ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and Information Systems)

ΕΡΓΑΣΙΑ (PROJECT)

ΟΛΟΚΛΗΡΩΜΕΝΗ ΑΝΑΠΤΥΞΗ ΚΑΙ ΔΟΚΙΜΗ ΑΛΓΟΡΙΘΜΟΥ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ K-MEANS ΣΕ PYTHON

ΙΩΑΝΝΗΣ ΚΑΡΑΓΙΑΝΝΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

ΔΡ. ΠΑΠΑΚΩΝΣΤΑΝΤΙΝΟΥ ΣΩΤΗΡΙΟΣ

ΕΠΙΒΛΕΠΩΝ ΒΟΗΘΟΣ ΔΙΔΑΣΚΑΛΙΑΣ

ΑΡΓΥΡΙΟΥ ΑΘΑΝΑΣΙΟΣ

ΑΘΗΝΑ

2021

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Οικονομικών Επιστημών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος

Ο υποψήφιος βεβαιώνει ότι η υποβληθείσα εργασία είναι προσωπική εκτός από τα σημεία όπου γίνεται αναφορά στις εργασίες άλλων.

Ιωάννης Καραγιάννης

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον καθηγητή κ. Ιωάννη Δημητρίου, διότι μέσα από την παρακολούθηση των μαθημάτων που μας δίδαξε, προέκυψε το ενδιαφέρον μου για τους αλγορίθμους Μηχανικής Μάθησης, με το αποτέλεσμα να είναι η εκπόνηση της εργασίας αυτής. Θα ήθελα επίσης να ευχαριστήσω τον βοηθό διδασκαλίας Αργυρίου Αθανάσιο, για την πολύτιμη βοήθεια και καθοδήγησή του, καθ' όλη τη διάρκεια εκπόνησης της εργασίας. Ένα μεγάλο ευχαριστώ στον καθηγητή κ. Παπακωνσταντίνου Σωτήριο για τη συνεισφορά του στην ολοκλήρωση της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω θερμά την οικογένειά μου, για την συμπαράστασή τους όλο αυτό το διάστημα.

Με εκτίμηση,

Ιωάννης Καραγιάννης

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΚΟΝΕΣ

ΔΙΑΓΡΑΜΜΑΤΑ

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια, συμβαδίζοντας με το σύγχρονο τρόπο ζωής, η ανθρωπότητα παράγει ολοένα και περισσότερα δεδομένα, λόγω της εξάπλωσης των κοινωνικών δικτύων, των νέων μέσων επικοινωνίας, της εξέλιξης της ψηφιακής φωτογραφίας αλλά και του Διαδικτύου των Πραγμάτων. Η αξιοποίηση των δεδομένων αυτών ώστε να καταλήξουν σε χρήσιμη πληροφορία και έπειτα σε γνώση, αποτελεί αντικείμενο αρκετών επιστημών. Η Μηχανική Μάθηση, η οποία εντάσσεται στο πεδίο της Τεχνητής Νοημοσύνης, είναι μια από αυτές.

Στη συγκεκριμένη εργασία θα ασχοληθούμε συγκεκριμένα με τη μέθοδο συσταδοποίησης k-means, η οποία ανήκει στην κατηγορία αλγορίθμων μη επιτηρούμενης μάθησης. Θα αναφερθούμε σε θεωρητικό επίπεδο στις διάφορες μεθόδους Μηχανικής Μάθησης και στη συνέχεια θα αναλύσουμε τον αλγόριθμο k-means με τους περιορισμούς του αλλά και κάποια πρακτικά παραδείγματα εφαρμογής. Στη συνέχεια θα περιγράψουμε τον αλγόριθμο που αναπτύχθηκε στα πλαίσια της εργασίας σε γλώσσα προγραμματισμού python, ο οποίος παρέχει οπτικοποίηση και αναλυτική πληροφορία στο χρήστη για τα ενδιάμεσα στάδια και τις επαναλήψεις του k-means, πράγμα που τον καθιστά κατάλληλο για σκοπούς εκπαίδευσης ή επίδειξης.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, k-means, συσταδοποίηση, python

SUMMARY

In recent years, keeping pace with the modern way of life, humanity is producing more and more data, due to the spread of social networks, new communication methods, the evolution of digital photography and the Internet of Things. Utilizing this data to produce useful information and make it into knowledge is the subject of several sciences. Machine Learning which is a field of Artificial Intelligence is one of them.

In this paper we will deal specifically with the k-means clustering method, which belongs to the category of unsupervised learning algorithms. We will refer at a theoretical level to the various methods of Machine Learning and then we will analyze k-means algorithm with its limitations and some practical examples of application. Next, we will refer to the algorithm developed in python programming language, in the context of this project, which provides visualization and analytical information to the user for the intermediate stages and iterations of k-means, which makes it suitable for training or demonstration purposes.

Keywords: Artificial Intelligence, Machine Learning, k-means, clustering, python

1 Εισαγωγή

Στην παρούσα εργασία δημιουργήθηκε εκ του μηδενός ένας αλγόριθμος που εκτελεί τη μέθοδο συσταδοποίησης (clustering) k-means, σε γλώσσα προγραμματισμού python. Αυτή τη στιγμή υπάρχουν διαθέσιμες κάποιες γνωστές βιβλιοθήκες που περιέχουν έτοιμες εντολές που εκτελούν αλγόριθμο k-means σε python, όπως για παράδειγμα η scikit-learn (scikit-learn, 2020) ή η scipy (SciPy.org, 2020). Συνήθως η χρήση αυτών των έτοιμων πακέτων απαιτεί από τον χρήστη να είναι εξοικειωμένος με τη χρήση της Python, για να καταφέρει να τα χρησιμοποιήσει αποτελεσματικά και αποδοτικά, αλλά και να γνωρίζει τον τρόπο που λειτουργεί ο αλγόριθμος k-means ώστε να φτάσει σε κάποιο αποτέλεσμα.

Είναι δεδομένο πως τα τελευταία χρόνια η διάδοση των αλγορίθμων Μηχανικής Μάθησης (Machine Learning) αυξάνεται με ταχείς ρυθμούς, κυρίως λόγω του ρόλου που διαδραματίζει πλέον η συλλογή και επεξεργασία δεδομένων σε διάφορους τομείς της οικονομίας. Η μέθοδος k-means ανήκει σε αυτή την ομάδα αλγορίθμων και μάλιστα είναι από τις σχετικά απλές και εύκολα κατανοητές μεθόδους συσταδοποίησης δεδομένων χωρίς επίβλεψη (unsupervised).

Για τους παραπάνω λόγους, επιλέχθηκε η υλοποίηση της μεθόδου k-means σε python με σκοπό να προσφέρει μεγαλύτερη διαφάνεια και πληροφόρηση σε σχέση με κάποια έτοιμη βιβλιοθήκη, ιδιαίτερα στους χρήστες που δεν έχουν μεγάλη εξοικείωση με τη χρήση της python, αλλά χρειάζονται την πληροφόρηση αυτή. Ως εκ τούτου, ο αλγόριθμος που αναπτύχθηκε, καθίσταται ιδανικός για εκπαιδευτικούς σκοπούς, ως μέσο επίδειξης, μιας και όπως θα δούμε είναι σε θέση να παρέχει βήμα προς βήμα ενημέρωση για τα αποτελέσματά του σε κάθε επανάληψη, συνοδευόμενα από γραφικές παραστάσεις και αποτελέσματα υπολογισμών.

Στο τρίτο κεφάλαιο θα αναφερθούμε συνοπτικά στη Μηχανική Μάθηση και τις κατηγορίες των αλγορίθμων ενώ στο τέταρτο κεφάλαιο θα περιγράψουμε τον αλγόριθμο k-means. Έπειτα, στο πέμπτο κεφάλαιο θα αναλύσουμε τον αλγόριθμο που αναπτύχθηκε στα πλαίσια της εργασίας. Στην ενότητα 5.1 υπάρχουν αναλυτικές οδηγίες για τη χρήση του αλγορίθμου. Κλείνοντας θα αναφέρουμε κάποια συμπεράσματα που προέκυψαν από την εργασία.

2 Μεθοδολογία και Βιβλιογραφική επισκόπηση

Η ανάπτυξη του κώδικα πραγματοποιήθηκε σε εικονικό περιβάλλον jupyter lab version 2.2.8 με τη χρήση της γλώσσας προγραμματισμού Python version 3.8.2 (32bit). Επιπλέον, χρησιμοποιήθηκαν οι βιβλιοθήκες numpy (v1.19.2 (NumPy, 2020)), pandas (v1.1.2) (Pandas, 2020), matplotlib (v3.3.2) (matplotlib, 2020), seaborn (v0.11.0) (seaborn, 2020), scikit-learn (v0.23.2) (scikit-learn, 2020) και warnings.py (github, n.d.).

