Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών

Τίτλος

Εκπόνηση Εργασίας:

Θεοδωρίδης Ξενοφών

Επιβλέπων Καθηγητής:

Νικόλαος Πιτσιάνης

Θεσσαλονίκη, Ημερομηνία

Ευχαριστίες

Ονοματεπώνυμο

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η δημιουργία μιας βιβλιοθήκης αλγορίθμων ομαδοποίησης της τεχνιτής μηχανικής μάθησης στο περιβάλλον Matlab.

Στο πρώτο κομμάτι γίνεται μία εκτενής ανάλυση στην μηχανική μάθηση και στους κυριότερους αλγορίθμους ομαδοποίησης όπου πραγματοποείται μία ιστορική αναδρομή και παρουσίαση του κύριου σώματος του κάθε αλγορίθμου καθώς και των βασικών συστατικών που τους συντελούν.

Στο δεύτερο μέρος παρουσιάζονται οι υλοποιήσεις των αλγορίθμων στο περιβάλλον Matlab μαζί και ολοκληρώνεται με την παρουσίαση των χρόνων επίδοσης και ευστοχίας τους κάθε αλγόριθμου σε διάφορα σύνολα δεδομένων.

Τα τελευταία χρόνια η εξέλιξη τόσο στους αλγορίθμους όσο και στο υλικό των Η/Υ έχει επιτρέψει στην μηχανική μάθηση να γνωρίσει επιτυχίες που κάποτε φάνταζαν ακατόρθωτες. Οι τρεις κύριοι κλάδοι της μηχανικής μάθησης είναι:

1) Ενισχυτική Μάθηση
2) Μάθηση με Επίβλεψη
3) Μάθηση χωρίς Επίβλεψη

Στη μάθηση με επίβλεψη το μοντέλο καλείται να μάθει μέσα από ένα σετ δεδομένων μαζί με ετικέτες για κάθε εγγραφή μία έννοια ή συνάρτηση, η οποία αντιστοιχίζει τις εγγραφές στις αντίστοιχες ετικέτες.

Στη μάθηση χωρίς επίβλεψη το σύστημα ανακαλύπτει της συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων χωρίς την ύπαρξη ετικετών, δημιουργώντας πρότυπα χωρίς να είνα γνωστό αν υπάρχουν.

Η ενισχυτική μάθηση ασχολείται με το πως νοήμονες πράκτορες(μοντέλα) αλληλεπιδρούν με το περιβάλλον με πράξεις για τις οποίες ανταμίβονται προκειμένου να μεγιστοποιήσουν το συνολικό έπαθλο.

Η μάθηση με επίβλεψη , κυρίως χάρη στην ανάπτυξη των νευρωνικών δικτύων, έχει καταφέρει να λύσει προβλήματα με ποσοστά επιτυχίας που κάποτε φάνταζαν ακατόρθωτα. Μερικά από τα πιο πρόσφατα είναι: Ο διαγωνισμός Imagenet για αντιστοίχιση ετικέτων σε φωτογραφίες, Αναγνώριση E-mail Spam, Αναγνώριση συναλλαγών απάτης, Δίπλωμα Πρωτεΐνων.

Από την άλλη μεριά η μάθηση χωρίς επίβλεψη δεν έχει γνωρίσει την ίδια επιτυχία σε σχέση με τις δύο προηγούμενες ουτέ λαμβάνει το απαραίτητο ενδιαφέρον από την επιστημονική κοινώτητα. Ένας από τους λόγους είναι ότι το πρόβλημα που καλεί να λύσει, δηλαδή η εξαγωγή προτύπων χωρίς την ύπαρξη ετικέτων και περιβάλλοντος, είναι τάξης πιο δύσκολο από αυτό που λύνουν η μάθηση με επίβλεψη και η ενισχυτική αντίστοιχα.Όμως η επιτυχή ανάπτηξη της μάθησης χωρίς επίβλεψη είναι εξίσου αν όχι μεγαλύτερης σημασίας σε σχέση με αυτή των άλλων δύο.Κάνοντας παράθεση τον Yann LeCun έναν από τους πατέρες της μηχανικής μάθησης:

Η Μάθηση χωρίς επίβλεψη είναι το ιερό δισκοπότηρο
προκειμένου να ξεκλειδώσουμε τη γενική μηχανική μάθηση

’Ενας από τους επιμέρους κλάδους της Μάθησης χωρίς Επίβλεψης είναι η Ομαδοποίσης. Σκοπός της παρούσας διπλωματικής είναι η ανάπτυξη αλγορίθμων ομαδοποίησης στο περιβάλλον Matlab, τόσο με γνώμονα τη χρηστικότητα όσο και την επίδοση.

Κεφάλαιο 1

Εισαγωγή

Κάθε μέρα παράγονται μεγάλες ποσότητες απο δεδομένα και πληροφορίες. Υπολογίζεται ότι περίπου 2.5 ZB δημιουργούνται καθημερινά και με την συνεχώς αυξανόμενη χρήση του IoT και Η/Υ ο συγκεκριμένος αριθμός ολοένα και αυξάνεται. Εύλογο επακόλουθο είναι ότι και ο χρόνος και το κόστος επεξεργασίας του όγκου δεδομένων αυξάνονται επίσης. Στο σημείο αυτό επεμβαίνει η μηχανική μάθηση επιτρέποντας την επεξεργασία τους σε λογικά χρονικά περιθώρια. Η μηχανική μάθηση είναι ένα παρακλάδι της τεχνιτής νοημοσύνης. Σκοπός της είναι η επίλυση προβλημάτων χωρίς συγγραφή κώδικα που τα επιλύει, αλλά μέσω από προηγούμενες εμπειρίες και παραδείγματα να σκεφτεί ο υπολογιστής σαν ”άνθρωπος”. Πιο τεχνικά, δεδομένου ένα σετ δεδομένων καλείται ο εκάστοτε αλγόριθμος μηχανικής μάθησης να χτύσει ένα μαθηματικό μοντέλο ώστε να προσαρμόζει τα δεδομένα ώστε να είναι κατανοητά και να μπορούν να χρησιμοποιηθούν από ανθρώπους.

Η έννοια της μηχανικής μάθησης πρωτο-υιοθετήθηκε το 1956 στο κολλέγιο του Ντάρθμουθ από τους Newell, Simon, McCarthy και Samuel. Μαζί με την ομάδα τους μελέτησαν την δυνατότητα λύσης προβλημάτων από ηλεκτρονικούς υπολογιστές χωρίς ρητή διατύπωση κώδικα.

1.1

1.1.1 Subsection 1

1.1.2 Subsection 2

1.2 Main Section 2

$$K(x,y) = \sum_{k=0}^{p-1} \frac{1}{k!} \frac{\partial^k}{\partial x^k} K(c_s, y)(x - c_s)^k \tag{1.1}$$

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Αλγόριθμοι Ομαδοποίησης

Η ομαδοποίηση είναι μία στατιστική μέθοδος επεξεργασίας δεδομένων. Σκοπός της είναι η οργάνωση των δεδομένων σε ομάδες βάση μια μετρικής ομοιότητας μεταξύ. 'Οπως έχει ήδη αναφερθεί η ομαδοποίηση είναι μια μέθοθος μη επιτηρούμενης μάθησης και χρησιμοποιείται κυρίως σε περιπτώσεις που δεν υπάρχει κάποια αρχική γνώση για την δομή των δεδομένων. Οι αλγόριθμου ομαδοποίησης χωρίζονται στις παρακάτω κατηγορίες:

1. Ιεραρχική Ομαδοποίηση
2. Ομαδοποίηση Βάση Κέντρων
3. Ομαδοποίηση Βάση Κατανομής
4. Ομαδοποίηση Βάση πυκνότητας

Αναφέρεται ότι θα αναλυθούν μόνο οι αλγόριθμοι που χρησιμοποιούνται στην παρούσα διπλωματική εργασία.

