Υπολογιστική Νοημοσύνη

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2025

Περίληψη

1. Ομαδοποίηση
   - Βασικά στοιχεία
   - Κριτήρια ομοιότητας
2. Ο αλγόριθμος KNN
   - Βασικός αλγόριθμος
   - Επεκτάσεις
3. Δένδρα απόφασης
4. Ο αλγόριθμος KMEANS

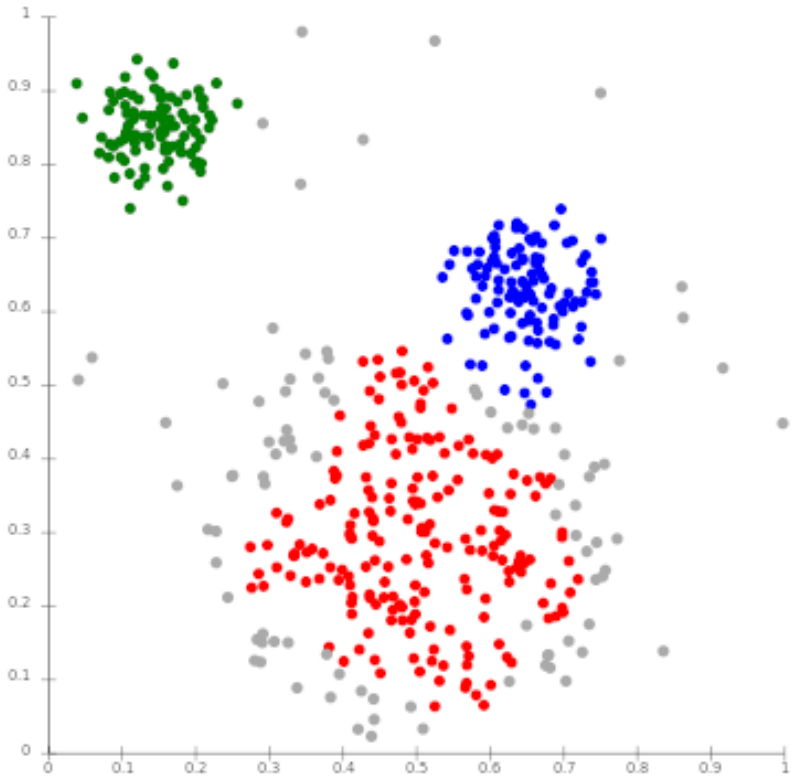
Ορισμός

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)

Προϋποθέσεις

1. Η ομάδα να είναι 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



Παράδειγμα ομαδοποίησης

Παραδείγμα.

- Δεδομένα που ανήκουν σε τρεις κατηγορίες
- Σκοπός της μεθόδου θα πρέπει να είναι ο διαχωρισμός των γκρίζων σημείων σε κάποια από τις περιοχές αυτές
- Κάποια σημεία είναι ξεκάθαρο σε ποια ομάδα πρέπει να μπουν
- Κάποια σημεία βρίσκονται ανάμεσα στις περιοχές και δεν είναι ξεκάθαρο σε ποια περιοχή πρέπει να μπουν
- Σε κάποιες περιπτώσεις ενδεχομένως να χρειαστεί να αλλάξει το πλήθος των ομάδων (αύξηση ή μείωση).

Ορισμοί.

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.

Ευκλείδεια 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

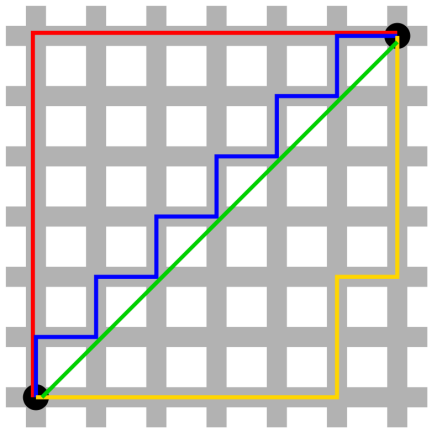
1. Είναι ο πίνακας των αποστάσεων μεταξύ διανυσμάτων.
2. Χρησιμοποιείται σε πολλές μεθόδους.
3. $A = (a_{ij}),\ a_{ij} = d_{ij}^2 = \|x_i - x_j\|^2$
4. $A_{ij} = 0,\ \forall i = j$
5. $A_{ij} = A_{ji}$, συμμετρικός πίνακας
6. $A_{ij} \geq 0$

Απόσταση Manhattan.