Θα πρέπει να αναφερθεί ότι ο κώδικας που υλοποιεί τον αλγόριθμο, πιθανόν να μην είναι ο πλέον αποδοτικός όσον αφορά τον χρόνο εκτέλεσης ή της γενικότερης αξιοποίησης των υπολογιστικών πόρων. Επιπλέον, υπολείπεται της αποδοτικότητας των ευρέως διαδεδομένων αλγορίθμων (όπως π.χ. των βιβλιοθηκών σε γλώσσα python, scikit-learn (scikit-learn, 2020) ή scipy (SciPy.org, 2020)), οι οποίοι μέσα από την πολυετή χρήση και τη συμβολή ανατροφοδότησης (feedback) από χρήστες έχουν φτάσει σε υψηλό επίπεδο λειτουργίας και ευχρηστίας. Έμφαση δόθηκε στην παροχή όσο το δυνατόν περισσότερης πληροφορίας στον αρχάριο χρήστη σε κάθε βήμα εκτέλεσης του αλγορίθμου και στην δυνατότητα προσομοίωσης, της μεθοδολογίας εκτέλεσής του βήμα προς βήμα σε κάποιο υπολογιστικό περιβάλλον (π.χ. excel). Επιπλέον, δίνεται η δυνατότητα για επέμβαση στον αλγόριθμο και πιθανή βελτίωση και εξέλιξή του.

3 Αλγόριθμοι Μηχανικής Μάθησης (Machine Learning)

Η Μηχανική Μάθηση αποτελεί ένα υποσύνολο της Τεχνητής Νοημοσύνης και σχετίζεται κυρίως με ανάπτυξη αλγορίθμων που επιτρέπουν σε έναν υπολογιστή να μαθαίνει με τη βοήθεια δεδομένων, να βελτιώνει τις επιδόσεις του καθώς και να προβλέπει αποτελέσματα, χωρίς να έχει προγραμματιστεί ρητά για αυτό. Ο όρος επινοήθηκε το 1959 από τον Arthur Samuel. Αρκετά αργότερα και συγκεκριμένα το 1997, ο Mitchell πρότεινε ένα νέο ορισμό:

«*Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία Ε ως προς μια κλάση εργασιών Τ και ένα μέτρο επίδοσης Ρ, αν η επίδοσή του σε εργασίες της κλάσης Τ, όπως αποτιμάται από το μέτρο Ρ, βελτιώνεται με την εμπειρία Ε*» (Mitchell, 1997).

Τα μοντέλα και οι προβλέψεις των αλγορίθμων βασίζονται σε δεδομένα και η ακρίβεια της πρόβλεψης επηρεάζεται από το πλήθος αυτών, δηλαδή, ένας μεγάλος όγκος δεδομένων θα βοηθήσει ώστε να καταλήξουμε σε ένα μοντέλο με μεγαλύτερη ακρίβεια στις προβλέψεις του. Τα μοντέλα αυτά είναι σε θέση να καταλήξουν σε συμπεριφορά που ο συντάκτης τους δεν είχε αρχικά προβλέψει.

Οι αλγόριθμοι Μηχανικής Μάθησης συνήθως κατατάσσονται σε 3 βασικές κατηγορίες τις οποίες θα αναφέρουμε παρακάτω.

3.1 Επιτηρούμενη μάθηση ή μάθηση με επίβλεψη (Supervised learning)

Προς απλοποίηση της επεξήγησης θα θεωρήσουμε ένα απλό σύστημα Μηχανικής Μάθησης που δέχεται δεδομένα ως είσοδο και παράγει κάποια έξοδο ως αποτέλεσμα. Αν διαθέτουμε ήδη ένα σύνολο δεδομένων που περιέχει κάποια έξοδο για κάθε δεδομένη είσοδο, τότε η μάθηση με αυτόν τον τρόπο ονομάζεται επιτηρούμενη. Δηλαδή, γνωρίζουμε από πριν τα αποτελέσματα της διαδικασίας και προσπαθούμε να εκπαιδεύσουμε (train) τον αλγόριθμό μας. Μόλις έχουμε εκπαιδεύσει τον αλγόριθμο μπορούμε να τον εφαρμόσουμε σε ευρύτερο σύνολο δεδομένων.

Ένα δημοφιλές παράδειγμα της συγκεκριμένης κατηγορίας είναι η κατάταξη μιας εικόνας, π.χ. σκύλος ή γάτα; Έχοντας τροφοδοτήσει τον αλγόριθμό μας με όσο το δυνατόν περισσότερες εικόνες με σκύλους και γάτες κατά τη διαδικασία εκπαίδευσης, μπορούμε να του παρουσιάσουμε μια νέα εικόνα και να την κατηγοριοποιήσει.

3.2 Μη επιτηρούμενη μάθηση ή μάθηση χωρίς επίβλεψη (Unsupervised learning)

Αντίθετα από την επιτηρούμενη μάθηση, στη μη επιτηρούμενη τα δεδομένα μας δεν είναι κατηγοριοποιημένα (unlabeled data). Δεν παρέχουμε δηλαδή, στον αλγόριθμο κάποια πληροφόρηση και πρέπει ο ίδιος να αποφασίσει ανακαλύπτοντας συσχετίσεις ή / και μοτίβα.

Στο παραπάνω παράδειγμα θα τροφοδοτούσαμε τον αλγόριθμο με εικόνες από σκύλους και γάτες, χωρίς όμως να του δώσουμε τη σωστή κατηγορία της κάθε εικόνας. Ο αλγόριθμος θα μπορούσε να διακρίνει τις δύο κατηγορίες με τα διαφορετικά χαρακτηριστικά και να οργανώσει τις εικόνες σε συστάδες (clusters), χωρίς όμως να δώσει ονομασίες στις κατηγορίες παρά μόνο συστάδα 1 και συστάδα 2. Αν στη συνέχεια του παρουσιάσουμε μια νέα εικόνα, θα την κατηγοριοποιήσει με βάση τη συστάδα με τα πλησιέστερα χαρακτηριστικά.

3.3 Ενισχυτική μάθηση (Reinforcement learning)

Σε αυτή την περίπτωση ο αλγόριθμος μαθαίνει μέσω αλληλεπίδρασης με το περιβάλλον του, προς την επίτευξη κάποιου τελικού στόχου. Πολλές φορές μπορεί να υπάρχει σύστημα ανταμοιβής ή επιβράβευσης χωρίς όμως να είναι γνωστό στον αλγόριθμο. Εφαρμογές της συγκεκριμένης κατηγορίας μπορούμε να συναντήσουμε στη βιομηχανία (ρομποτικά συστήματα), σε προσωποποιημένες προτάσεις στις διαδικτυακές πλατφόρμες (π.χ. πρόταση για μουσική παρόμοια με τις προτιμήσεις μας), στον τομέα παιχνιδιών (π.χ. ο δημοφιλής αλγόριθμος της Google, AlphaGo (ALPHABET INC, 2021)) κ.α..

4 Αλγόριθμος k-means

Η μέθοδος k-means ανήκει στην κατηγορία αλγορίθμων μη επιτηρούμενης μάθησης. Ο όρος k-means χρησιμοποιήθηκε για πρώτη φορά από τον MacQueen (MacQueen, 1967). Ωστόσο ο αλγόριθμος είχε ήδη χρησιμοποιηθεί από τον Lloyd το 1957 ο οποίος το δημοσίευσε αρκετά χρόνια αργότερα (S. Lloyd, 1982), αλλά και από τον Forgy (Forgy, 1965) λίγα χρόνια νωρίτερα. Αποτελεί μέθοδο ομαδοποίησης ή συσταδοποίησης και πρόκειται για έναν από τους παλαιότερους και περισσότερο χρησιμοποιούμενους αλγόριθμους αυτής της κατηγορίας. Στην ουσία κατατάσσει *n* παρατηρήσεις σε *k* αριθμό συστάδων, με την κάθε παρατήρηση να ανήκει στη συστάδα *C* με το πλησιέστερο κέντρο μ_j (centroid).

Ο αριθμός των συστάδων *k* αποτελεί είσοδο του αλγορίθμου και καθορίζεται από το χρήστη. Για να ξεκινήσει ο αλγόριθμος τα αρχικά κέντρα των συστάδων επιλέγονται τυχαία, είτε από σημεία των παρατηρήσεων είτε από τυχαία σημεία εντός του ίδιου χώρου. Ωστόσο, έχουν αναπτυχθεί κάποιες μέθοδοι που μας κατευθύνουν με βάση διάφορα κριτήρια. Μια από τις πιο γνωστές και διαδεδομένες μεθόδους αποτελεί το διάγραμμα elbow chart, το οποίο μας παρέχει μια ένδειξη της εγγύτητας των παρατηρήσεων από τα κέντρα για ένα προκαθορισμένο εύρος συστάδων.