2.1.1 Ιεραρχική Ομαδοποίηση

Αν δίνεται ένα σύνολο δεδομένων $D = x_1, ..., x_n$ όπου $x_i \in R^d$, μία συσταδοποίηση $C = C_1, ..., C_k$ αποτελεί διαμέριση του D, δηλαδή κάθε συστάδα είναι ένα σύνολο σημείων $C_i \subseteq D$ τέτοιο ώστε οι συστάδες να είναι ξένες ανα ζεύγη. Υπάρχουν δύο κύριες κατηγορίες αλγορίθμων ιεραρχικής ομαδοποίησης: οι συσσωρευτικές (agglomerative) και οι διαιρετικές (divisive). Οι συσσωρευτικές μέθοδοι λειτουργούν συνχονευτικά (bottom-up). Δηλαδή ξεκινούν με μία ξεχωριστή συστάδα για καθένα από τα *n* σημεία και συγχωνεύουν επανειλημμένα το ζεύγος των συστάδων με τη μεγαλύτερη ομοιότητα. Η διαδικασία σταματά όταν όλα τα σημεία ανήκουν στην ίδια συστάδα. Οι διαιρετικές μέθοδοι λειτουργούν αντίθετα (top-down). Ξεκινούν με μία συστάδα που περιέχει όλα τα σημεία και διαμερίζουν αναδρομικά τις συστάδες, με τη διαδικασία να σταματά όταν όλα τα σημεία ανήκουν σε διαφορετικές συστάδες. Σε επόμενη ενότητα θα αναλυθούν οι κυριότεροι από τους αλγορίθμους συσσωρετικής ιεραρχικής ομαδοποίησης.

2.1.2 Συσταδοποίση βάση αντιπροσώπων

Αν δίνεται ένα σύνολο δεδομένων με *n* σημεία σε έναν *d*-δίαστατο χώρο, $D = x_1, ...x_n$, καθώς και το πλήθος των επιθυμητών συστάδων *k*, ο στόχος της βασισμένης σε αντιπροσώπους συσταδοποίσης είναι ο διαμερισμός του συνόλου σε *k* ομάδες, η οποία ονομάζεται *συσταδοποίηση* και συμβολίζεται με $C = C_1, C_2, ..., C_k$. Επιπλέον, για κάθε συστάδα C_i υπάρχει ένα αντιπροσωπευτικό σημείο που τη συνοψίζει. Μία δημοφιλής επιλογή γι'αυτό το σημείο είναι ο μέσος $\boldsymbol{\mu}_i$ όλων των σημείων της συστάδας, που ονομάζεται επίσης *κέντρο βάρους* centroid:

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$$

όπου $n_i = |C_i|$ είναι το πλήθος των σημείων που ανήκουν στη συστάδα C_i.

'Ατυπα, κάθε σημείο μπορεί να αντιστοιχιστεί σε οποιαδήποτε από τις *k* συστάδες, οπότε υπάρχουν το πολύ k^n πιθανές συσταδοποιήσεις. Όμως οποιαδήποτε μετάθεση των *k* συστάδων εντός μιας δεδομένης συσταδοποίησης παράγει μια ισοδύναμη συσταδοποίηση και συνεπώς υπάρχουν $O(k^n/k!)$ συσταδοποιήσεις των *n* σημείων σε *k* ομάδες. Είναι προφανές ότι η εξαντλητική απαρίθμηση και βαθμολόγηση όλων των πιθανών συσταδοποιήσεων είνα πρακτικά ανέφικτη. Στο επόμενο κεφάλαιο θα αναλυθεί ο αλγόριθμων K μέσων (K-means algorithm), ένας άπληστος αλγόριθμος για την εύρεση συστάδων βάση αντιπροσώπων.