1. Βασίζεται σε παρατηρήσεις σχετικά με τις αποστάσεις στην τετραγωνισμένη περιοχή του **Manhattan**
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 Μανηατταν διστανσε. Πράσινο: Απευθείας μετακίνηση. Λοιπά χρώματα: Ισοδύναμες αποστάσεις Μανηατταν

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

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

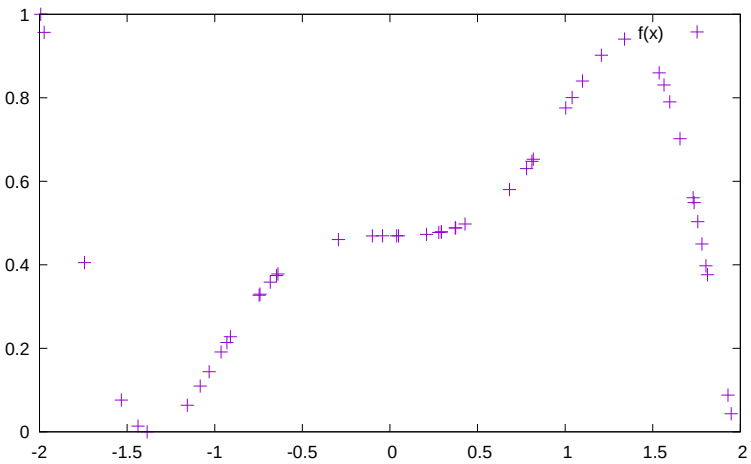
Ορισμοί

- Η μέθοδος αναπτύχθηκε από τους Fix και Hodges το 1951[1].
- Μη παραμετρική μέθοδος, δηλαδή δεν υπάρχουν παράμετροι που πρέπει να εκτιμηθούν.
- Τα δεδομένα κατατάσσονται στην κατηγορία που πλειοψηφεί ανάμεσα στους K κοντινότερους γείτονες τους.
- Παίζει σημαντικό ρόλο η τιμή του K (αριθμός γειτόνων)
- Παίζει λιγότερο σημαντικό ρόλο το είδος της απόστασης που θα χρησιμοποιηθεί.
- Μπορεί να χρησιμοποιηθεί τόσο για κατηγοριοποίηση δεδομένων όσο και για μάθηση συναρτήσεων.
- Είναι ανεκτικός αλγόριθμος σε παρουσία θορύβου και όταν ακόμα λείπουν τιμές από χαρακτηριστικά.

KNN για κατηγοριοποίηση

1. Δημιουργία συνόλου εκπαίδευσης $S = \{X_1, X_2, ..., X_N\}$, όπου $X_i \in R^d$
2. Καθορισμός της παραμέτρου K. Συνήθως οι τιμές αυτής της παραμέτρου είναι μονοί αριθμοί.
3. Για κάθε νέο πρότυπο X_i
   1. Δημιουργία του συνόλου S_x με τους K κοντινότερους γείτονες από το σύνολο S. Για την εύρεση των γειτόνων χρησιμοποιούνται διάφορα κριτήρια απόστασης με το πιο συνηθισμένο την **Ευκλείδια** απόσταση.
   2. Εύρεση της κατηγορίας Υ που πλειοψηφεί στο σύνολο S_x
   3. Ανάθεση του προτύπου στην κατηγορία Υ.
4. Ο αλγόριθμος βασίζεται στο K. Επίσης είναι σχετικά αργός αλγόριθμος, αφού απαιτεί ταξινόμηση για κάθε πρότυπο.
5. Πιθανή λύση η δημιουργία πίνακα απόστάσεων (όπως ο πίνακας Ευκλείδιας απόστασης).