Η εγγύτητα των παρατηρήσεων από τα κέντρα υπολογίζεται συνήθως με βάση την Ευκλείδεια απόσταση (μπορεί να διαφέρει ανάλογα με τον τύπο του προβλήματος που αντιμετωπίζουμε), οπότε με αυτό τον τρόπο ελαχιστοποιείται το εντός συστάδας άθροισμα των τετραγώνων των αποστάσεων (Within Cluster Sum of Squares - WCSS) ή αδράνεια (inertia, αν θεωρήσουμε την κάθε συστάδα μια μάζα):

$$\sum_{i=0}^{n} \min_{\mu_j \in C} || x_i - \mu_j ||^2$$

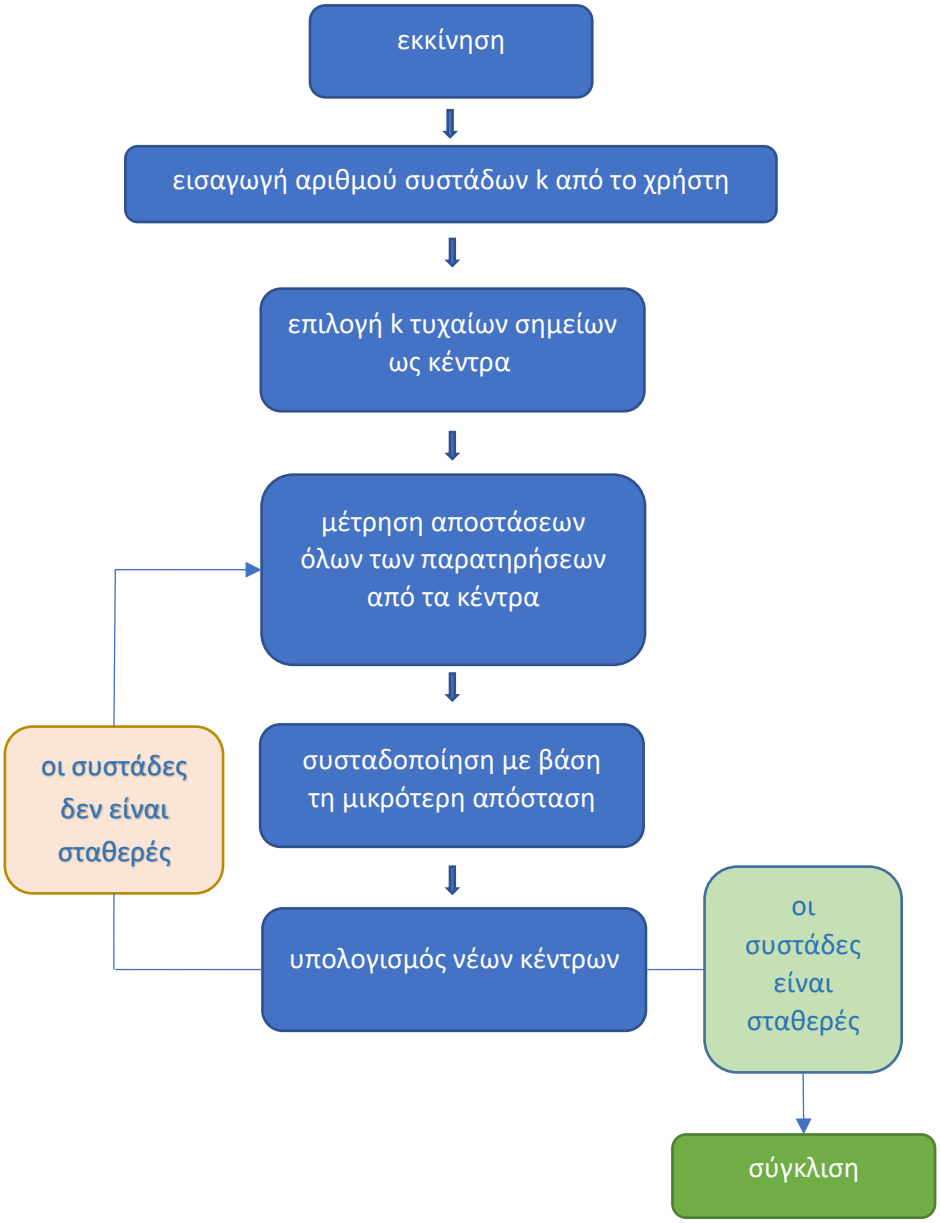
Ο τρόπος υπολογισμού της απόστασης δεν είναι πάντα η Ευκλείδεια απόσταση αλλά καθορίζεται από τη φύση του προβλήματος που θέλουμε να επιλύσουμε (π.χ. Manhattan ή City Block για προβλήματα αποστάσεων σε αστικό οδικό δίκτυο). Κάθε φορά που γίνεται ανάθεση των σημείων στο κοντινότερο κέντρο τους, υπολογίζουμε εκ νέου το κέντρο της κάθε συστάδας το οποίο είναι ο μέσος (mean) των σημείων της, χωρίς όμως να απαιτείται να είναι κάποιο από τα σημεία της. Έπειτα αναθέτουμε εκ νέου τις παρατηρήσεις στο πλησιέστερο κέντρο που τους αντιστοιχεί.

Η διαδικασία αυτή συνεχίζεται με τη μορφή επαναλήψεων (iterations) έως ότου ο αλγόριθμος είτε να συγκλίνει (convergence), (δηλαδή να μην υπάρχει πλέον μετατόπιση των

κέντρων ή ανάθεση των παρατηρήσεων σε νέα κέντρα), είτε μέχρι να φτάσει τις μέγιστες καθορισμένες επαναλήψεις (αν έχουν καθοριστεί από το χρήστη).

Δεδομένου ότι ο αλγόριθμος βασίζει την απόφασή του στις αποστάσεις, τα δεδομένα που θα εισάγουμε προς συσταδοποίηση θα πρέπει να τεθούν στην ίδια κλίμακα (scaling), ώστε τα μεγέθη προς εξέταση να είναι συγκρίσιμα. Σε διαφορετική περίπτωση μπορεί να μειωθεί η απόδοση του αλγορίθμου και να μην είναι εύκολο να αναπαρασταθούν γραφικά τα αποτελέσματα.

Όπως αναφέραμε παραπάνω ο αλγόριθμος προσπαθεί στην ουσία να ελαχιστοποιήσει το εντός κάθε συστάδας άθροισμα των τετραγώνων των αποστάσεων (Within Cluster Sum of Squared Error - WCSSE). Αν προσθέσουμε όλα αυτά τα αθροίσματα έχουμε μια εκτίμηση του σφάλματος (Sum of Squared - SSE). Το σφάλμα αυτό μειώνεται όσο αυξάνεται ο αριθμός των συστάδων k (overfitting) και όταν ο αριθμός των συστάδων $k = n$ (πλήθος παρατηρήσεων) τότε το σφάλμα είναι ίσο με μηδέν. Σε αυτή την περίπτωση, κάθε συστάδα περιλαμβάνει μια μόνο παρατήρηση, η οποία αποτελεί το κέντρο της, οπότε δεν υφίσταται σφάλμα.

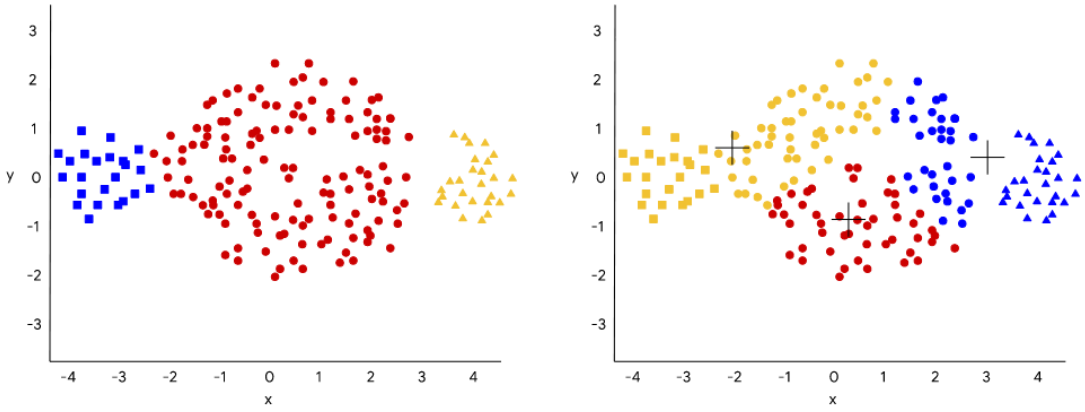


Διάγραμμα 1 – Απλοποιημένη αναπαράσταση του αλγορίθμου k-means

4.1 Μειονεκτήματα και περιορισμοί της μεθόδου k-means

Με βάση τα όσα αναφέραμε παραπάνω γίνεται αντιληπτό ότι υπάρχουν κάποια μειονεκτήματα και περιορισμοί της μεθόδου.

- Ο αλγόριθμος τείνει να μην αποδίδει επαρκώς σε δεδομένα που αποτελούνται από συστάδες διαφορετικής πυκνότητας και μεγέθους (π.χ. μη σφαιρικές συστάδες) όπως για παράδειγμα στο Διάγραμμα 2. Μια μέθοδος που έχει αναπτυχθεί με βάση την k-means είναι η elliptical k-means που σχηματίζει επιμηκυμένες συστάδες (Poggio, 1994).



Διάγραμμα 2 – Ένα από τα βασικότερα μειονεκτήματα της μεθόδου k-means (Machine Learning Crash Course, 2021)

- Η αρχική επιλογή των κέντρων μπορεί να επηρεάσει το αποτέλεσμα, για παράδειγμα ο αλγόριθμος να συγκλίνει σε κάποιο τοπικό ελάχιστο το οποίο να μην αποτελεί βέλτιστη λύση. Αυτό σημαίνει ότι θα πρέπει να επιλέξουμε να επαναλάβουμε τον αλγόριθμο και για διαφορετικά αρχικά κέντρα. Ωστόσο, έχουν αναπτυχθεί διάφορες μέθοδοι, όπως για παράδειγμα η k-means++ σύμφωνα με την οποία τα αρχικά κέντρα τίθενται με κάποια βαρύτητα. Κατά τα λοιπά είναι όμοια με την k-means αλλά καταλήγει σε αποτελέσματα γρηγορότερα και με μεγαλύτερη ακρίβεια από τη δεύτερη (David Arthur, 2007).