2.1.3 Συσταδοποίηση βασισμένη στην πυκνότητα

Οι μέθοδοι συσταδοποίησης που βασίζονται σε αντιπροσώπους, όπως ο αλγόριθμος K μέσων, είναι κατάλληλες για την εύρεση συστάδων με ελλειψοειδές σχήμα ή, στην καλύτερη περίπτωση, κυρτών συστάδων. Ωστόσο, στην περίπτωση κοίλων συστάδων οι συγκεκριμένες μέθοδοι δυσκολεύονται να βρουν τις πραγματικές συστάδες επειδή δύο σημεία από διαφορετικές συστάδες ενδέχεται να βρίσκονται πιο κοντά από ό,τι δύο σημεία που ανήκουν στην ίδια συστάδα. Οι μέθοδοι που βασίζονται στην πυκνότητα (density), είναι σε θέση να εξορύξουν τέτοιες κοίλες συστάδες.

Στη συσταδοποίηση που βασίζεται στην πυκνότητα, δεν χρησιμοποιείται μόνο η απόσταση των σημείων στον προσδιορισμό των συστάδων, αλλά αξιοποείται και η τοπική πυκνότητα των σημείων. Ορίζουμαι μια μπάλα ακτίνας *ε* γύρο από ένα σημείο $\mathbf{x} \in R^d$, η οποία ονομάζεται *ε*-γειτονιά (*ε*-neighborhood) του **x**, ως εξής:

$$N_\varepsilon(x) = B_d(\mathbf{x},\varepsilon) = \{\mathbf{y} | \boldsymbol{\delta}(\mathbf{x}, \mathbf{y}) \leq \varepsilon\}$$

Εδώ, το $\boldsymbol{\delta}(\mathbf{x}, \mathbf{y})$ αναπαριστά την απόσταση των σημείων **x** και **y**. Συνήθως υποθέτουμε ότι πρόκειται για την Ευκλείδων απόσταση των σημείων. Ωστόσο, μπορούν να χρησιμοποιηθούν και άλλες μετρικές της απόστασης.

2.1.4 Φασματική συσταδοποίηση και συσταδοποίηση γραφημάτων

Ας υποθέσουμε ότι δίνεται ένα σύνολο δεδομένων $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^n$ που αποτελείται από *n* σημεία στον $\mathbf{x} \in R^d$. 'Εστω ότι συμβολίζουμε με **A** τη συμμετρική *μήτρα ομοιότητας*

(similarity matrix), διαστάσεων $n \times n$, μεταξύ των σημείων, η οποία ορίζεται ως

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{2.1}$$

όπου το **A***(i,j)*=a_{ij} αναπαριστά την ομοιότητα ή συνάφεια (affinity) των σημείων $\mathbf{x}_i$ και $\mathbf{x}_j$. Απαιτούμε από την ομοιότητα να είναι συμμετρική και μη αρνητική, δηλαδή να ισχύουν οι σχέσεις $\alpha_{ij} = \alpha_{ji}$ και $\alpha_{ij} \geq 0$, αντίστοιχα. Η μήτρα **A** μπορεί να εκληφθεί ως *σταθμισμένη μήτρα γειτνίασης* (weighted adjacency matrix) του σταθμισμένου (μη κατευθυνόμενου) γραφήματος *G = (V,E)*,όπου κάθε κορυφή είναι ένα σημείο και κάθε ακμή ενώνει ένα ζεύγος σημείων. Επιπλέον η μήτρα ομοιότητα **A** παρέχει το βάρος για κάθε ακμή, δηλαδή το στοιχείο α_{ij} αναπαριστά το βάρος της ακμής $(\mathbf{x}_i, \mathbf{x}_j)$. Αν όλες οι τιμές συνάφειας είναι 0 ή 1, τότε η μήτρα **A** αναπαριστά την κανονική σχέση γειτνίασης μεταξύ των κορυφών.