Μάθηση συναρτήσεων

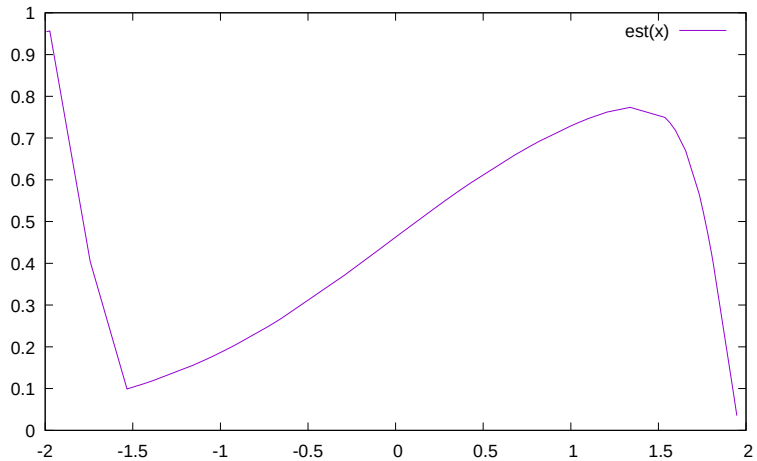
1. Με τον όρο μάθηση συναρτήσων μιλάμε για εύρεση της καμπύλης (συνάρτησης) που πιθανόν να βρίσκεται πίσω από δεδομένα.
2. Σκοπός ενός μοντέλου που κάνει μάθηση συναρτήσεων είναι η εκτίμηση της συνάρτησης που περνά από αυτά τα σημεία αλλά και από άλλα ενδιάμεσα σημεία και πιθανόν και από σημεία εκτός του διαστήματος.
3. 

Χρήση KNN για μάθηση συναρτήσεων

1. Δημιουργία συνόλου εκπαίδευσης $S = \{X_1, X_2, ..., X_N\}$, όπου $X_i \in R^d$
2. Καθορισμός της παραμέτρου K. Συνήθως οι τιμές αυτής της παραμέτρου είναι μονοί αριθμοί.
3. Για κάθε νέο πρότυπο X_i
   1. Δημιουργία του συνόλου S_x με τους K κοντινότερους γείτονες από το σύνολο S. Για την εύρεση των γειτόνων χρησιμοποιούνται διάφορα κριτήρια απόστασης με το πιο συνηθισμένο την **Ευκλείδια** απόσταση.
   2. Υπολογισμός της τιμής

$$Y(x) = \frac{1}{K} \sum_{i=1}^{K} X_i, \forall\ X_i \in S_x$$

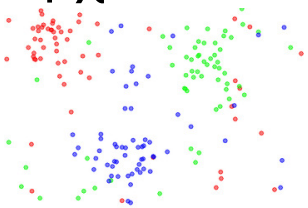
Παράδειγμα μάθησης συναρτήσεων



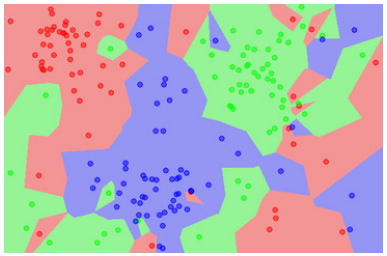
Για το προηγούμενο σύνολο δεδομένων με χρήση KNN

Παράδειγμα εύρεσης κατηγοριών

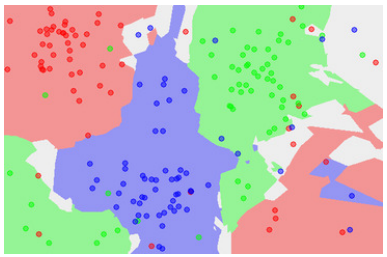
1. Αρχικό σύνολο δεδομένων



2. Χρήση ενός γείτονα



3. Χρήση 5 γειτόνων



Ιδιότητες

1. Για K=1 δεν έχουμε τόσο καλή συμπεριφορά, καθώς κάνει πολλά λάθη
2. Για μεγαλύτερες τιμές του K, με K περιττό παρατηρείται καλύτερη συμπεριφορά
3. Για πολύ μεγάλες τιμές γίνονται πάλι λάθη, καθώς συμμετέχουν στην ψηφοφορία και πολύ μακρινά σημεία

KNN με χρήση βαρών

- Έχει παρουσιάσει καλύτερες ικανότητες μάθησης σε αρκετά παραδείγματα
- Συσχετίζουμε κάθε γείτονα με ένα βάρος
- Έχουν αναπτυχθεί αρκετές τεχνικές KNN με βάρη
- Στην συνέχεια παρουσιάζεται η τεχνική Inverse Weighted KNN[2]

KNN με βάρη

1. Για ένα πρότυπο x_i εύρεση των αποστάσεων $d_j, j = 1..K$
2. Για κάθε απόσταση d_j υπολογισμός της ποσότητας $V_j = \frac{1}{d_j}$
3. Υπολογισμός των βαρών $w_j = \frac{V_j}{\sum V_k}$
4. Ανάθεση του προτύπου στην κατηγορία με το μεγαλύτερο άθροισμα βαρών

Ομαδοποίηση πλησιέστερων 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.

