Μηχανική μάθηση

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2022

Περίληψη

Μέσος όρος τιμών

1. Ο μέσος όρος ορίζεται ως το άθροισμα των στοιχείων προς το πλήθος των στοιχείων.
2. Είναι πάντα δεκαδικός αριθμός
3. Μπορεί να γραφεί

$$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

Διακύμανση

- Δίνει ένα μέτρο της διαφοράς των στοιχείων από την μέση τιμή τους.
- Δηλαδή πόσο διαφορετικά είναι τα στοιχεία μεταξύ τους.
- Μπορεί να γραφεί ως

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{(n-1)}$$

- Είναι πάντα θετικός αριθμός.

Συνδιακύμανση

- Μετράει την συσχέτιση δύο διαφορετικών διανυσμάτων.
- Θεωρούμε πως και τα δύο διανύσματα έχουν το ίδιο πλήθος στοιχείων.
- Μπορεί να γραφεί ως

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{(n-1)}$$

Πράξεις διανυσμάτων

- Η **πρόσθεση** δύο διανυσμάτων $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$ και $\overrightarrow{y} = (y_1, y_2, \ldots, y_n)$ γίνεται στοιχείο προς στοιχείο.
- Για να γίνει πρόσθεση θα πρέπει και τα δύο διανύσματα να έχουν το ίδιο πλήθος στοιχείων.
- Αποτέλεσμα πρόσθεσης: $\overrightarrow{z} = \overrightarrow{x} + \overrightarrow{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$.
- Το **εσωτερικό γινόμενο** διανυσμάτων είναι αριθμός και είναι πολλαπλασιασμός στοιχείου με στοιχείο.
- Αποτέλεσμα εσωτερικού γινομένου: $P = \overrightarrow{x}^T \overrightarrow{y} = (x_1 y_1, x_2 y_2, \ldots, x_n y_n)$.
- Ο πολλαπλασιασμός αριθμού με δίανυσμα γίνεται πολλαπλασιάζοντας κάθε στοιχείο του διανύσματος με αυτόν τον αριθμό. Το αποτέλεσμα είναι και πάλι διάνυσμα.

Νόρμες διανυσμάτων (ιδιότητες)

- Μια νόρμα $\|.\| : R^n \rightarrow R$ είναι μια συνάρτηση με τις εξής ιδιότητες
  - $\|x\| \geq 0$
  - $\|x\| \Leftrightarrow x = (0, 0, \ldots, 0)$
  - $\|ax\| = |a| \|x\|, \ \forall a \in R$
  - $\|x + y\| \leq \|x\| + \|y\|$

Παραδείγματα νόρμας

1. Η νόρμα l_1:
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
2. Η νόρμα l_2:
$$\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$$
3. Η νόρμα l_∞:
$$\|x\|_\infty = \max_i |x_i|$$

Ορισμοί

- Ένας πίνακας θεωρείται μια δίαταξη m γραμμών και n στηλών.
- Συμβολίζεται ως : $A = [a_{ij}]_{m \times n}$
- Αν $m = n$, ο πίνακας ονομάζεται **τετραγωνικός**.

Ειδικοί πίνακες

- Ενας τετραγωνικός πίνακας για τον οποίο έχει μη μηδενικά στοιχεία μόνο στην κύρια διαγώνιό του ονομάζεται **διαγώνιος** πίνακας.
- Ένας διαγώνιος πίνακας A για τον οποίο ισχύει $A_{ii} = 1$, ονομάζεται μοναδίαιος πίνακας και συμβολίζεται με I.
- Ένας διαγώνιος πίνακας με μη - μηδενικά στοιχεία πάνω από την κύρια διαγώνιο ονομάζεται **άνω τριγωνικός**.
- Ένας διαγώνιος πίνακας με μη - μηδενικά στοιχεία κάτω από την κύρια διαγώνιο ονομάζεται **κάτω τριγωνικός**.

