Σύγκριση διαφορετικών τεχνικών clustering

Μαρία Παππά

AM: 361

Εισαγωγή

Το Iris dataset είναι ένα διάσημο dataset. Σε αυτή την εργασία εφαρμόζουμε διαφορετικούς αλγόριθμους clustering και αξιολογούμε την επίδοσή τους. Οι αλγόριθμοι που θα αξιολογηθούν είναι οι: K-nearest neighbors (KNN) καο ο αλγόριθμος K-Means.

Δεδομένα

Το πραγματικό σετ δεδομένων που χρησιμοποιήσαμε ονομάζεται Iris. Το Iris είναι ένα κλασσικό σετ δεδομένων που χρησιμοποιείται στην επίλυση προβλημάτων μηχανικής μάθησης. Στο εσωτερικό του μπορούμε να αναγνωρίσουμε ποιο είναι το είδος ενός λουλουδιού βασιζόμενοι σε διαφορετικές μετρήσεις, όπως το μήκος και το πλάτος των πέταλων του. Αυτό το dataset περιέχει τρεις διαφορετικούς τύπους λουλουδιών. Και οι τρεις τύποι είναι είδη του Iris – setosa, versicolo και virginica. Το dataset μας δίνει 50 παραδείγματα για κάθε τύπο λουλουδιού, δηλαδή 150 παραδείγματα στο σύνολο. Παρατηρούμε ότι υπάρχουν τέσσερις παράμετροι που χρησιμοποιούνται για να περιγράψουν κάθε παράδειγμα. Αυτές οι παράμετροι είναι το μήκος και το πλάτος του φύλου κάλυκος του ανθού και του πέταλου του λουλουδιού. Οι πρώτες τέσσερις στήλες μας δίνουν τις μετρήσεις και η τελευταία στήλη μας δίνει την ετικέτα του, δηλαδή σε ποιο είδος ανθού ανήκει κάθε στήλη. Χρησιμοποιούμε το scikit-learn που παρέχει ένα σύνολο από σύνολα δεδομένων για χρήση, συμπεριλαμβανομένου και του Iris. Το dataset περιέχει το πίνακα που εμφανίζεται στο Wikipedia όπως επίσης και κάποια άλλα επιπλέον δεδομένα. Τα επιπλέον δεδομένα μας δίνουν πληροφορία για τις μετρήσεις και τα ονόματα των διαφορετικών τύπων λουλουδιών.

Μεθοδολογία και Πειράματα

Ομαδοποίηση

Ένας απλός ορισμός για την ομαδοποίηση ή συσταδοποίηση (clustering): ομαδοποίηση ονομάζεται η διαδικασία που οργανώνει πρότυπα (παρατηρήσεις, δεδομένα ή διανύσματα χαρακτηριστικών) σε ομάδες (συστάδες-clusters), όπου τα μέλη μιας ομάδας είναι παρόμοια μεταξύ τους σύμφωνα με κάποιο κριτήριο. Σκοπός είναι να προσδιοριστούν οι ομάδες που ανήκουν διάφορες ποσότητες δεδομένων, με βάση κάποια κριτήρια ομοιογένειας. Η τεχνική της ομαδοποίησης υπάγεται στην ευρύτερη κατηγορία των τεχνικών μάθησης χωρίς επίβλεψη. Η διαφορά της ομαδοποίησης δεδομένων (data

clustering) από την ταξινόμηση δεδομένων (data classification) είναι ότι, στην ταξινόμηση οι ομάδες στις οποίες θα τοποθετηθούν τα δεδομένα είναι προκαθορισμένες. Αυτό σημαίνει, ότι είναι εκ των προτέρων γνωστός ο αριθμός των ομάδων, τα ονόματα και οι ταυτότητες τους. Είναι και αυτό ένα σύστημα μάθησης μιας και οι ετικέτες που δίνονται από τα διαθέσιμα πρότυπα χρησιμοποιούνται ώστε να μάθει το σύστημα ταξινόμησης την περιγραφή κάθε κλάσης και να είναι σε θέση να ταξινομήσει ένα νέο πρότυπο. Αντίθετα, στην ομαδοποίηση δεδομένων τονίζεται ιδιαίτερα ότι οι ομάδες δεν προϋπάρχουν αλλά αποφασίζονται από τον αλγόριθμο κατά δυναμικό τρόπο. Στην ομαδοποίηση δεδομένων δηλαδή, υπάρχει ένα σύνολο δεδομένων το οποίο πρέπει να διαχειριστεί ώστε από αυτό να προκύψουν δυναμικά οι ομάδες (είναι δηλαδή data driven). Σκοπός είναι να δημιουργηθούν ομάδες, που η καθεμία από αυτές θα συγκεντρώνει ομοιογενή στοιχεία. Κάθε μία από αυτές τις ομάδες διατηρεί ένα κέντρο, συνήθως το πιο κεντρικό στοιχείο της.