Αυτός ο αλγόριθμος στηρίζεται στις παραμέτρους k και M

Η τεχνική Radius Neighbors Classifier/Regressor

1. Radius Neighbors Classifier/Regressor
2. Ο χρήστης ορίζει μια απόσταση R.
3. Ο αλγόριθμος εντοπίζει **όλους** τους γείτονες που βρίσκονται σε απόσταση R από το πρότυπο.
4. Το πρότυπο ανατίθεται στην κατηγορία που πλειοψηφεί σε αυτό το σύνολο γειτόνων.

Περίληψη

1. Κατάλληλα για δημιουργία κανόνων απόφασης
2. Κατάλληλα για Data mining
3. Απαιτούν την ύπαρξη θετικών και αρνητικών περιπτώσεων για την δημιουργία κανόνων απόφασης.

Αναπαράσταση

1. Τα δένδρα απόφασης είναι μοντέλα της μηχανικής μάθησης που χρησιμοποιούνται στην ταξινόμηση δεδομένων.
2. Σε αυτά τα μοντέλα δημιουργείται ένα σύνολο κανόνων απόφασης σε δενδρική δομή
3. Στους εσωτερικούς κόμβους του δένδρου βρίσκονται χαρακτηριστικά από το πρόβλημα
4. Στα φύλλα βρίσκονται αποφάσεις, δηλαδή η κατηγορία που θα επιλεχθεί.

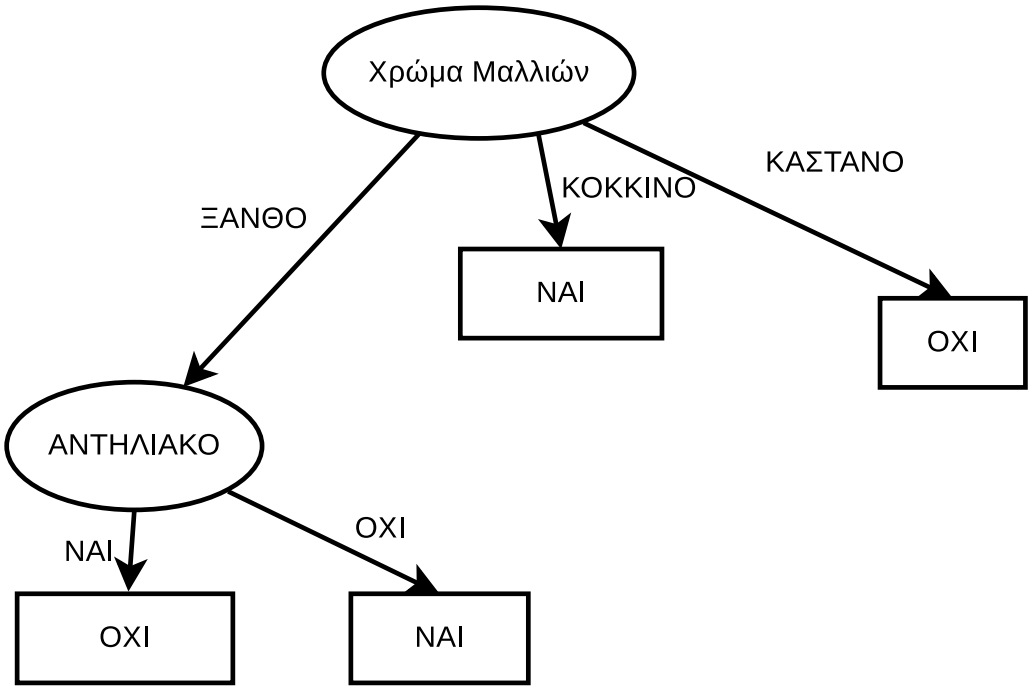
Παράδειγμα εφαρμογής εγκαυμάτων

Όνομα	Μαλλιά	Ύψος	Βάρος	Αντηλιακό	Κάηκε
Σάρα	Ξανθά	Μέτριο	Ελαφρύ	Όχι	Ναι
Άννα	Ξανθά	Ψηλό	Μέτριο	Ναι	Όχι
Νίκος	Καστανά	Κοντό	Μέτριο	Ναι	Όχι
Άλεξανδρος	Ξανθά	Κοντό	Μέτριο	Όχι	Ναι
Νίκη	Κόκκινα	Μέτριο	Βαρύ	Όχι	Ναι
Τάκης	Καστανά	Ψηλό	Βαρύ	Όχι	Όχι
Καίτη	Καστανά	Μέτριο	Βαρύ	Όχι	Όχι
Γιάννης	Ξανθά	Κοντό	Ελαφρύ	Ναι	Όχι