Πρόσθεση πινάκων

- Για να γίνει πρόσθεση πινάκων A και B θα πρέπει να έχουν τον ίδιο αριθμό γραμμών και στηλών.
- Το αποτέλεσμα της πράξης:
  $C = A + B \Leftrightarrow C_{ij} = A_{ij} + B_{ij},\ i = 1, .., n,\ j = 1, ..m$
- Ισχύουν τα εξής:
  - $A + B = B + A$
  - $(A + B) + C = A + (B + C)$
  - $A + 0 = A$ (0 είναι ο πίνακας με μηδέν σε κάθε στοιχείο του).
  - $A + (-A) = 0$

Πολλαπλασιασμός στοιχείου με πίνακα

- Γίνεται πολλαπλασιασμός κάθε στοιχείου με τον αριθμό και το αποτέλεσμα: $\lambda \cdot A = \lambda A_{ij},\ i = 1, .., n,\ j = 1, ..m$
- Ισχύουν τα ακόλουθα:
  - $(k + l) \cdot A = k \cdot A + l \cdot A$
  - $k \cdot (A + B) = k \cdot A + k \cdot B$
  - $k \cdot (l \cdot A) = (kl) \cdot A$
  - $1 \cdot A = A$

Πολλαπλασιασμός πινάκων

- Το γινόμενο μεταξύ πινάκων A, B επιτρέπεται μόνο ο A είναι πίνακας $m \times n$ και ο B είναι πίνακας $n \times k$.
- Το αποτέλεσμα του γινομένου είναι πίνακας $m \times k$.
- Το αποτέλεσμα ορίζεται ως:

$$C = AB \Rightarrow [c_{ij}] = \left[\sum_{p=1}^{n} A_{ip} B_{pj} \right]$$

Αντίστροφος πίνακα

- Αν υπάρχει πίνακας Β για τον τετραγωνικό πίνακα Α για τον οποίο ισχύει: $AB = I$, τότε ο πίνακας Β συμβολίζεται με A^{-1} και ονομάζεται αντίστροφος του Α.
- Αν δεν υπάρχει αντίστροφος πίνακας για τον Α, τότε ο Α ονομάζεται ιδιάζων (singular).

Θετικά ορισμένοι πίνακες

1. Θεωρούμε πίνακες συμμετρικούς, $A \in R^{n \times n}$
2. Ένας συμμετρικός πίνακας λέγεται **θετικά ορισμένος** αν: $x^T Ax > 0, \ \forall x \neq 0$
3. Ένας συμμετρικός πίνακας λέγεται **θετικά ημιορισμένος** αν: $x^T Ax \geq 0, \ \forall x \neq 0$

Ορίζουσα πίνακα

1. Ορίζεται για τετραγωνικούς πίνακες.
2. Συμβολίζεται με det(A)
3. Είναι δεκαδικός αριθμός
4. Παρέχει σημαντικές πληροφορίες για ένα πίνακα και έχει πολλές εφαρμογές
5. Μπορεί να υπολογιστεί αναδρομικά.

Υπολογισμός ορίζουσας για 2X2 πίνακα

- Έστω ο πίνακας

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

- $det(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$
- Η ορίζουσα ορίζεται ως $det(A) = a_{11}a_{22} - a_{21}a_{12}$

Υπολογισμός ορίζουσας για 3X3 πίνακα

- Έστω ο πίνακας

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- $det(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$