Οι αλγόριθμοι κατηγοριοποίησης είναι αρκετοί. Η ερώτηση που γεννιέται τώρα σχετίζεται με το ποιος είναι ο καλύτερος. Η επίδοση των αλγορίθμων εξετάζεται με την εκτίμηση της ακρίβειας (accuracy) της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης αφού κάτι τέτοιο μας δείχνει πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Η εκτίμηση της ακρίβειας είναι επίσης θεμιτή αφού μας επιτρέπει την σύγκριση των διαφόρων αλγόριθμων κατηγοριοποίησης.

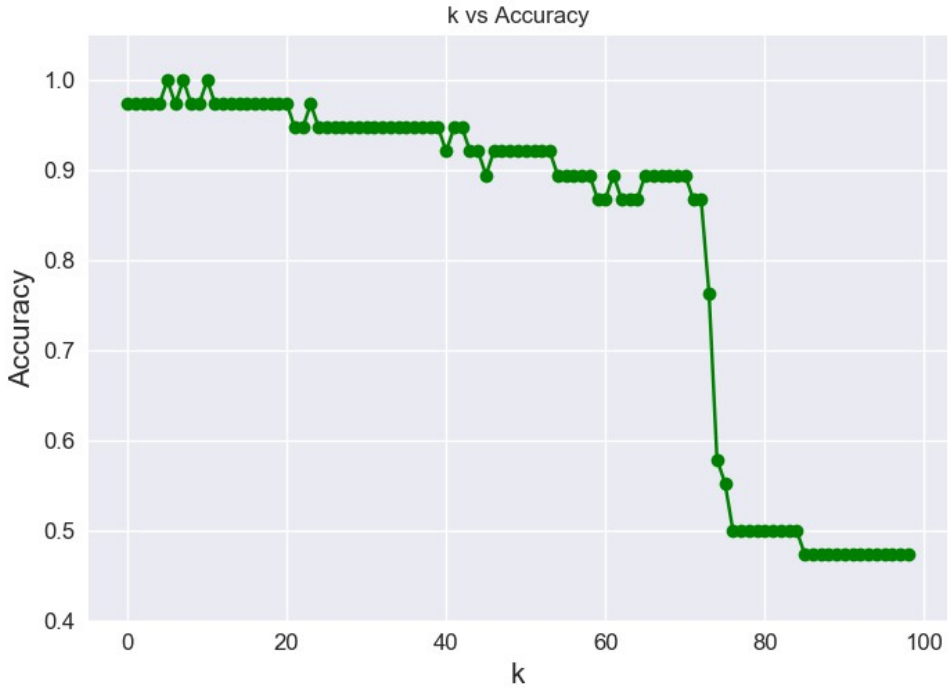
Κ κοντινότεροι γείτονες (KNN)

Ο Αλγόριθμος Κ κοντινότεροι γείτονες (K Nearest Neighbors - KNN) είναι μία πολύ γνωστή και ευρεία χρησιμοποιούμενη τεχνική κατηγοριοποίησης που στηρίζεται στη χρήση μέτρων βασισμένων στην απόσταση. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k πιο «κοντινών» στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους «γείτονες» του. Δύο ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

1. Ο ορισμός της απόστασης μεταξύ δύο στιγμιότυπων, δηλαδή μιας τιμής πάνω στο χώρο των στιγμιότυπων, που θα εκφράζει την εγγύτητα, ήαλλιώς την «ομοιότητα» μεταξύ των στιγμιότυπων.

2. Η τιμή του k.

