Μηχανική μάθηση

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2022

Περίληψη

Χαρακτηριστικά

- Κάθε τιμή που αναπαριστά μια ιδιότητα
- Τα χαρακτηριστικά μπορούν να είναι
  - Συνεχείς τιμές
  - Διακριτές τιμές
  - Αλφαριθμητικές τιμές

Παραδείγματα

- Η θερμοκρασία από έναν αισθητήρα (συνεχής τιμή)
- Η ηλικία ενός ανθρώπου (διακριτή τιμή)
- Η πιστοληπτική ικανότητα ενός δανειολήπτη (αλφαριθμητική τιμή)

Μετατροπές τιμών

- Οι αλφαριθμητικές τιμές αν είναι πεπερασμένες σε πλήθος μετατρέπονται σε διακριτές πχ. με απαρίθμηση.
- Οι συνεχείς τιμές μπορούν να μετατραπούν σε διακριτές τιμές με χρήση ορίων.
- Συνήθως οι διακριτές τιμές είναι κατάλληλες για δένδρα απόφασης, ενώ οι συνεχείς είναι περισσότερο κατάλληλες σε τεχνητά νευρωνικά δίκτυα.

Παραδείγματα μετατροπής

- Μετατροπή χρώματος φρούτων:
  - αρχικές τιμές: ΚΙΤΡΙΝΟ, ΠΡΑΣΙΝΟ, ΚΟΚΚΙΝΟ
  - Διακριτές τιμές: 0, 1, 2
- Συνήθως δεν υπάρχει κάποιο θέμα με τις αριθμητικές τιμές που επιλέγονται

Παράδειγμα μετατροπής

- Συνεχείς τιμές θερμοκρασιών πχ 16.7
  - Αρχικές τιμές: Συνεχείς τιμές στο διάστημα [-20,40]
  - Δημιουργία 6 ομάδων:
    [-20,-10],(-10,0],(0,10],(10,20],(20,30],(30,40].
  - Ανάθεση σε κάθε ομάδα μιας διακριτής τιμής: [0,1,2,3,4,5]
  - Για παράδειγμα η θερμοκρασία 16.7 είναι στην τέταρτη ομάδα και έτσι παίρνει την τιμή 3.
- Είναι κρίσιμο το εύρος του διαστήματος και σε πολλές περιπτώσεις απαιτείται και η συμβουλή ενός ειδικού στο πεδίο για τον καθορισμό του.

Πρότυπα

- Τα πρότυπα είναι σύνολα χαρακτηριστικών.
- Κάθε πρότυπο είναι ένα μια ξεχωριστή καταγραφή.
- Δεν είναι υποχρεωτικό όλα τα χαρακτηριστικά να είναι αποκλειστικά συνεχή ή αποκλειστικά διακριτα.
- Το σύνολο προτύπων ονομάζεται Dataset.
- Συνήθως μαζί με κάθε πρότυπο υπάρχει και ένας χαρακτηρισμός όπως για παράδειγμα η ποιότητα ενός μπουκαλιού κρασιού.

Το dataset Lenses

- Χρησιμοποιείται για να διαχωρίσει τα φακούς που πρέπει να φορέσουν άτομα με προβλήματα όρασης.
- 4 Χαρακτηριστικά κανονικοποιημένα
  - Ηλικία ασθενούς σε 3 κλίμακες (1-νέος, 2-για προ πρεσβυωπία, 3-για μετά από πρεσβυωπία). **Σημείωση**: εδώ χρειάστηκε η γνώμη του ειδικού για την κλίμακα των ηλικιών.
  - Διάγνωση: 1-μυωπία, 2-πρεσβυωπία
  - Αστιγματισμός: 1-όχι, 2-ναι
  - Παραγωγή δακρύων: 1-μειωμένο, 2-κανονικό
- 3 πιθανές κατηγορίες
  - 1-ο ασθενής χρειάζεται γυαλιά με πολλούς βαθμούς, 2-ο ασθενής χρειάζεται γυαλιά για μειωμένους βαθμούς, 3-ο ασθενής δεν χρειάζεται γυαλιά.

