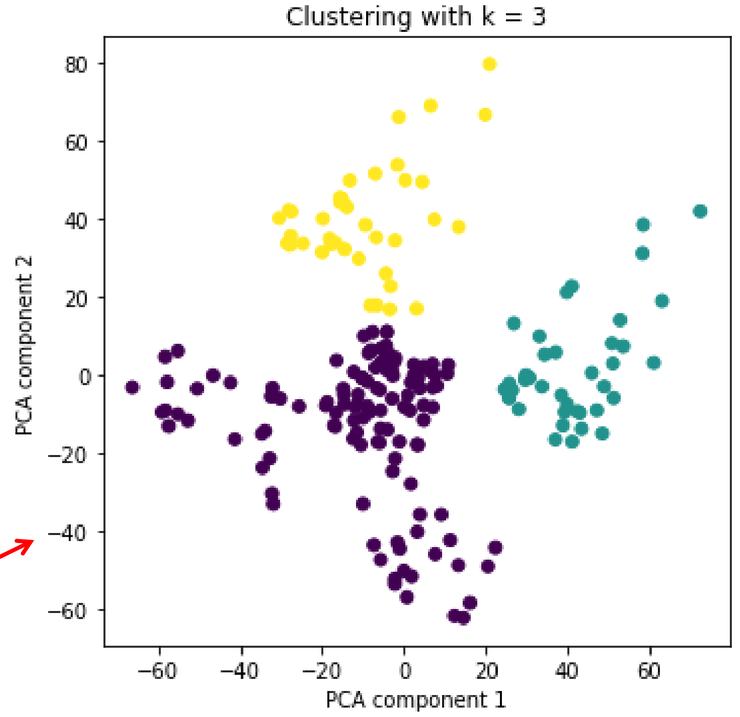
Clustering δεδομένων υψηλής διάστασης

Δεδομένα υψηλής διάστασης: δεκάδες έως χιλιάδες διαστάσεις
Παραδείγματα: εικόνες, έγγραφα κειμένου

Το πρόβλημα: Καθώς η διαστασιότητα αυξάνεται, τα σημεία δεδομένων γίνονται πιο αραιά, έτσι, η απόσταση από το πιο απομακρυσμένο σημείο γίνεται πιο κοντά στην απόσταση του πλησιέστερου σημείου.

Έτσι, οι μετρήσεις απόστασης στους αλγόριθμους ομαδοποίησης καθίστανται άνευ νοήματος

Άμβλυνση: Μείωση της διάστασης των δεδομένων και στη συνέχεια συστάδα
— Επιλογή χαρακτηριστικών
— Μείωση της διάστασης π.χ. με τη χρήση PCA



This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Το clustering μπορεί να βοηθήσει στην εποπτευόμενη μάθηση

Εύρεση των αρχικών κέντρων για τα δίκτυα RBF

- Τα δίκτυα RBF με ρυθμιζόμενο κέντρος έχουν μια φάση εκκίνησης όπου επιλέγουμε τις αρχικές συντεταγμένες των κέντρων
- Στη συνέχεια, ο αλγόριθμος προχωρά στην προσαρμογή των κεντρικών συντεταγμένων χρησιμοποιώντας το Gradient Descent (GD)

- **Πώς να επιλέξετε τα αρχικά κέντρα:**
  - **Προσέγγιση 1**: Επιλογή κέντρων ομοιόμορφα τυχαία από τα σημεία δεδομένων
    - Απλό, ωστόσο, εάν τα κέντρα δεν επιλέγονται σωστά, GD μπορεί να κολλήσει στο τοπικό βέλτιστο
  - **Προσέγγιση 2**: Χρήση ομαδοποίησης
    - Οδηγεί σε καλύτερες αρχικές συντεταγμένες, γεγονός που βοηθά το GD
    - Ορισμένες μέθοδοι ομαδοποίησης μπορούν ακόμη και να αποφασίσουν τον βέλτιστο αριθμό κέντρων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Το clustering μπορεί να βοηθήσει στην εποπτευόμενη μάθηση