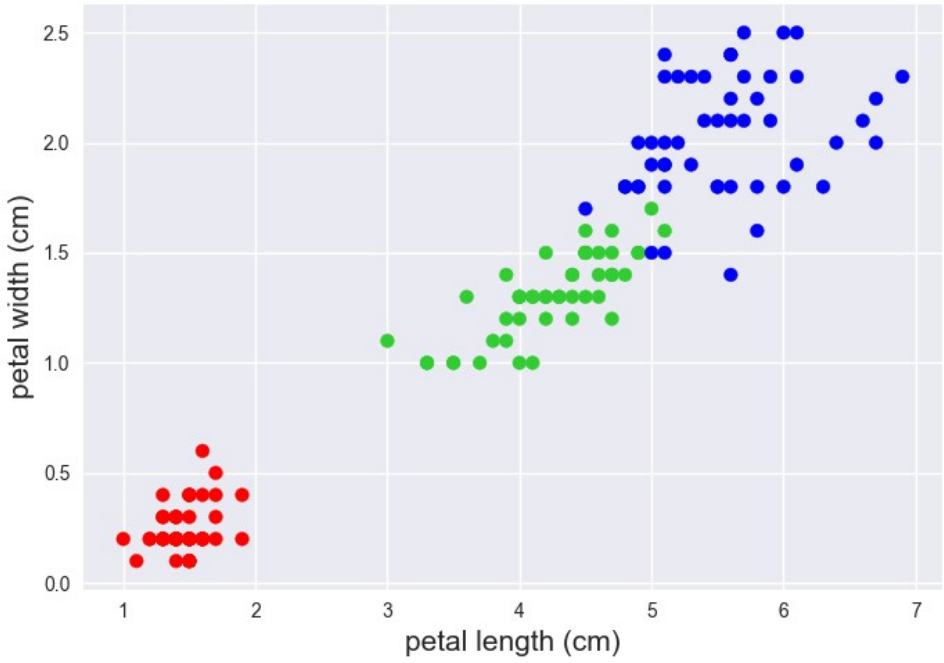
Η KNN τεχνική clustering είναι ίσως ο πιο απλός αλγόριθμος clustering. Για κάθε νέα εμφάνιση κάποιου αντικειμένου που δεν έχει ομαδοποιηθεί, υπάρχουν k πλησιέστεροι ομαδοποιημένοι γείτονες που ανήκουν σε μία ομάδα που πιθανολογείται να ομαδοποιηθεί το καινούργιο αντικείμενο. Μια βασική παράμετρος εισόδου για τον αλγόριθμο KNN είναι ο αριθμός των γειτόνων k. Για k = 1, η πρόβλεψη θα μπορούσε να είναι πολύ ευαίσθητη σε δεδομένα εκπαίδευσης λόγω του ότι το υποψήφιο αντικείμενο θα μπορούσε να ανήκει σε πολλαπλά clusters. Για k = (n_observations-1), η μεγαλύτερη δυνατή αξία, η πρόβλεψη θα είναι ίδια για όλες τις εισόδους (ίδια με την πολυπληθέστερη κατηγορία). Τυπικά, για μία ενδιάμεση k τιμή μας δίνει το καλύτερο αποτέλεσμα.



Σχήμα 1: Η ακρίβεια του αλγορίθμου KNN σε σχέση με το κ εισόδου.

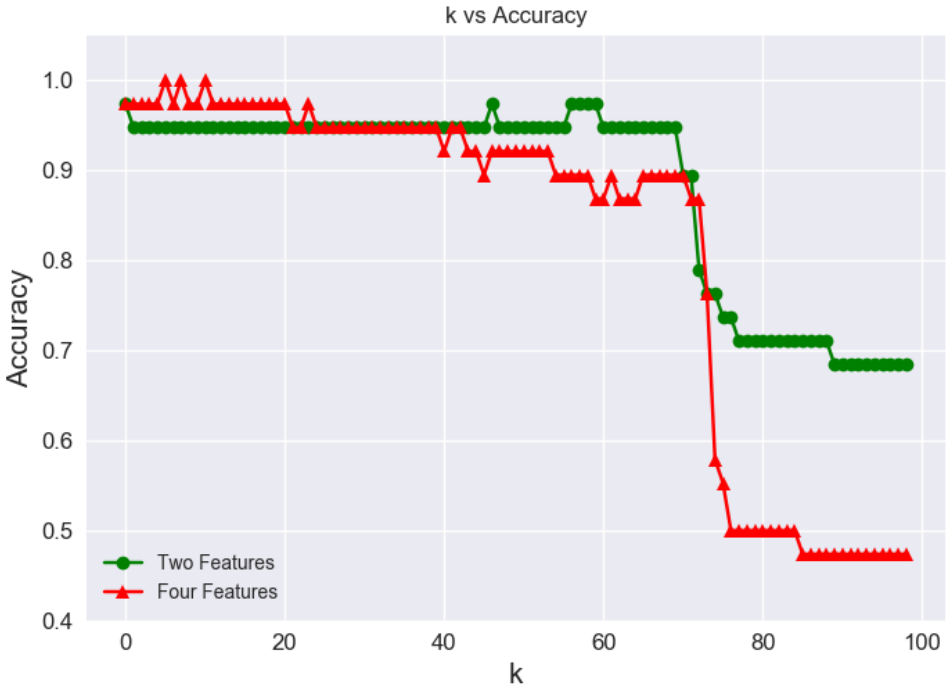
Ο αλγόριθμος KNN φαίνεται να λειτουργεί πολύ καλά ακόμη και για την τιμή k = 1. Η ομαδοποίηση στο Iris είναι σχετικά απλή και εύκολη. Η ερώτηση που γεννάται είναι αν μπορούμε να χρησιμοποιήσουμε λιγότερα χαρακτηριστικά από τα δεδομένα που προσφέρει το Iris και να αποκομίσουμε παρόμοια ισχύ στην ικανότητα να προβλέψουμε.