Παράδειγμα κατηγορίας εγκαυμάτων

1. Οι στήλες ΟΝΟΜΑ, ΜΑΛΛΙΑ, ΥΨΟΣ, ΒΑΡΟΣ, ΑΝΤΗΛΙΑΚΟ είναι χαρακτηριστικά.
2. Η στήλη ΚΑΗΚΕ είναι η επιθυμητή κατηγορία.
3. Η πρώτη στήλη ΟΝΟΜΑ δεν αναμένενεται να έχει κάποιο αποτέλεσμα στην μάθηση και μπορεί να αφαιρεθεί.

Ενδεικτικό σχήμα δένδρου για εγκαύματα



Κανόνες απο το παραπάνω δένδρο

Από το δένδρο αυτό μπορούν να εξαχθούν οι επόμενοι κανόνες:

1. Αν κάποιος έχει κόκκινο χρώμα μαλλιών **καίγεται**.
2. Αν κάποιος έχει καστανό χρώμμα μαλλιών **δεν καίγεται**.
3. Αν κάποιος έχει ξανθό χρώμα μαλλιών και δεν φορά αντηλιακο **καίγεται**.
4. Αν κάποιος έχει ξανθό χρώμα μαλλιών και φορά αντηλιακό τότε **δεν καίγεται**.

Προφανώς σε αυτό το δένδρο απόφασης δεν έχουν ληφθεί υπόψη χαρακτηριστικά τα οποία δεν έχουν σχέση με το αποτέλεσμα όπως το βάρος και το ύψος κάτι που στις περισσότερες περιπτώσεις δεν συμβαίνει.

Αλγόριθμος κατασκευής δένδρου

Ο τυπικός αλγόριθμος κατασκευής δένδρων έχει ως ακολούθως:

1. Έστω $x_1, x_d, ..., x_n$ τα χαρακτηριστικά του προβλήματος.
2. Έστω S τα δεδομένα εκπαίδευσης.
3. Επιλογή ενός χαρακτηριστικού x_k από το σύνολο των χαρακτηριστικών του προβλήματος
4. Για κάθε μία επιλογή $f_1, f_2, .., f_M$ του χαρακτηριστικού x_k κάνε
   1. Αν τα δεδομένα που περιέχουν μόνον στο χαρακτηριστικό x_k το f_i κατατάσσονται σε μια κατηγορία, τότε βάλε στο δένδρο σαν τερματικό φύλλο αυτή την τιμή της κατηγορίας
   2. Αλλιώς δημίουργησε νέο υποδένδρο με κάποιο άλλο χαρακτηριστικό (που δεν το έχουμε λάβει ήδη) και θέσε σαν ρίζα του την τιμή f_i
5. Η παραπάνω διαδικασία συνεχίζεται μέχρι να δημιουργηθούν μόνο τερματικοί κόμβοι.

Ο αλγόριθμος ID3

1. Ο παραπάνω αλγόριθμος επιλέγει τα χαρακτηριστικά με τυχαίο τρόπο και πολλές φορές αυτό δεν είναι αποδοτικό.
2. Για το παράδειγμα των εγκαυμάτων θα οδηγούσε στην επιλογή όλων των χαρακτηριστικών ακόμα και αυτών που δεν συνεισφέρουν κάτι στην επιλογή της σωστής κατηγορίας.
3. Σε αυτήν την περίπτωση θα οδηγούσε σε μεγάλο δένδρο με αργό χρόνο εκπαίδευσης και απόκρισης.
4. Ο αλγόριθμος ID3 επιλέγει κάθε φορά το καλύτερο χαρακτηριστικό βρίσκοντας την συνεισφορά του στο τελικό αποτέλεσμα μέσω της εντροπίας (entropy) και του κέρδους πληροφορίας (information gain).