- Ο αλγόριθμος είναι ευαίσθητος σε έκτοπα σημεία (outliers). Πολλές φορές η ύπαρξη ακραίων τιμών μπορεί να δημιουργήσει μια επιπλέον συστάδα που να περιέχει τις τιμές αυτές. Στην περίπτωση αυτή, τίθεται ένα ελάχιστο όριο παρατηρήσεων ανά συστάδα ώστε αν η συστάδα δεν είναι αρκετά μεγάλη, η διαδικασία να επαναληφθεί (Ameet V Joshi, 2020) ή εναλλακτικά να αποκλειστούν κάποιες ακραίες παρατηρήσεις εξαρχής.

- Η επιλογή του αριθμού των συστάδων k γίνεται από τον χρήστη. Δεδομένης της φύσης του μη επιτηρούμενου αλγορίθμου, αυτό σημαίνει ότι δεν υπάρχει σωστό ή λάθος, ο

χρήστης θα πρέπει μέσα από τις δοκιμές να καταλήξει στο βέλτιστο κατά την κρίση του αριθμό. Σε αυτό το πεδίο έχουν αναπτυχθεί αλγόριθμοι που βοηθούν στην επιλογή του αριθμού συστάδων όπως η μέθοδος elbow (inertia) (Thorndike, 1953), η μέθοδος silhouette (J.Rousseeuw, 1987) και η μέθοδος gap statistic (Robert Tibshirani, 2000).

- Όπως σε πολλούς αλγορίθμους Μηχανικής Μάθησης, τα προς επεξεργασία δεδομένα θα πρέπει να τεθούν στην ίδια κλίμακα (με κάποιο scaler). Όπως αναφέραμε και προηγουμένως, αν τα δεδομένα δεν βρίσκονται στην ίδια κλίμακα, τότε μπορεί να μειωθεί σημαντικά η απόδοση του αλγορίθμου και να μην είναι εύκολο να αναπαρασταθούν γραφικά τα αποτελέσματα.

4.2 Παραδείγματα εφαρμογής αλγορίθμου k-means

Παρακάτω θα αναφέρουμε κάποια ενδεικτικά παραδείγματα της εφαρμογής του αλγορίθμου k-means που μπορούμε να συναντήσουμε στην καθημερινότητα ή σε επίπεδο μελετών.

- Συμπίεση δεδομένων με απώλειες (lossy), στην οποία δεχόμαστε ορισμένα σφάλματα στην ανακατασκευή σε αντάλλαγμα για υψηλότερα επίπεδα συμπίεσης από ότι μπορεί να επιτευχθεί στην περίπτωση χωρίς απώλειες (lossless).

- Κατάτμηση εικόνων, ανάλογα με το χρώμα (Siddheswar Ray), σχήμα ή περιεχόμενο που διαθέτουν, ή ακόμη και για να διαχωρίσουν τμήματα ενδιαφέροντος της εικόνας από το παρασκήνιο (Nameirakpam Dhanachandra, 2015). Επίσης, η μέθοδος k-means μπορεί να βελτιώσει την τμηματοποίηση εικόνων από μαγνητικούς τομογράφους (H.P. Ng, 2006)

- Ταξινόμηση εγγράφων βασισμένη σε ετικέτες, θέματα και γενικό περιεχόμενο. Πρόκειται για ένα συχνό πρόβλημα που απαιτεί λύση και η μέθοδος k-means είναι αρκετά διαδεδομένη σε αυτόν τον τομέα. Τα έγγραφα θα πρέπει προηγουμένως να έχουν μετατραπεί σε διανύσματα (vectors), τα οποία στη συνέχεια χωρίζονται σε συστάδες.

- Τμηματοποίηση πελατών/αγοράς. Η τμηματοποίηση αυτή μπορεί να γίνει βάσει ιστορικού αγορών, τα ενδιαφέροντα ή τη δραστηριότητα των πελατών (Bacila, 2012) ή ακόμη και με βάση την κερδοφορία που αυτοί αποφέρουν (Arumawadu, 2015).

- Βελτιστοποίηση τρόπου αποστολής ανάλογα με τις ιδιότητες του αντικειμένου (βάρος/όγκος) και το όχημα μεταφοράς (Yudhanegara Mokhammad, 2020).

5 Ανάπτυξη αλγορίθμου k-means

Ο αλγόριθμος που αναπτύχθηκε έχει ως σκοπό να αυτοματοποιήσει την εφαρμογή της μεθόδου k-means σε περιβάλλον python, τόσο για χρήστες αρχάριους στην python όσο και στον αλγόριθμο τον ίδιο ή ακόμη και την επιστήμη των δεδομένων γενικότερα. Γι' αυτό το λόγο έχει απλοποιηθεί η εισαγωγή δεδομένων προς ανάλυση και σε κάθε επανάληψή του ενημερώνει το χρήστη με τα αποτελέσματα της συνοδευόμενα από τη γραφική αναπαράσταση των συστάδων με τα κέντρα τους.

Προαιρετικά πριν την εκκίνηση του κυρίως μέρους του αλγορίθμου, ο χρήστης μπορεί να επιλέξει τον υπολογισμό και εμφάνιση του scree plot (elbow plot), δηλαδή, της ποσότητας που προκύπτει από το άθροισμα των τετραγώνων των αποστάσεων εντός όλων των συστάδων (Within Cluster Sum of Squares – WCSS ή inertia), σε σχέση με τον αριθμό των συστάδων k. Με αυτή την επιπλέον πληροφόρηση, μπορεί ο χρήστης να αποφασίσει από πριν τον βέλτιστο – σύμφωνα πάντα με το κριτήριο WCSS – αριθμό των συστάδων k που θα εκτελέσει έπειτα στον αλγόριθμο. Σε αυτό το βήμα ο χρήστης εισάγει το μέγιστο αριθμό συστάδων k, για τον οποίο επιθυμεί να υπολογιστεί το WCSS. Για την ολοκλήρωση των υπολογισμών σε αυτό το στάδιο γίνεται η χρήση του αλγορίθμου k-means που αναπτύχθηκε για κάθε αριθμό k, χωρίς φυσικά τα ενδιάμεσα στάδια ενημέρωσης για το χρήστη, είναι δηλαδή μια απλοποιημένη μορφή του βασικού αλγορίθμου.

Για την μετασχηματισμό των δεδομένων σε κλίμακα (scaling) έχει γίνει επιλογή του MinMaxScaler που περιέχεται στη βιβλιοθήκη sci-kit learn (scikit-learn developers, 2020). Ο scaler αυτός θέτει τα δεδομένα τυπικά σε κλίμακα 0 έως 1 οπότε απλοποιεί αρκετά τους υπολογισμούς και την οπτικοποίηση. Όπως αναφέραμε και παραπάνω είναι αρκετά σημαντικό για τη χρήση του k-means, διότι στην περίπτωση δεδομένων σε μη όμοια κλίμακα, μπορεί να είχαμε μεγάλη ευαισθησία στις διαφορές που παρουσιάζουν οι κλίμακές τους, μιας και ο αλγόριθμος βασίζεται σε μέτρηση αποστάσεων μεταξύ των σημείων. Ωστόσο, αυτή η μέθοδος είναι ευαίσθητη σε ακραίες τιμές (outliers). Ο μετασχηματισμός γίνεται με βάση την παρακάτω διατύπωση:

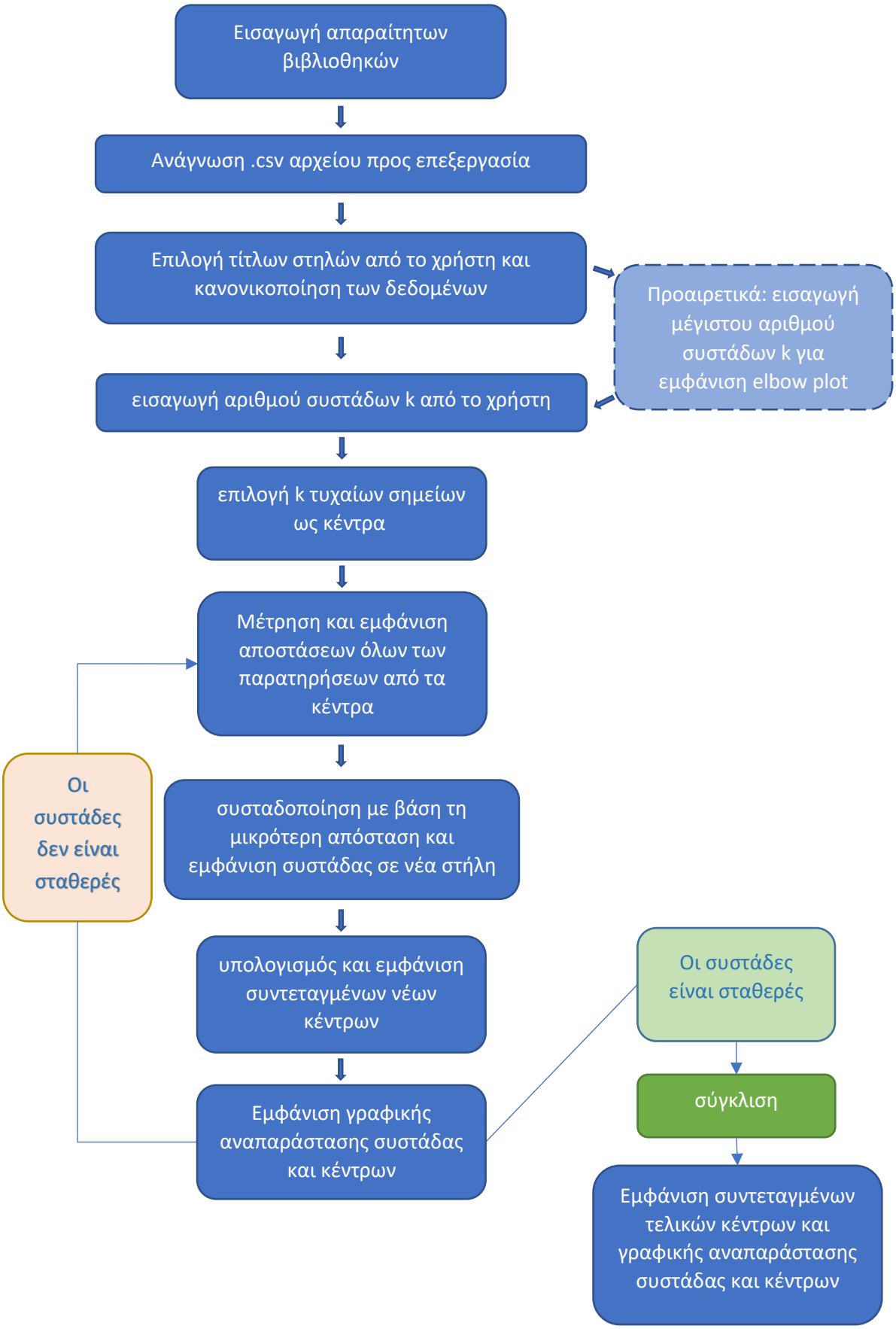
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

*X_scaled = X_std * (max - min) + min*

όπου min, max το εύρος του δείγματός μας.

Ακόμη μια ευρέως χρησιμοποιούμενη μέθοδος μετασχηματισμού δεδομένων η οποία περιλαμβάνεται στη βιβλιοθήκη scikit learn είναι η StandardScaler, η οποία αφαιρεί από κάθε παρατήρηση x το μέσο u, και διαιρεί με την τυπική απόκλιση του δείγματος s.

$$z = (x - u) / s$$



Διάγραμμα 3 – Απλοποιημένη αναπαράσταση του αλγορίθμου που αναπτύχθηκε

5.1 Οδηγίες χρήσης του αλγορίθμου που αναπτύχθηκε

Ο αλγόριθμος αναπτύχθηκε σε εικονικό περιβάλλον Jupyter notebook και το αρχείο .ipynb (interactive python notebook) στο οποίο περιέχεται, απαιτεί το περιβάλλον αυτό για να τρέξει. Στο κεφάλαιο 2 αναφέρονται οι ακριβείς εκδόσεις του απαραίτητου λογισμικού και των βιβλιοθηκών που χρησιμοποιήθηκαν. Το αρχείο .ipynb που περιέχει τον κώδικα, μπορεί να βρεθεί στο αποθετήριο github, το οποίο εξυπηρετεί το διαμοιρασμό του, στην παρακάτω διεύθυνση:

github.com/karajohn85/k-means_edu

Για να ξεκινήσει ο αλγόριθμος, απαιτεί την εισαγωγή των προς επεξεργασία δεδομένων από τον χρήστη σε αρχείο .csv. Όπως θα δούμε και παρακάτω, κατά την εισαγωγή αλλά και καθ' όλη τη διάρκεια των υπολογισμών και των αποτελεσμάτων του αλγορίθμου, τα δεδομένα έχουν μορφή pandas dataframe.

	A	B	C	D
1	,"Murder"	"Assault","UrbanPop","Rape"		
2	Alabama,13.2,236,58,21.2			
3	Alaska,10,263,48,44.5			
4	Arizona,8.1,294,80,31			
5	Arkansas,8.8,190,50,19.5			
6	California,9,276,91,40.6			
7	Colorado,7.9,204,78,38.7			

Εικόνα 1 - Απόσπασμα από την εμφάνιση των δεδομένων σε πρόγραμμα MS Excel

Το πρώτο βήμα στην εκτέλεση του αλγορίθμου είναι η εισαγωγή των απαραίτητων βιβλιοθηκών. Σε αυτό το βήμα δεν χρειάζεται κάποια παρέμβαση του χρήστη διότι γίνεται αυτόματα. Το επόμενο βήμα αποτελεί την εισαγωγή και ανάγνωση του προς επεξεργασία .csv αρχείου. Το αρχείο .csv θα πρέπει να βρίσκεται στον ίδιο φάκελο με τα αρχεία του αλγορίθμου που τρέχουμε. Το αρχείο αυτό προς διευκόλυνση μας θα πρέπει να περιέχει επικεφαλίδες (headers) στην πρώτη γραμμή των δεδομένων. Προς διευκόλυνση των παρακάτω στιγμιότυπων και απεικονίσεων χρησιμοποιούμε το δημοφιλές σύνολο δεδομένων (dataset) USArrests (McNeil, 1977). Σε αυτό λοιπόν το σημείο ο χρήστης καλείται να εισάγει το όνομα του αρχείου (χωρίς την κατάληξη):

```
Please enter the name of the .csv file (must be in the same path):
```

Εικόνα 2 – Πεδίο συμπλήρωσης ονόματος αρχείου .csv προς ανάλυση από τον αλγόριθμο.

Τα δεδομένα του αρχείου μετατρέπονται σε pandas dataframe και το αποτέλεσμα που εμφανίζεται είναι το σχήμα (shape) του καθώς και οι 5 πρώτες σειρές δεδομένων με τις επικεφαλίδες (εντολή head):

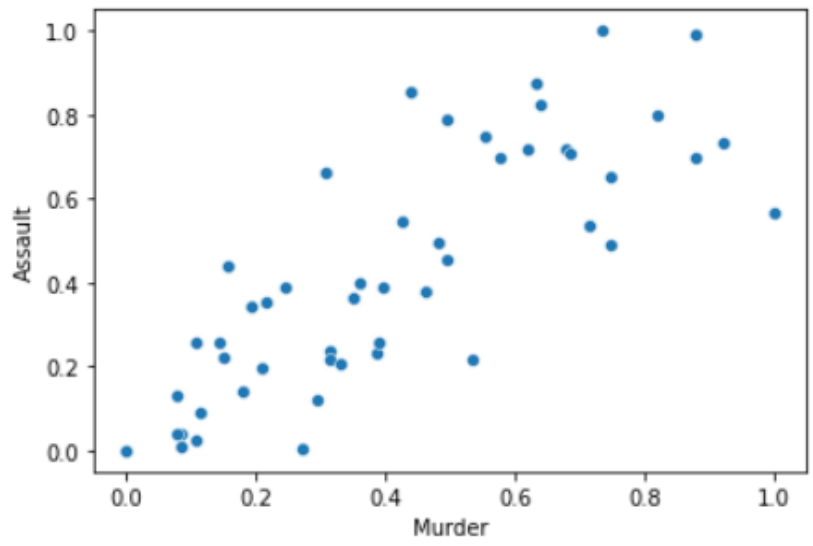
```
Please enter the name of the .csv file (must be in the same path):
 USArrests
Dataframe shape: rows, columns:
(50, 5)
Showing the first 5 rows of the dataframe:
```

	Unnamed: 0	Murder	Assault	UrbanPop	Rape
0	Alabama	13.2	236	58	21.2
1	Alaska	10.0	263	48	44.5
2	Arizona	8.1	294	80	31.0
3	Arkansas	8.8	190	50	19.5
4	California	9.0	276	91	40.6

Εικόνα 3 – Αποτελέσματα αφού ο χρήστης εισάγει το όνομα αρχείο και ο αλγόριθμος διαβάσει τα δεδομένα.