Υλοποιούμε το εξής πείραμα επιλέγοντας δύο από τις τέσσερις στήλες που μας προσφέρει το Iris dataset για αξιολογήσουμε αν οι κλάσεις που μας επιστρέφει ο αλγόριθμος KNN με υπολειπόμενα χαρακτηριστικά εισόδου διαχωρίζονται καλά.



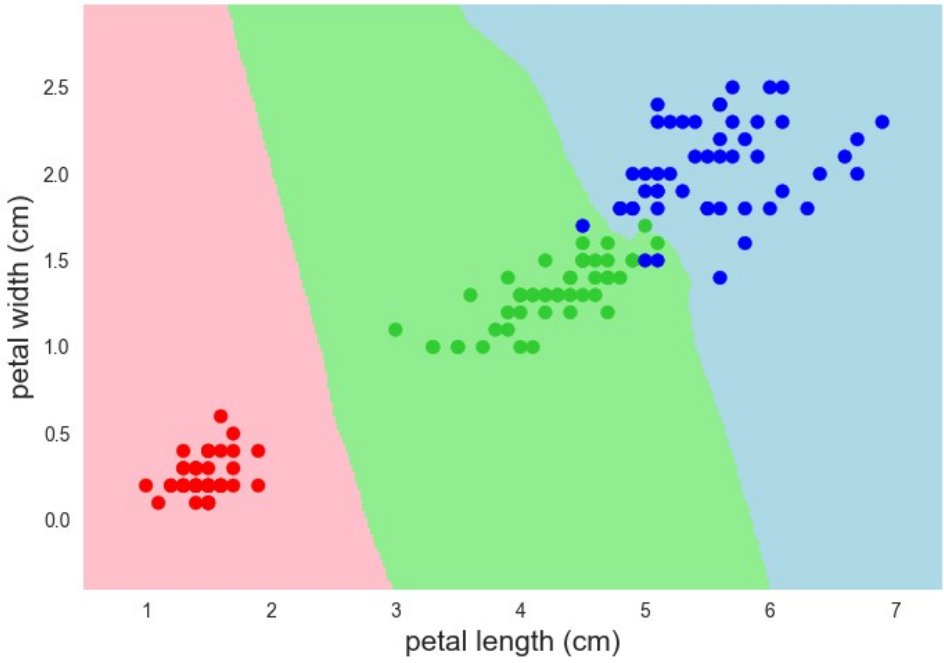
Σχήμα 2: Ομαδοποίηση κάνοντας χρήση του KNN

Στην συνέχεια εφαρμόζοντας την τεχνική KNN στο ίδιο training dataset παίρνουμε συγκριτικά αποτελέσματα που αφορούν την ακρίβεια του αλγορίθμου για δύο περιπτώσεις. Η πρώτη περίπτωση περιλαμβάνει ως πληροφορία και τα τέσσερα χαρακτηριστικά μετρήσεων που μας δίνουν τα δεδομένα του Iris. Στη δεύτερη περίπτωση έχουμε ως πληροφορία εισόδου τις δύο πρώτες στήλες του dataset που περιλαμβάνουν πληροφορίες σχετικά με το πλάτος και το μήκους του ανθού. Τα αποτελέσματα που λάβαμε ως έξοδο αντικατοπτρίζονται στο παρακάτω σχήμα.