Εντροπία

- Η εντροπία για τα δεδομένα στο S δίνεται από την εξίσωση:
- $$H(S) = \sum_{c \in C} -p(c) \log_2 p(c) \tag{5}$$

  όπου
  - S είναι τα δεδομένα στο σύνολο εκπαίδευσης
  - ῾ είναι το σύνολο των κατηγορίων, στο παράδειγμα με το έγκαυμα είναι (ΝΑΙ,ΟΧΙ)
  - π(ς) είναι το κλάσμα των δεδομένων εκπαίδευσης που ανήκουν στην κατηγορία ς από το σύνολο Σ.
- Για το παράδειγμα με τα εγκαύματα η εντροπία είναι

$$-\frac{3}{8} \log_2 \left(\frac{3}{8}\right) - \frac{5}{8} \log_2 \left(\frac{5}{8}\right) = 0.954434$$

καθώς 3 από τα 8 ανήκουν στην κατηγορία ΝΑΙ και 5 από τα 8 στην κατηγορία ΟΧΙ.

Κέρδος πληροφορίας

- Το κέρδος πληροφορίας για ένα χαρακτηριστικό A από το σύνολο Σ ορίζεται ως:

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

όπου

- H(S) είναι η εντροπία του συνόλου Σ, όπως υπολογίστηκε παραπάνω.
- T είναι τα υποσύνολα του S αν το διαχωρίσουμε με βάση τις διαφορετικές τιμές που λαμβάνει το χαρακτηριστικό A, $S = \cup_{t \in T} t$
- p(t) το κλάσμα των δεδομένων που βρίσκονται στο t προς τον συνολικό αριθμό δεδομένων που είναι στο S.
- H(t) είναι η εντροπία του υποσυνόλου t.

Παράδειγμα υπολογισμού κέρδους

- Για παράδειγμα ας υπολογίσουμε στο παράδειγμα το κέρδος από το χαρακτηριστικό **Βάρος**
  - IG(Βάρος)=H(S)-p(Βάρος=ελαφρύ)H(Βάρος=ελαφρύ)-p(Βάρος=Μέτριο)H(Βάρος=Μέτριο)-p(Βάρος=Βαρύ)H(Βάρος=Βαρύ) = 0.9544-0.25-0.344361-0.344361=0.015678
- Το κέρδος από το χαρακτηριστικό **Μαλλιά** είναι
  - IG(Μαλλιά)=H(S)-p(Μαλλιά=Ξανθά)H(Μαλλιά=Ξανθά)-p(Μαλλιά=Καστανά)H(Μαλλιά=Καστανά)-p(Μαλλιά=Κόκκινα)H(Μαλλιά=Κόκκινα)=0.9544-0.5-0.0-0.0=0.50
- Επομένως το χαρακτηριστικό **Μαλλιά** είναι πιθανότερο να επιλεγεί από ότι το χαρακτηριστικό **βάρος**.

Ορισμοί

1. Ο αλγόριθμος KMEANS χρησιμοποιείται για την δημιουργία εκπροσώπων από ομάδες.
2. Έχει υλοποιηθεί αρχικά από τον MacQueen[3].
3. Το πλήθος των ομάδων θεωρείται δεδομένο.
4. Έχουν αναπτυχθεί δεκάδες παραλαγές του αλγορίθμου από τότε.

Ο αλγόριθμος

1. **Αρχικοποίηση** των K κέντρων $c_i,\ i = 1..K$, όπου K είναι το εκτιμούμενο πλήθος ομάδων. Κάθε κέντρο c_i θεωρούμε πως έχει n στοιχεία, όπου n είναι η διάσταση των προτύπων εισόδου.
2. **Επανέλαβε**
   1. $S_i = \{\},\ i = 1..K$
   2. Εύρεση της ομάδας που ανήκει το κάθε στοιχείο: α) εύρεση $j^* = \min_{i=1}^{K} \{D(x_i, c_j)\}$ β) $S_{j^*} = S_{j^*} \cup x_i$
   3. Ανανέωση του κέντρου της ομάδας

$$c_j = \frac{1}{M_j} \sum_{x_i \in S_j} x_i \tag{6}$$

   όπου M_j το πλήθος των μελών της ομάδας θ.
3. **Αν** τα κέντρα δεν έχουν αλλάξει **τότε τερματισμός**, **αλλιώς** μετάβαση στο βήμα 2.

Σύνοψη

- Παρουσιάστηκαν η έννοια της κατηγοριοποίησης
- Παρουσιάστηκε ο αλγόριθμος KNN και οι επεκτάσεις του
- Δόθηκε μια σύντομη παρουσίαση του αλγορίθμου KMEANS

Βιβλιογραφία I


Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas.

https://visualstudiomagazine.com/articles/2019/04/01/weighted-k-nn-classification.aspx


MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281-297, 1967.