Όταν υπάρχουν περισσότερα μη επισημασμένα παραδείγματα από τα επισημασμένα

- Η εποπτευόμενη μάθηση λειτουργεί σε επισημασμένα παραδείγματα
  - $D_{labelled} = \{(x^{(i)}, y^{(i)})\}_{i=1:l}$
- Τα μη επισημασμένα παραδείγματα, ωστόσο, είναι συχνά πιο εύκολο να συλλεχθούν
  - $D_{unlabelled} = \{(x^{(j)})\}_{j=1:u}$
- Συνήθως $u >> l$
  - Για παράδειγμα: για μια εργασία αναγνώρισης ζώων, μπορεί να έχουμε άφθονες εικόνες για διάφορα ζώα, αλλά μόνο πολύ λίγες ετικέτες
- Επίσης, είναι συχνά δαπανηρή η επισήμανση
  - Για παράδειγμα: ιατρικές εικόνες, πρέπει να προσλάβουμε ειδικούς για να τις επισημάνουμε σωστά

Co-financed by the European Union
Connecting Europe Facility

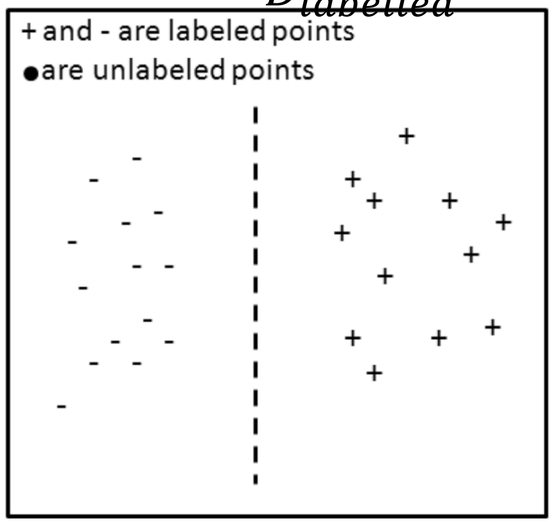
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Το clustering μπορεί να βοηθήσει στην εποπτευόμενη μάθηση

Όταν υπάρχουν περισσότερα μη επισημασμένα παραδείγματα από τα επισημασμένα

Αν λάβουμε υπόψη μόνο τα επισημασμένα παραδείγματα:

$D_{labelled}$

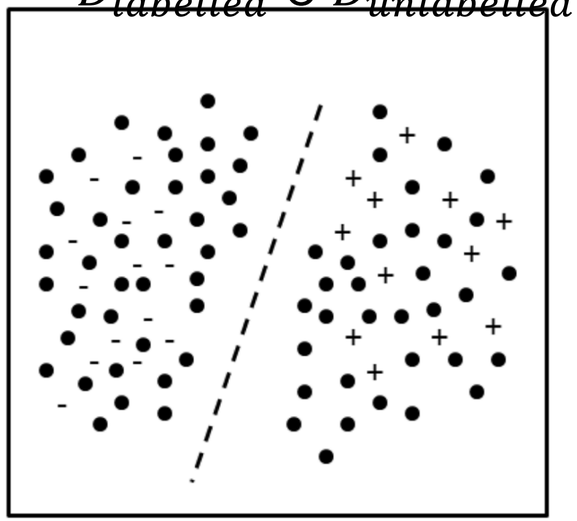


Διαφορετικό όριο απόφασης

Βελτιωμένη απόδοση γενίκευσης

Εάν θεωρήσουμε όλα τα δεδομένα ως δεδομένα κατάρτισης:

$D = D_{labelled} \cup D_{unlabelled}$



Πηγή εικόνων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επόμενη Διάλεξη

- Μείωση των διαστάσεων

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Σας ευχαριστούμε

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423