Κεφάλαιο 3

Ανάλυση των Αλγορίθμων

3.1 K-Μέσοι

3.1.1

'Εστω ότι δίνεται μια συσταδοποίηση $C = C_1, C_2, ..., C_k$. Χρειαζόμαστε κάποια συνάρτηση βαθμολόγησης η οποία θα αξιολογεί την ποιότητα της συσταδοποίησης. Η συνάρτηση αυτή βασίζεται στο *άθροισμα των τετραγώνων των σφαλμάτων* (sum of squared errors,SSE) και ορίζεται ως

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2 \tag{3.1}$$

Στόχος είναι να η εύρεση της συσταδοποίησης που ελαχιστοποιεί τη βαθμολογία SSE:

$$C^* = \arg\min_{C} \{SSE(C)\}$$

Ο αλγόριθμος K μέσων χρησιμοποιεί μια άπληστη(greedy) επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την (3.1) και κατά συνέπεια, μπορεί να συγκλίνει σε τοπικά βέλτιστα και όχι σε ολικά. Ο συγκεκριμένος αλγόριθμος καθορίζει τις αρχικές τιμές των μέσων για τις συστάδες παράγοντας με τυχαίο τρόπο *k* σημεία στο χώρο δεδομένων. Κάθε επανάληψη του αλγορίθμου K μέσων αποτελείται από δύο βήματα:(1) την αντιστοίχιση των σημείων σε συστάδες και (2) την ενημέρωση των κέντρων βάρους. Κάθε σημείο $\mathbf{x}_j \in \mathbf{D}$, στο πρώτο βήμα του αλγορίθμου, αντιστοιχίζεται στον πλησιέστερο μέσο. Αυτό προκαλεί μία συσταδοποίηση, με κάθε συστάδα C_i να περιλαβάνει τα σημεία που βρίσκονται πιο κοντά στο κέντρο βάρους $\boldsymbol{\mu}_i$ σε σύκριση με το κέντρο βάρους οποιασδήποτε άλλης συστάδας. Δηλαδή, κάθε σημείο $\mathbf{x}_j$ αντιστοιχίζεται στη συστάδα C_{j^*}, όπου

$$j^* = \arg\min_{i=1}^{k} \{||\mathbf{x}_j - \boldsymbol{\mu}_i||^2\} \tag{3.2}$$

Για ένα καθορισμένο σύνολο συστάδων C_i, *i=1,...,k*, στο δεύτερο βήμα του αλγορίθμου υπολογίζονται οι νέες μέσες τιμές για κάθε συστάδα από τα σημεία του συνόλου C_i. Τα βήματα της αντιστοίχισης σε συστάδες και της ενημέρωσης των κέντρων βάρους εκτελούνται επαναληπτικά μέχρι να καταλήξουμε σε ένα σταθερό σημείο ή σε τοπικά ελάχιστα. Πιο συγκεκριμένα, θεωρούμε ότι ο αλγόριθμος K μέσων έχει συγκλίνει αν τα

κέντρα βάρους δεν αλλάζουν από τη μία επανάληψη στην επόμενη. Για παράδειγμα, ο αλγόριθμος σταματάει την εκτέλεσή του αν ισχύει $\sum_{i=1}^{k} ||\mathbf{\mu}_i^t - \mathbf{\mu}_i^{t-1}||^2 \leq \varepsilon$, όπου ε>0 είναι το κατώφλι σύγκλισης (convergence threshold), το *t* συμβολίζει την τρέχουσα επανάληψη και το $\mathbf{\mu}_i^t$ συμβολίζει τον μέσο για τη συστάδα C_i στην επανάληψη *t*.