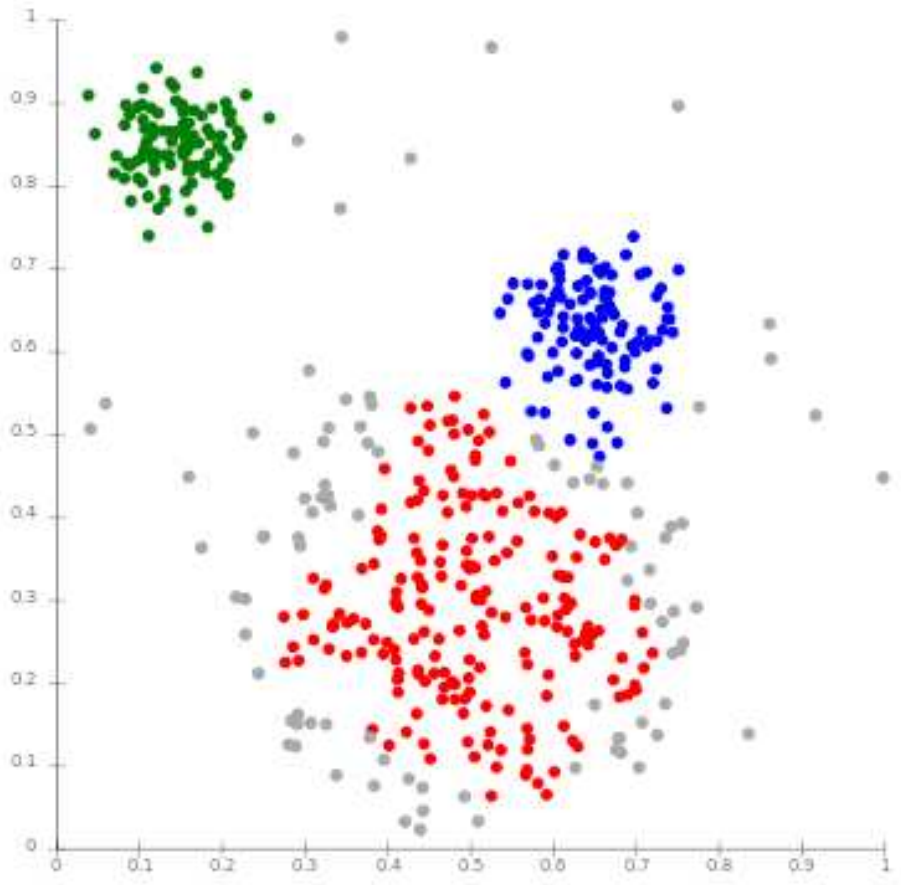
Βασικά στοιχεία ομαδοποίησης

- Στην ομαδοποίηση "έξυπνοι" αλγόριθμοι χρησιμοποιούνται για την κατάταξη δεδομένων σε ένα προκαθορισμένο αριθμό κατηγοριών.
- Για παράδειγμα αν διαθέτουμε πολλά δείγματα κρασιών και θέλουμε να τα κατατάξουμε σε κατηγορίες με βάση χαρακτηριστικά τους (πχ οξύτητα, χρώμα κτλ)
- Συνήθως ο αριθμός των κατηγοριών είναι εκ των προτέρων γνωστός.

Κριτήρια σχηματισμού ομάδας

1. Η ομάδα να ειναι ομοιογενής, δηλαδή τα στοιχεία που απαρτίζουν μια ομάδα να είναι όσο το δυνατόν πιο κοντά μεταξύ τους
2. Οι ομάδες να απέχουν, δηλαδή να μην είναι "κοντά", γιατί αλλιώς θα πρέπει να ενωθούν σε μια.

Παράδειγμα ομάδων



Κριτήρια ομοιότητας

Για να μπορέσουμε να αξιολογήσουμε πόσο κοντά βρίσκονται τα δεδομένα θα πρέπει να υπάρχει κάποιο κριτήριο ομοιότητας. Σε όλες τις εκφράσεις που ακολουθούν ο αριθμός n εκφράζει την διάσταση (πλήθος χαρακτηριστικών) κάθε προτύπου. Μερικά γνωστά κριτήρια παρουσιάζονται στην συνέχεια.

Ευκλείδια απόσταση.

- Σε όλες τις εκφράσεις θεωρούμε πως στους τύπους έχουμε διανύσματα
- Η μεταβλητή n αναπαριστά την διάσταση του προβλήματος (διάσταση προτύπων)
- Είναι το πιο γνωστό κριτήριο απόστασης.