Το dataset lenses

- Οι τρεις πρώτες εγγραφές
-

ΗΛΙΚΙΑ	ΔΙΑΓΝΩΣΗ	ΑΣΤΙΓΜΑΤΙΣΜΟΣ	ΔΑΚΡΥΑ	ΚΑΤΗΓΟΡΙΑ
1	2	1	1	3
1	1	1	2	2
1	1	2	1	3

Ιστοσελίδες με πρότυπα

- `https://archive.ics.uci.edu/ml/index.php` UCI, το παλαιότερο και πιο ενημερωμένο.
1. `https://www.kaggle.com/datasets`. Kaggle, το πιο σύγχρονο με πολλούς διαγωνισμούς.

Ιστοσελίδες με πρότυπα

- `https://archive.ics.uci.edu/ml/index.php` UCI, το παλαιότερο και πιο ενημερωμένο.
1. `https://www.kaggle.com/datasets`. Kaggle, το πιο σύγχρονο με πολλούς διαγωνισμούς.

Χαμένες τιμές σε πρότυπα(missing values)

- Η έλλειψη τιμών σε ορισμένα χαρακτηριστικά.
- Προκύπτει από λαθός καταχωρήσεις πολλές φορές
- Μπορεί να προκύψει από δεδομένα στα οποία έχουν γίνει κατα λάθος διαγραφές
- Πολλές φορές προκαλείται από αστοχία υλικού σε περίπτωση αισθητήρων για παράδειγμα

Παράδειγμα χαμένων τιμών

- Πιστοληπτική ικανότητα
-

Ετήσιο εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι

Τρόποι επίλυσης

1. Διαγραφή ολόκληρης της γραμμής. Μπορεί να μειώσει αρκετά τις εγγραφές και δεν χρησιμοποιείτα συχνά.
2. Αναζήτηση της πραγματικής τιμής. Αυτό μπορεί να γίνει από τον ειδικό που έφτιαξε το σύνολο δεδομένων.
3. Χρήση σταθεράς στις χαμένες τιμές. Αντικατάσταση χαμένων τιμών με κάποια σταθερά πχ 0.0 αλλά μπορεί να προκαλέσει θόρυβο στα δεδομένα.
4. Αντικατάσταση με τον μέσο όρο. Αντικαθίστανται οι χαμένες τιμές με τον μέσο όρο της στήλης. Είναι η πιο κοινή μέθοδος.

Θόρυβος στα δεδομένα

- Παρουσία λανθασμένων τιμών στα χαρακτηριστικά.
- Πιθανή λανθασμένη εισόδου τιμών από τον χρήστη.
- Πιθανή επίσης και η κακή λειτουργία συσκευών που καταγράφουν δεδομένα (πχ συσκευών ανάγνωσης ετικετών RFID)
- Πιθανόν και από προβλήματα στη μετάδοση των δεδομένων μέσω ενός δικτύου.
- Σε πολλές περιπτώσεις παρουσιάζονται και δεδομένα με ακραίες τιμές (πολύ μικρές ή πολύ μεγάλες) τα οποία δεν βοηθούν τον αλγόριθμο μηχανικής μάθησης, καθώς περιγράφουν σπάνιες και μεμονωμένες περιπτώσεις.

Αντιμετώπιση θορύβου

- Μια λύση είναι η διαγραφή των γραμμών που περιέχουν θόρυβο
- Μια δεύτερη λύση είναι η αντικατάσταση με άλλες τιμές, για παράδειγμα με μέσους όρους

Αντιμετώπιση θορύβου με κετακερματισμό

- Σε αυτήν την περίπτωση οι τιμές ενός χαρακτηριστικού ταξινομούνται και χωρίζονται σε διαστήματα. Στην συνέχεια επιλέγεται
  1. Είτε να αντικατασταθούν οι τιμές με τον μέσο όρο κάθε διαστήματος
  2. Είτε να γίνει αντικατάσταση των οριακών τιμών, δηλαδή κάθε τιμή αν είναι κοντά στην μικρότερη τιμή αντικαθίσταται με αυτήν αλλιώς με την μεγαλύτερη τιμή του διαστήματος.