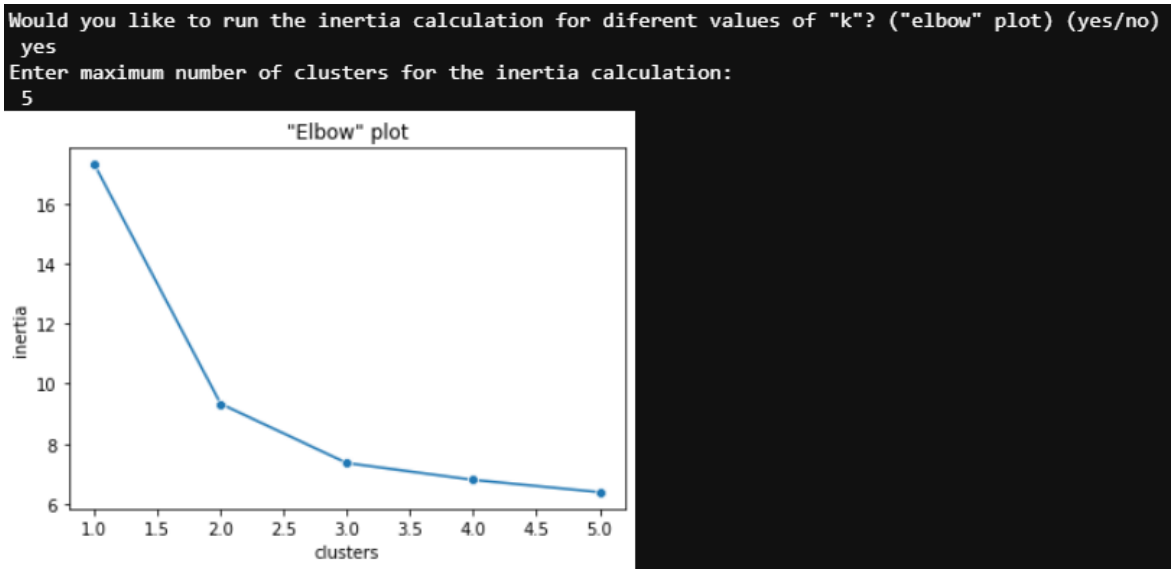
Στο επόμενο βήμα ο χρήστης εισάγει τα ονόματα των επικεφαλίδων για το x και y άξονα που θα επιλέξει να χρησιμοποιηθούν στην ανάλυση. Ταυτόχρονα με την εισαγωγή αυτή, γίνεται ανάθεση των δεδομένων των στηλών αυτών σε ένα νέο dataframe. Το dataframe αυτό υπόκειται αυτόματα σε αυτό το βήμα σε κανονικοποίηση με τον MinMaxScaler της βιβλιοθήκης sklearn και τα δεδομένα μας εμφανίζονται σε ένα διάγραμμα διασποράς (scatterplot).

```
Enter x axis label: Murder
Enter y axis label: Assault
x axis label is: Murder
y axis label is: Assault
<AxesSubplot:xlabel='Murder', ylabel='Assault'>
```



Εικόνα 4 – Αποτελέσματα που εμφανίζονται αφού ο χρήστης εισάγει το όνομα των στηλών που θα αναλύσει.

Στη συνέχεια τρέχουν τα βασικά τμήματα του αλγορίθμου για τον υπολογισμό της αδράνειας (inertia) και του k-means, αλλά ως συναρτήσεις (functions) χωρίς αποτέλεσμα, με σκοπό να ερωτηθεί στη συνέχεια ο χρήστης αν θέλει να εκτελέσει τον υπολογισμό της αδράνειας πριν το βασικό κομμάτι του κώδικα. Σε περίπτωση που η απάντηση στην ερώτηση είναι θετική, ο χρήστης ερωτάται για το μέγιστο αριθμό k συστάδων που θέλει να εκτελέσει τον υπολογισμό της αδράνειας. Το αποτέλεσμα είναι η εμφάνιση του elbow plot για των αριθμό συστάδων που επέλεξε ο χρήστης. Σε περίπτωση που ο χρήστης απαντήσει αρνητικά στον υπολογισμό και εμφάνιση του elbow plot, ο αλγόριθμος προχωρά στο επόμενο και κυρίως μέρος του υπολογισμού του k-means.



Εικόνα 5 – Διάγραμμα «elbow» που εμφανίζεται με βάση το μέγιστο αριθμό υπολογισμού συστάδων k που επιλέγει ο χρήστης.

Επιπλέον ο χρήστης λαμβάνει ως αποτέλεσμα τις ακριβείς τιμές της αδράνειας για κάθε βήμα σε μορφή λίστας:

```
This is the inertia value list for each number of clusters k:

[17.312633442609815, 9.329732291408202, 7.352893195093407, 6.794718091068642, 6.382477781492242]
```

Εικόνα 6 – Οι τιμές αδράνειας σε μορφή λίστας για κάθε τιμή αριθμού συστάδων που επέλεξε ο χρήστης.

Έπειτα ο χρήστης ερωτάται τον αριθμό k για τον οποίο επιθυμεί να τρέξει τον αλγόριθμο:

```
Enter number of clusters k to run the k-means algorithm:
 3
```

Εικόνα 7 – Πεδίο όπου ο χρήστης εισάγει τον αριθμό συστάδων k για τον οποίο επιθυμεί να εκτελέσει τον αλγόριθμο.

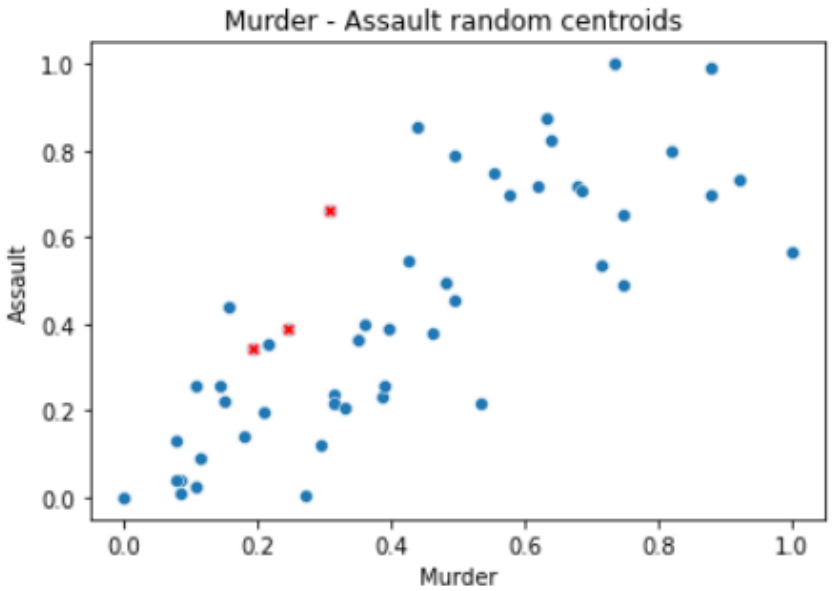
Με τη βοήθεια του παραπάνω elbow plot, επιλέγουμε τον αριθμό 3 που φαίνεται ότι θα εξηγεί αρκετά καλά τα συγκεκριμένα δεδομένα και πλέον ο αλγόριθμος τρέχει το βασικό κομμάτι του. Αρχικά επιλέγονται k τυχαία σημεία (εδώ 3 σημεία) από τις παρατηρήσεις τα οποία θα αποτελούν τα αρχικά κέντρα.

Initial random centroids coordinates:

	Murder	Assault
36	0.246988	0.390411
46	0.192771	0.342466
7	0.307229	0.660959

Εικόνα 8 – Η επιλογή των k αρχικών κέντρων από τον αλγόριθμο και η εμφάνιση των συντεταγμένων τους.

Τα τυχαία αυτά σημεία αυτά εμφανίζονται με το σύνολο των δεδομένων σε χρώμα κόκκινο:



Διάγραμμα 4 – Εμφάνιση αρχικών τυχαίων κέντρων από υφιστάμενα σημεία του dataset και αναπαράστασή τους με κόκκινο X στο διάγραμμα διασποράς.

Στη συνέχεια ο αλγόριθμος ενημερώνει το χρήστη για τους υπολογισμούς της πρώτης επανάληψης:

Iteration number 1 calculation results:

	Murder	Assault	dist_centr_0	dist_centr_1	dist_centr_2	min_dist
0	0.746988	0.654110	0.565276	0.635829	0.439812	dist_centr_2
1	0.554217	0.746575	0.470364	0.542169	0.261406	dist_centr_2
2	0.439759	0.852740	0.500908	0.566906	0.233118	dist_centr_2
3	0.481928	0.496575	0.257813	0.327660	0.239878	dist_centr_2
4	0.493976	0.791096	0.470693	0.540364	0.227618	dist_centr_2
5	0.427711	0.544521	0.237509	0.309876	0.167552	dist_centr_2
6	0.150602	0.222603	0.193519	0.127064	0.465498	dist_centr_1
7	0.307229	0.660959	0.277174	0.338435	0.000000	dist_centr_2
8	0.879518	0.993151	0.873722	0.946051	0.661715	dist_centr_2

Εικόνα 9 – Απόσπασμα από τα αποτελέσματα των υπολογισμών του αλγορίθμου για τις αποστάσεις των σημείων από τα κέντρα και την ανάθεσή της κάθε παρατήρησης στο κοντινότερο κέντρο.

Στις δύο πρώτες στήλες εμφανίζονται οι παρατηρήσεις μας και στις τρεις επόμενες είναι η Ευκλείδεια απόσταση της κάθε παρατήρησης από το εκάστοτε κέντρο που αντιπροσωπεύεται με τις στήλες (π.χ. dist_centr_0 -> είναι η απόσταση των παρατηρήσεων από το κέντρο 0). Στη στήλη «min_dist» εμφανίζεται το όνομα του κέντρου για το οποίο η κάθε παρατήρηση εμφανίζει την ελάχιστη απόσταση, άρα και η ανάθεση στην αντίστοιχη συστάδα (π.χ. dist_centr_2 -> είναι η ανάθεση παρατήρησης στη συστάδα 2).