Σχήμα 3: Η ακρίβεια σε σχέση με το k για 2 και 4 χαρακτηριστικά

Με μόνο δύο χαρακτηριστικά ως είσοδο παίρνουμε σχεδόν την ίδια ακρίβεια. Επίσης για μεγάλες τιμές του k, τις οποίες συνήθως δεν τις προτιμάμε, ο αλγόριθμος KNN που παίρνει ως είσοδο δύο χαρακτηριστικά παρουσιάζει μία εξαιρετική συμπεριφορά. Αυτό ευθύνεται στο γεγονός, ότι ο μεγαλύτερος αριθμός χαρακτηριστικών μας δίνει μεγαλύτερη πιθανότητα να έχει την επιλογή να ομαδοποιήσει σε περισσότερες ομάδες (overfitting).



Σχήμα 4 : Αξιολόγηση των cluster με βάση τα όρια που ανήκουν

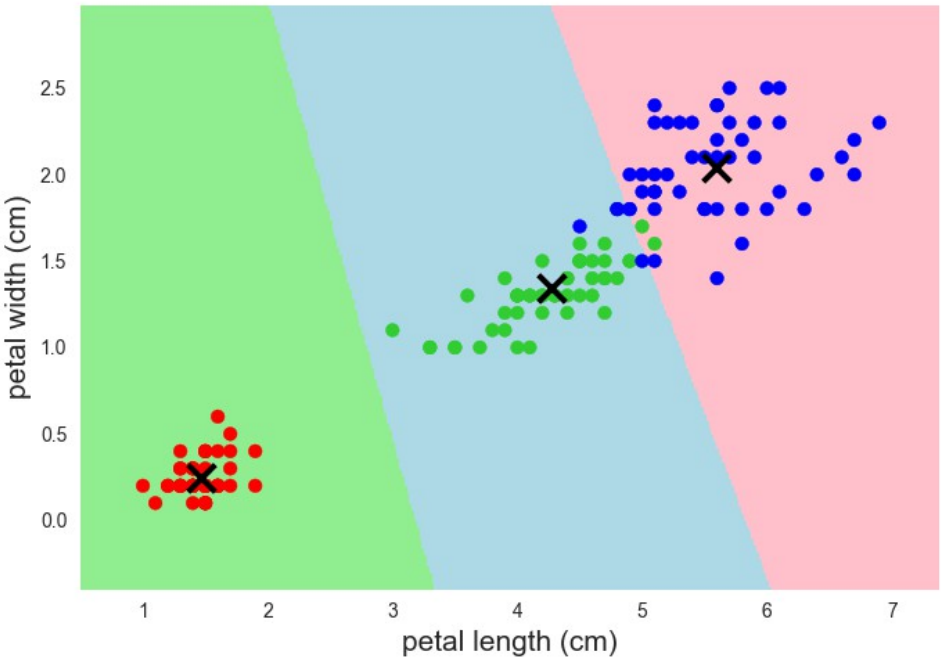
Ο αλγόριθμος ομαδοποίησης K-means

Ο αλγόριθμος k-means είναι ένας από τους δημοφιλέστερους αλγορίθμους ομαδοποίησης εξαιτίας της απλότητας της υλοποίησης του. Είναι ένας αλγόριθμος γραμμικής πολυπλοκότητας η οποία είναι της τάξης O(n) (όπου n είναι το σύνολο των στοιχείων του). Ανήκει στην ευρύτερη κατηγορία των αλγορίθμων τεχνικής μάθησης χωρίς επίβλεψη. Η διαδικασία της ομαδοποίησης ενός συνόλου δεδομένων με βάση τον k-means είναι εύκολη, αρκεί να είναι εκ των προτέρων καθορισμένος ο αριθμός (k) των clusters (ομάδων) που θα προκύψουν. Η κύρια ιδέα είναι να προσδιοριστούν αρχικά k centroids (κεντροειδή), ένα για κάθε cluster. Τα αρχικά centroids με βάση τις διαφορετικές αρχικές τους θέσεις δίνουν διαφορετικά αποτελέσματα. Αυτό σημαίνει ότι η αρχική θέση των centroids επηρεάζει το αποτέλεσμα που θα δώσει ο αλγόριθμος. Συνήθως η καλύτερη επιλογή των centroids είναι να απέχουν μεταξύ τους όσο περισσότερο γίνεται. Στο επόμενο βήμα πραγματοποιείται η επιλογή κάθε στοιχείου από το σύνολο δεδομένων και στη συνέχεια η συσχέτιση του με το κοντινότερο σε αυτό centroid. Όταν αυτή η διαδικασία ολοκληρωθεί για κάθε στοιχείο του συνόλου δεδομένων, το πρώτο βήμα έχει ολοκληρωθεί και μία πρώτη και «πρόχειρη» ομαδοποίηση έχει ήδη προκύψει. Στη συνέχεια, απαιτείται να υπολογιστούν εκ νέου τα k νέα centroids. Τα K νέα centroids αποτελούν το κέντρο βάρους για κάθε ένα cluster που προέκυψε από το προηγούμενο βήμα. Αφού λοιπόν οριστούν τα νέα k centroids, ακολουθεί και πάλι η ίδια διαδικασία ανάθεσης καθενός από τα στοιχεία του συνόλου δεδομένων στο κοντινότερο με αυτό, νέο πλέον, centroid. Έτσι, γίνεται μια επανάληψη της ίδιας διαδικασίας. Αποτέλεσμα αυτής της επανάληψης είναι ότι σε κάθε βήμα τα centroids αλλάζουν θέση (ορίζονται νέα) και τα στοιχεία ανατίθενται στο κατάλληλο cluster κάθε φορά με βάση το κοντινότερο centroid. Όταν σε κάποια επανάληψη δεν σημειωθούν