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

Πίνακας ευκλίδειας απόστασης

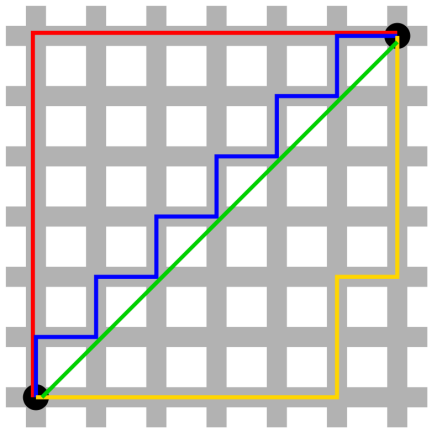
1. Είναι ο πίνακας των αποστάσεων μεταξύ διανυσμάτων.
2. Χρησιμοποιείται σε πολλές μεθόδους.
3. $A = (a_{ij}), \ a_{ij} = d_{ij}^2 = \|x_i - x_j\|^2$
4. $A_{ij} = 0, \ \forall i = j$
5. $A_{ij} = A_{ji}$, συμμετρικός πίνακας
6. $A_{ij} \geq 0$

Απόσταση Manhattan.

1. Βασίζεται σε παρατηρήσεις σχετικά με τις αποστάσεις στην τετραγωνισμένη περιοχή του Μανχάττσαν
2. Η απόσταση αυτή δίνεται από την εξίσωση

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

Απόσταση Manhattan (σχήμα)



Κόκκινο χρώμα: Απόσταση Manhattan distance. Πράσινο: Απευθείας μετακίνηση. Λοιπά χρώματα: Ισοδύναμες αποστάσεις Manhattan

Μέγιστης διαφοράς.

1. Το κριτήριο αυτό βασίζεται στην εύρεση της μέγιστης διαφοράς σε όλες τις διαστάσεις των προτύπων
2. Χρησιμοποιείται αρκετά τακτικά όπως και της Ευκλείδιας απόστασης.
3. Δίνεται από την εξίσωση

$$D(x, y) = max_{i=1}^{n} |x_i - y_i| \tag{3}$$

Συνημιτονοειδής ομοιότητα.

Ορίζεται από την εξίσωση

$$D(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4}$$

Ο αλγόριθμος Kmeans

1. **Αρχικοποίηση** των K κέντρων $c_i,\ i = 1..K$, όπου K είναι το εκτιμούμενο πλήθος ομάδων.
2. **Επανέλαβε**
   1. $S_i = \{\},\ i = 1..K$
   2. Εύρεση της ομάδας που ανήκει το κάθε στοιχείο $x_i,\ \iota = 1..N$ ως ακολούθως: α) εύρεση $j^* = \min_{i=1}^{K} \{D(x_i, c_j)\}$ β) $S_{j^*} = S_{j^*} \cup x_i$
   3. Έστω M_j το πλήθος των μελών της ομάδας. Ανανέωση του κέντρου της ομάδας

$$c_j = \frac{1}{M_j} \sum_{i=1}^{M_j} x_i \tag{5}$$

   όπου x_i τα μέλη της ομάδας.
3. **Αν** τα κέντρα δεν έχουν αλλάξει **τότε τερματισμός**, **αλλιώς** μετάβαση στο βήμα 2.

Kmeans

1. Η μέθοδος είναι δοκιμασμένη και χρησιμοποιείται και στα τεχνητά νευρωνικά δίκτυα RBF
2. Δεν παίζει ρόλο η αρχικοποίηση, μπορεί να γίνει και τυχαία.
3. Η απόδοση του αλγορίθμου εξαρτάται σε μεγάλο βαθμό από την επιλογή του K.

Ο αλγόριθμος των πλησιεστέρων γειτόνων

1. Για κάθε πρότυπο x_i δημιούργησε την λίστα $L(x_i)$ με τους k κοντινότερους γείτονες.
2. Για κάθε ζεύγος σημείων x_i και x_j
   1. Αν $L(x_i) \cap L(x_j) \geq M$, τοποθέτησε τα δύο σημεία x_i και x_j στην ίδια ομάδα
3. Η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχουν πλέον άλλα σημεία εκτός ομάδας.

Αν και αυτός ο αλγόριθμος δεν απαιτεί εκ των προτέρων γνώση για το πλήθος των ομάδων στηρίζεται στις παραμέτρους k και M.

Σύνοψη

- Παρουσιάστηκαν βασικές έννοιες μαθηματικών.
- Παρουσιάστηκαν έννοιες ομαδοποίησης τιμών.