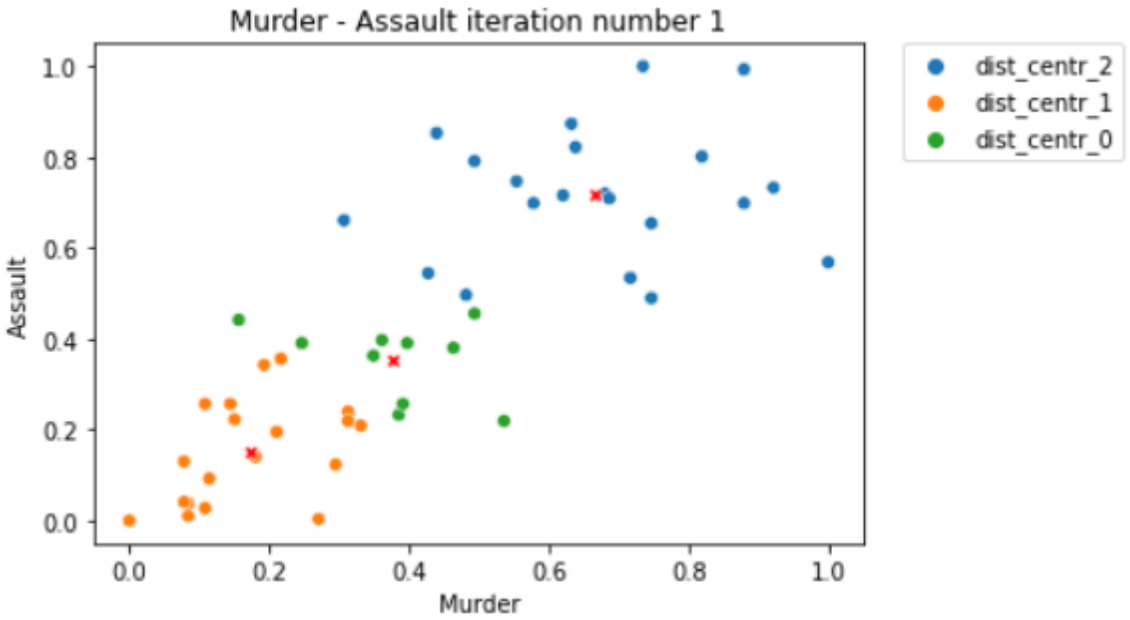
Μετά από τους υπολογισμούς, ο χρήστης λαμβάνει πληροφόρηση για τις συντεταγμένες των νέων κέντρων:

New centroids coordinates after iteration 1:

min_dist	Murder	Assault
dist_centr_0	0.378313	0.352740
dist_centr_1	0.172479	0.152848
dist_centr_2	0.666093	0.719178

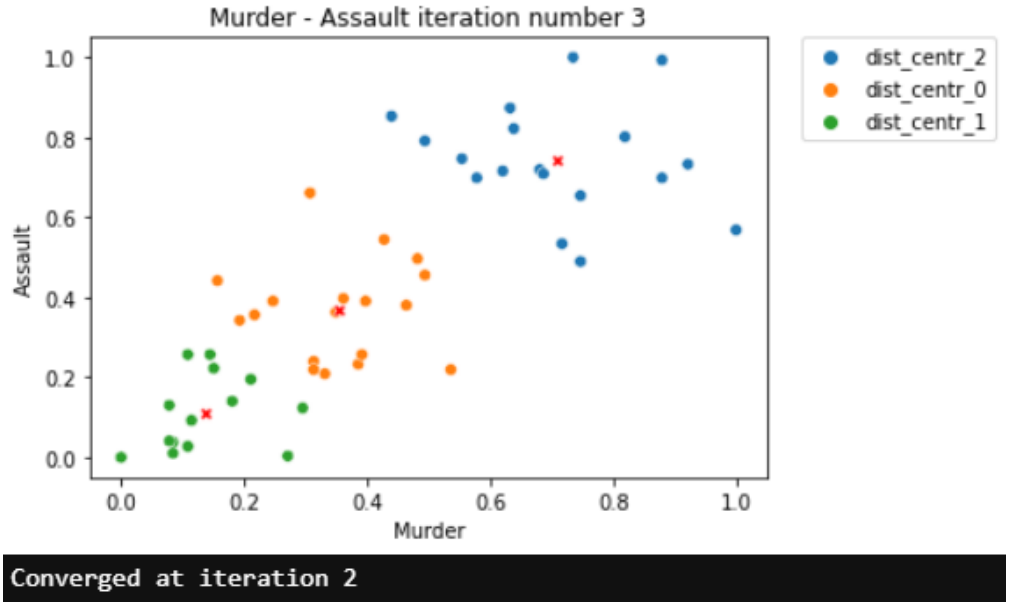
Εικόνα 10 – Εμφάνιση των νέων κέντρων που προέκυψαν από τους υπολογισμούς της πρώτης επανάληψης.

Εδώ βλέπουμε ότι από την πρώτη επανάληψη έχουν προκύψει νέα κέντρα για τις τρεις συστάδες. Τα νέα αυτά κέντρα εμφανίζονται σε διάγραμμα διασποράς μαζί με τις παρατηρήσεις όπου πλέον φαίνεται με διαφορετικό χρώμα και η ανάθεσή τους στις συστάδες:



Διάγραμμα 5 – Εμφάνιση κέντρων που προέκυψαν από την πρώτη επανάληψη με κόκκινο Χ, καθώς και χρωματική ανάθεση των σημείων σε συστάδες.

Η παραπάνω διαδικασία επαναλαμβάνεται για όσες φορές χρειαστεί μέχρι να υπάρξει σύγκλιση, δηλαδή τα σημεία να μην εμφανίζουν αλλαγή συστάδας σε κάθε επανάληψη και συγκεκριμένα οι συντεταγμένες των κέντρων να μην αλλάξουν για δύο διαδοχικές επαναλήψεις. Σε αυτή την περίπτωση ο χρήστης λαμβάνει το τελικό διάγραμμα διασποράς καθώς και την ενημέρωση για τη σύγκλιση και σε ποια επανάληψη αυτή εμφανίστηκε:



Διάγραμμα 6 – Τελικό διάγραμμα διασποράς που προέκυψε από την τρίτη επανάληψη του αλγορίθμου, όπου εμφανίστηκε η σύγκλιση.

Στο παραπάνω διάγραμμα διασποράς παρατηρούμε ότι ο τίτλος αναφέρει τρίτη επανάληψη (iteration number 3), ενώ το αποτέλεσμα αναφέρει σύγκλιση στη δεύτερη επανάληψη (converged at iteration 2). Αυτό συμβαίνει διότι ο k-means αλγόριθμος χρειάζεται να κάνει επαναλήψεις μέχρι να καταλήξει σε αποτέλεσμα όπου τα κέντρα των συστάδων δεν μετακινούνται πλέον και όλα τα σημεία παραμένουν στις ίδιες ακριβώς συστάδες. Στη συγκεκριμένη περίπτωση έφτασε μέχρι και την τρίτη επανάληψη και αφού σύγκρινε τα αποτελέσματά της με αυτά της δεύτερης επανάληψης, κατέληξε σε αποτέλεσμα. Δηλαδή τα κέντρα που προέκυψαν και οι αναθέσεις παρατηρήσεων σε αυτά της δεύτερης επανάληψης, είναι τα τελικά και δεν αλλάζουν πλέον.

Τα διαγράμματα διασποράς που θα προκύψουν αποθηκεύονται στο φάκελο εκτέλεσης της python με ονομασία αρχείου «*αριθμός επανάληψης*.jpeg», οπότε για παράδειγμα με την τρίτη επανάληψη θα προκύψει το αρχείο «3.jpeg». Το αρχείο «0.jpeg» θα είναι πάντα αυτό με τις αρχικές τυχαίες αναθέσεις σημείων σε κέντρα.

5.2 Περιορισμοί του αλγορίθμου που αναπτύχθηκε

Δεδομένου του στόχου για τον οποίο αναπτύχθηκε ο αλγόριθμος, υπάρχουν κάποιοι λειτουργικοί περιορισμοί ώστε να εξασφαλιστεί η απλότητά και η ευχρηστία του. Άλλωστε παρόμοιοι αλγόριθμοι που διατίθενται ήδη σε βιβλιοθήκες είναι αρκετά προχωρημένοι και έχουν δοκιμαστεί σε πλήθος δεδομένων για μεγάλο χρονικό διάστημα, πράγμα που δεν είναι εφικτό να γίνει στα πλαίσια ενός ακαδημαϊκού εξαμήνου. Οι βασικοί περιορισμοί εντοπίζονται παρακάτω:

- Δέχεται μόνο δεδομένα δύο διαστάσεων προς διευκόλυνση υπολογισμών και οπτικοποίησης των αποτελεσμάτων.