αντιμεταθέσεις στοιχείων, τότε τερματίζει η εκτέλεση του αλγορίθμου. Το αποτέλεσμα που προκύπτει είναι η ομαδοποίηση του συνόλου δεδομένων σε k συστάδες.

Ο αλγόριθμος στοχεύει να ελαχιστοποιήσει μία αντικειμενική συνάρτηση, την λεγόμενη συνάρτηση τετραγωνικού λάθους που ορίζεται ως εξής: όπου είναι ένα μέτρο απόστασης που χρησιμοποιείται για να μετρά την απόσταση κάθε στοιχείου από το centroid του κάθε cluster.

Ο k-means αλγόριθμος είναι πιο πολύπλοκος αλγόριθμος συγκριτικά με τον KNN. Απαιτεί ως είσοδο τον αριθμό k cluster. Η τιμή του k είναι γνωστή σε κάθε περίπτωση. Ο αλγόριθμος αναμένεται να κάνει ομαδοποίηση σε k=3 ομάδες. Παρόλα αυτά, η απονομή των ετικετών στα clusters είναι τυχαία και εξαρτάται από τις αρχικές τοποθεσίες των centroids.



Σχήμα 5: Ομαδοποίηση του Iris χρησιμοποιώντας τον K-means για k=3

Στο παραπάνω σχήμα οι κύκλοι είναι χρωματισμένοι με βάση τις πραγματικές κλάσεις και τα όρια που έχουν δημιουργηθεί κατά την πρόβλεψη του k-means. Τα centroids απεικονίζονται χρησιμοποιώντας τα “x”s. Παρατηρούμε ότι τα στοιχεία είναι χρωματισμένα με τυχαίο τρόπο για τρία χρώματα, που εξαρτώνται από τις αρχικές τιμές των centroids. Έχοντας μια συνολική εικόνα ο k-means ανταποκρίνεται ικανοποιητικά.
Επίσης παρατηρούμε τα όρια για τα οποία αποφασίζει ο αλγόριθμος k-means είναι λιγότερο πολύπλοκα από ότι ο KNN και ότι 4 σημεία του έχουν ομαδοποιηθεί με λάθος τρόπο. Ο αλγόριθμος k-means δεν ανταποκρίνεται καλά στο να διαχωρίζει πολύπλοκα σημεία τα οποία μπλέκονται μεταξύ τους και συνορεύουν, αυτό δεν μας εκπλήσσει ως αποτέλεσμα.

Σύνοψη

Το clustering είναι η διαδικασία κατά την οποία ένα σύνολο δεδομένων ομαδοποιείται με βάση κάποιο μέτρο ομοιότητας. Δεδομένα διαφορετικών εφαρμογών μπορεί να χρειάζεται να ομαδοποιηθούν διαφορετικά για αυτό το λόγο αποφασίζεται και ο κατάλληλος αλγόριθμος για την εκάστοτε εφαρμογή. Το Iris είναι μια συλλογή από δεδομένα που είναι από τη φύση του πολύ απλά. Αυτό σημαίνει ότι η αξιολόγηση που μπορούμε να κάνουμε είναι περιορισμένη καθώς δε μπορούμε να αναθέσουμε ετικέτες σε παραπάνω communities.

Αναφορές

[1] https://en.wikipedia.org/wiki/Iris_flower_data_set

[2] https://en.wikipedia.org/wiki/Cluster_analysis

[3] https://en.wikipedia.org/wiki/K-means_clustering