Υπολογιστική Νοημοσύνη

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2021

Περίληψη

1. Δεδομένα
   - Χαρακτηριστικά
   - Πρότυπα
2. Προεπεξεργασία τιμών
   - Χαμένες τιμές
   - Θόρυβος στα δεδομένα
   - Κανονικοποίηση δεδομένων

Ορισμός.

- 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.

Παραδείγμα 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

Παραδείγμα 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.

Ορισμοί.

- 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 για το συγκεκριμένο dataset

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

Ιστοσελίδες με πρότυπα.

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

Ορισμοί

- 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 αισθητήτων για παράδειγμα

Παράδειγμα χαμένων τιμών

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

Τρόποι επίλυσης χαμένων τιμών

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.

Ορισμοί

- 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

Ορισμοί

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].

Κανοκοποίηση ελαχίστου -μεγίστου

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, όπου γίνεται διαίρεση των τιμών των μεταβλητών με δυνάμεις του 10.

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

Σύνοψη

- Παρουσιάστηκαν οι έννοιες των χαρακτηριστικών και των προτύπων
- Είναι θεμελιώδεις έννοιες στην Υπολογιστική Νοημοσύνη
- Παρουσιάστηκαν προβλήματα δεδομένων, όπως οι χαμένες τιμής και η παρουσιά θορύβου
- Παρουσιάστηκαν τρόποι κανονικοποίησης δεδομένων.

Βιβλιογραφία Ι

- Ιωάννης Μπούταλης και Γεώργιος Συρακούλης, Υπολογιστική Νοημοσύνη & Εφαρμογές, Εκδόσεις Κρίκος.
- Ηλιάδης, Λάζαρος Σ., Υπολογιστική νοημοσύνη και ευφυείς πράκτορες, Εκδόσεις Τζιόλα.