Αντιμετώπιση θορύβου με στατιστική

- Σε αυτήν την περίπτωση εντοπίζονται οι ακραίες τιμές με στατιστική.
- Για κάθε χαρακτηριστικό x υπολογίζεται ο μέσος όρος μ_x και η τυπική απόκλιση σ_x.
- Αν υπάρχουν τιμές μικρότερες εκτός του διαστήματος $[\mu_x - k * \sigma_x, \mu_x + k * \sigma_x]$ θεωρούνται ακραίες.

Ορισμός

- Η κανονικοποίηση είναι μια διαδικασία στην οποία αριθμητικά δεδομένα αντικαθίστανται από άλλα πιο κατάλληλα για την μέθοδο μηχανικής μάθησης που χρησιμοποιείται.
- Για παράδειγμα στα τεχνητά νευρωνικά δίκτυα, η μέθοδος εκπαίδευσης του δικτύου αποκρίνεται καλύτερα αν τα δεδομένα είναι στο διάστημα [0,1].

Κανονικοποίηση ελαχίστου - μεγίστου

- Η Για παράδειγμα έστω ότι το χαρακτηριστικό x έχει ελάχιστο x_{min} και μέγιστο x_{max}.
- Αν θέλουμε η νέα μεταβλητή να έρθει στο διάστημα $[a, b]$, τότε αυτό μπορεί να γίνει με την ακόλουθη γραμμική σχέση

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} (b - a) + a$$

Κανονικοποίηση z - score

- Σε αυτήν την περίπτωση για κάθε χαρακτηριστικό x υπολογίζεται ο μέσος όρος μ_x και η τυπική απόκλιση σ_x.
- Στην συνέχεια γίνεται η κλιμάκωση

$$x' = \frac{x - \mu_x}{\sigma_x}$$

Κανονικοποίηση δεκαδικής κλιμάκωσης

- Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε εξαιρετικά μεγάλες τιμές.
- Γίνεται διαίρεση των τιμών των μεταβλητών με δυνάμεις του 10.

$$x' = \frac{x}{10^k}, \ k = 1, 2, 3...$$

Ο βασικός αλγόριθμος

- 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.

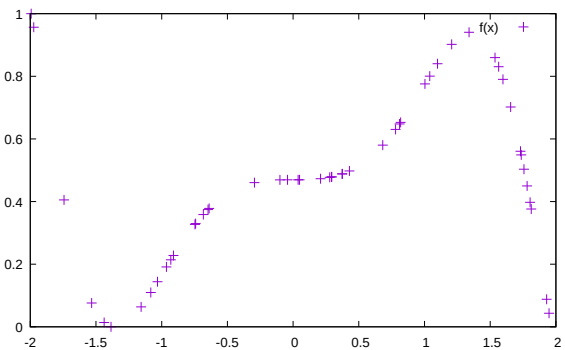
Χρήση KNN για κατηγοριοποίηση

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).

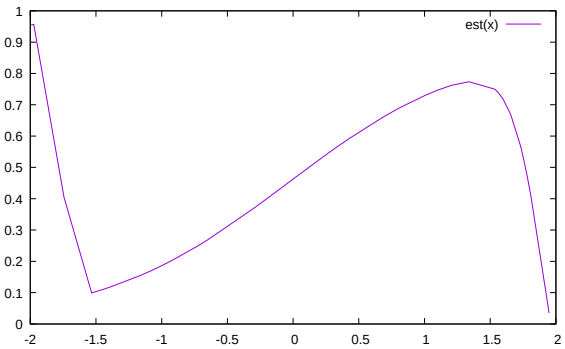
Επίδραση του K στην μάθηση

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

Χρήση KNN για 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. 

Παράδειγμα μάθησης συναρτήσεων

- Τι έμαθε το KNN;
- 

Σύνοψη

- Παρουσιάστηκαν βασικές έννοιες χαρακτηριστικών και προτύπων.
- Παρουσιάστηκαν έννοιες θορύβου και κανονικοποίησης
- Αναλύθηκε η μέθοδος KNN

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.

Βιβλιογραφία II


A. Georgouli, Μηχανική μάθηση, διαθέσιμο από https://repository.kallipos.gr/handle/11419/3382