3.1.2 Ψευδοκώδικας K μέσων

3.2 Συσσωρευτική Ιεραρχική Συσταδοποίηση

Στη συσσωρευτική ιεραρική συσταδοποίηση, ξεκινάμε με μια ξεχωριστή συστάδα για καθένα από τα *n* στοιχεία. Κατόπιν συγχωνεύουμε επανειλημμένα τις δύο πλησιέστερες συστάδες. Η διαδικασία σταματά όταν όλα τα σημεία ανήκουν πλέον στην ίδια συστάδα. Από μαθηματική άποψη, αν δίνεται ένα σύνολο συστάδων $C = \{C_1, C_2, ..., C_m\}$, βρίσκουμε το ζεύγος των *πλησιέστερων συστάδων* C_i και C_j και τις συγχωνεύουμε σε μία νέα συστάδα $C_{ij} = C_i \cup C_j$. Στη συνέχεια ενημερώνουμε το σύνολο των συστάδων διαγράφοντας τις συστάδες C_i και C_j, και προσθέτοντας στη συστάδα τη συστάδα $C = (C \backslash \{C_i, C_j\}) \cup \{C_{ij}\}$. Επαναλαμβάνουμε τη διαδικασία έως ότου το σύνολο C να περιέχει μόνο μία συστάδα. Επειδή το πλήθος των συστάδων μειώνεται κατά ένα σε κάθε βήμα, η διαδικασία αυτή παράγει μια ακολουθία *n* ένθετων συσταδοποιήσεων. Μπορούμε να σταματήσουμε τη διαδικασία συγχώνευσης όταν υπάρχουν ακριβώς *k* εναπομείνασες συστάδες.

3.2.1 Απόσταση μεταξύ συστάδων

Αναφέρθηκε ότι βρίσκουμε το ζεύγος των πλησιέστερων συστάδων. Για τον υπολογισμό της απόστασης μεταξύ δύο οποιωνδήποτε συστάδων μπορούν να χρησιμοποιηθούν αρκετά μέτρα τα οποία βασίζονται συνήθως στην ευκλείδεια απόσταση δύο σημείων. Τα βασικότερα μέτρα είναι:

Μοναδικός σύνδεσμος
Αν δίνονται δύο συστάδες δύο συστάδες C_i και C_j η μεταξύ τους απόσταση ορίζεται ως η ελάχιστη απόσταση ενός σημείου της συστάδας C_i από ένα σημείο της συστάδας C_j:

$$\delta(C_i, C_j) = \min\{\delta(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Πλήρης σύνδεςσμος
Η απόσταση δύο συστάδων ορίζεται ως η μέγιστη απόσταση μεταξύ ενός σημείου της συστάδας C_i και ενός σημείου της συστάδας C_j:

$$\delta(C_i, C_j) = \max\{\delta(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Μέσος όρος ομάδας
Η απόσταση δύο συστάδων ορίζεται ως ο μέσος όρος της απόστασης ανά ζεύγη μεταξύ σημείων της συστάδας C_i και της συστάδας C_j:

$$\delta(C_i, C_j) = \frac{\sum_{x \in C_i} \sum_{y \in C_j} \delta(\mathbf{x}, \mathbf{y})}{n_i \cdot n_j}$$

όπου $n_i = |C_i|$ είναι το πλήθος των σημείων της συστάδας C_i.

Μέση απόσταση
Η απόσταση δύο συστάδων ορίζεται ως η απόσταση των μέσων, ή κέντρων βάρους, των συστάδων:

$$\delta(C_i, C_j) = \delta(\mathbf{\mu}_i, \mathbf{\mu}_j)$$

όπου $\mu_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$

Μέθοδος του Ward
Η απόσταση δύο συστάδων ορίζεται ως η αύξηση που παρουσιάζει το άθροισμα των τετραγώνων των σφαλμάτων ότι οι δύο συστάδες συγχωνευθούν:

$$\delta(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j$$

Μπορούμε να δείξουμε ότι η απόσταση για το μέτρο του Ward απλοποιείται στην παρακάτω σχέση:

$$\delta(C_i, C_j) = \left(\frac{n_i \cdot n_j}{n_i + n_j} \right) ||\mathbf{\mu}_i - \mathbf{\mu}_j||^2$$

3.2.2 Υπολογιστική πολυπλοκότητα

Κεφάλαιο 4

Παρουσίαση Αποτελεσμάτων

[]

Παράρτημα Α΄

TEST

Write your Appendix content here.

Βιβλιογραφία

[1] W. Dehnen and J. Read. N-body simulations of gravitational dynamics. *The European Physical Journal Plus*, 126(55):1–28, 2011.