- Αν και ο αλγόριθμος έχει δοκιμαστεί σε σύνολο δεδομένων αρκετών παρατηρήσεων χωρίς προβλήματα ή σφάλματα (περίπου 100.000 παρατηρήσεις από το MovieLens Dataset ml-latest-small (Konstan., 2015)), δεδομένου ότι προορίζεται κυρίως για εκπαιδευτικούς σκοπούς ή επίδειξη, συνίσταται η χρήση συνόλου δεδομένων έως 60 παρατηρήσεων και αριθμού k έως 10 επαναλήψεις. Έτσι θα μπορεί να εμφανίζεται το σύνολο των πράξεων ενός dataframe με τις τυπικές ρυθμίσεις της python.

- Σε συνδυασμό με το παραπάνω, και δεδομένου ότι ο αλγόριθμος δεν έχει βελτιστοποιηθεί ως προς τον χρόνο εκτέλεσής του, όταν έχουμε μεγάλο όγκο δεδομένων προς επεξεργασία, το αποτέλεσμα είναι – ανάλογα και με τους διαθέσιμους υπολογιστικούς πόρους – η ύπαρξη μεγάλων χρόνων εκτέλεσης.

5.3 Αντίστοιχοι αλγόριθμοι k-means που έχουν αναπτυχθεί σε python

Όπως αναφέρθηκε και προηγουμένως, ο k-means αποτελεί έναν από τους πιο διαδεδομένους αλγορίθμους της κατηγορίας του. Είναι λοιπόν αναμενόμενο να υπάρχουν αντίστοιχες προσπάθειες ανάπτυξης του αλγορίθμου από την αρχή, σε περιβάλλον python. Κάποιες από τις υλοποιήσεις που μελετήθηκαν είναι οι παρακάτω:

- «The Most Comprehensive Guide to K-Means Clustering You'll Ever Need» (Sharma, 2019)

Η παραπάνω υλοποίηση αποτέλεσε μια βασική πηγή έμπνευσης για τον αλγόριθμο που αναπτύχθηκε στα πλαίσια της εργασίας αυτής, ωστόσο ο τρόπος διαχείρισης των δεδομένων διαφέρει αρκετά. Επιπλέον, το αποτέλεσμα του αλγορίθμου μόλις ολοκληρωθεί η εκτέλεσή του είναι μόνο οι τελικές συστάδες με κάποια βασικά χαρακτηριστικά και γραφική αναπαράσταση. Στο συγκεκριμένο παράδειγμα χρησιμοποιείται ένα σύνολο δεδομένων πρόβλεψης χορήγησης δανείου.

- «A Complete K Mean Clustering Algorithm from Scratch in Python: Step by Step Guide» (Sucky, 2020)

Στην παραπάνω ιστοσελίδα γίνεται επίσης μια εξαρχής ανάπτυξη του αλγορίθμου, όπου στο δεύτερο τμήμα της ο αλγόριθμος χρησιμοποιείται για μείωση διαστάσεων μιας εικόνας (Dimensionality Reduction). Η οργάνωση και διαχείριση των δεδομένων γίνεται σε NumPy arrays. Σε αυτή την υλοποίηση εμφανίζεται ένα διάγραμμα με τις συστάδες σε κάθε επανάληψη, οπότε υπάρχει μια επιπλέον πληροφόρηση για τον χρήστη.

- «K-means and hierarchical clustering with Python» (Grus, 2016)

Στο βιβλίο αυτό εκτός από την ανάπτυξη του αλγορίθμου από την αρχή και με πολύ αναλυτικό τρόπο, γίνεται επίσης αναφορά σε παραδείγματα πραγματικής εφαρμογής του.

6 Συμπεράσματα και προτάσεις

Τα τελευταία χρόνια παρατηρείται μια ταχεία διάδοση των αλγορίθμων Μηχανικής Μάθησης, κυρίως οδηγούμενη από τον διαθέσιμο όγκο δεδομένων που στις μέρες μας αυξάνεται με υψηλούς ρυθμούς. Η μετατροπή των δεδομένων αυτών σε πληροφορία και έπειτα γνώση, είναι ιδιαίτερα σημαντική ώστε τα δεδομένα τελικά να αξιοποιηθούν αποδοτικά. Η Μηχανική Μάθηση και γενικότερα η Επιστήμη Δεδομένων μας δίνουν τα απαραίτητα εργαλεία ώστε αυτό να είναι εφικτό.

Στη συγκεκριμένη εργασία ασχοληθήκαμε με ένα πολύ μικρό τμήμα της Μηχανικής Μάθησης που συνήθως δεν επαρκεί από μόνο του να ερμηνεύσει επαρκώς δεδομένα από τον πραγματικό κόσμο. Στις περισσότερες περιπτώσεις απαιτείται συνδυασμός περισσότερων του ενός εργαλείων, ώστε ο αναλυτής να καταλήξει σε αξιόπιστα συμπεράσματα. Ωστόσο, ο σκοπός της εργασίας είναι να αναπτυχθεί εξολοκλήρου και από την αρχή ο αλγόριθμος με τη χρήση της python, μιας γλώσσας προγραμματισμού που από μόνη της μπορεί να αποτελέσει ένα σημαντικό εργαλείο αναλύσεων. Δεδομένου ότι ο χρήστης ή αναλυτής θα πρέπει να έχει μια σαφή εικόνα για τον τρόπο λειτουργίας των αλγορίθμων που έχει τη δυνατότητα να επιλέξει, η παρούσα εργασία καθιστά λίγο πιο εύκολη την επιλογή αυτή. Ο αλγόριθμος που αναπτύχθηκε στην python απευθύνεται σε αρχάριους τόσο στο χώρο της Μηχανικής Μάθησης και Αναλυτικής Δεδομένων, όσο και στον προγραμματισμό σε python. Υπό αυτή τη σκοπιά η εργασία θα μπορούσε να αποτελέσει εργαλείο εκπαίδευσης και επίδειξης για χρήστες που εισέρχονται στα παραπάνω πεδία, αλλά και πηγή έμπνευσης ώστε ο αλγόριθμος να εμπλουτιστεί ή να βελτιωθεί ώστε να αποτελέσει ένα ολοκληρωμένο και αποδοτικό εργαλείο.

Βιβλιογραφία

ALPHABET INC. (2021, 1). *AlphaGo*. Ανάκτηση από deepmind.com: https://deepmind.com/research/case-studies/alphago-the-story-so-far

Ameet V Joshi. (2020). *Machine Learning and Artificial Intelligence.* Springer.

Arumawadu, H. R. (2015). Mining Profitability of Telecommunication Customers Using K-Means Clustering. *Journal of Data Analysis and Information Processing*, 63-71.

Bacila, M. &. (2012). Prepaid Telecom Customer Segmentation Using the K-Mean Algorithm. *Analele Universitatii din Oradea*, 1112-1118.

David Arthur, S. V. (2007, January). k-means++: The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007.

Demetriou, I. C. (2020, March 10). Machine Learning K Means Clustering.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*.

github. (χ.χ.). Ανάκτηση 2020, από https://github.com/python/cpython/blob/3.9/Lib/warnings.py

Grus, J. (2016). *K-means and hierarchical clustering with Python.* O'Reilly Media, Inc.

H.P. Ng, S. O. (2006). *MEDICAL IMAGE SEGMENTATION USING K-MEANS CLUSTERING AND IMPROVED WATERSHED ALGORITHM.* Singapore: IEEE.

J.Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 53-65.

Konstan., F. M. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 4: 19:1–19:19.

Machine Learning Crash Course. (2021, January). Ανάκτηση από Google Developers: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* University of California.

matplotlib. (2020). Ανάκτηση από https://matplotlib.org/

McNeil, D. R. (1977). *Interactive Data Analysis.* New York: Wiley.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill Education.

Nameirakpam Dhanachandra, K. M. (2015). Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm. *Eleventh International Multi-Conference on Information Processing.* Elsevier B.V.

NumPy. (2020). Ανάκτηση από https://numpy.org/

Pandas. (2020). Ανάκτηση από https://pandas.pydata.org/

Poggio, K.-K. S. (1994). *Example-based Learning for View-based Human Face Detection.*

Robert Tibshirani, G. W. (2000). *Estimating the number of clusters in a data set via the gap statistic.* Stanford University.

S. Lloyd. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 129 - 137.

scikit-learn. (2020). Ανάκτηση από https://scikit-learn.org/

scikit-learn developers. (2020). *preprocessing scaler*. Ανάκτηση από scikit-learn: https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler

SciPy.org. (2020). Ανάκτηση December 2020, από https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.vq.kmeans.html#scipy.cluster.vq.kmeans

seaborn. (2020). Ανάκτηση από https://seaborn.pydata.org/

Sharma, P. (2019, August 19). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. Ανάκτηση από Analytics Vidhya: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Siddheswar Ray, R. T. (χ.χ.). *Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation.* Victoria, 3168, Australia: School of Computer Science and Software Engineering Monash University.

Sucky, R. N. (2020, October 17). *A Complete K Mean Clustering Algorithm From Scratch in Python: Step by Step Guide*. Ανάκτηση από towardsdatascience.com: https://towardsdatascience.com/a-complete-k-mean-clustering-algorithm-from-scratch-in-python-step-by-step-guide-1eb05cdcd461

Thorndike, R. (1953). Who belongs in the family? *Psychometrika 18*, 267-276.

Yudhanegara Mokhammad, I. S. (2020). CLUSTERING FOR ITEM DELIVERY USING RULE-K-MEANS. *Journal of the Indonesian Mathematical Society*, 185-191.