Υπολογιστική Νοημοσύνη

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2025

Περίληψη

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.
3. `https://sci2s.ugr.es/keel/datasets.php`. KEEL Repository.

Διαμόρφωση δεδομένων σε CSV

1. Τα δεδομένα αποθηκεύονται χωριζόμενα με κόμμα ή κάποιον άλλο διαχωριστή (πχ ;)
2. Σε κάθε γραμμή υπάρχουν N+1 δεδομένα, όπου N ο αριθμός των προτύπων και 1 για την επιθυμητή έξοδο.
3. Σε πολλές περιπτώσεις υπάρχει και μια πρόσθετη γραμμή στην αρχή του αρχείου που είναι η επικεφαλίδα με πληροφορίες για κάθε χαρακτηριστικό, όπως το όνομα του.

Διαμόρφωση δεδομένων ARFF

1. Το αρχείο έχει μια σειρά από ενότητες.
   1. Στην πρώτη ενότητα παρουσιάζονται τα χαρακτηριστικά, το όνομα τους και η ιδιότητά τους (αριθμοί ή σύμβολα)
   2. Στην δεύτερη ενότητα μετά τον χαρακτηρισμό @Data ακολοθούν σε μορφή CSV τα πρότυπα και η έξοδος
2. Χρησιμοποιείται κυρίως από το λογισμικό WEKA.

Σφάλμα εκπαίδευσης

1. Ορίζεται στην μάθηση με επίβλεψη (γνωστή η επιθυμητή έξοδος)
2. Το σύνολο εκπαίδευσης ορίζεται:
   $T_R = \{x_i, y_i\}, \ i = 1, \ldots, M$
3. Το σφάλμα εκπαίδευσης για ένα μοντέλο μάθησης $M(x)$:

$$E\left(M(x)\right) = \sum_{i=1}^{M} \left(M\left(x_i\right) - y_i\right)^2$$

4. Ακόμα και να μηδενιστεί δεν είναι σίγουρο πως το μοντέλο έχει μάθει σωστά τα πρότυπα (υπερεκπαίδευση).

Σφάλμα ελέγχου

1. Το σύνολο ελέγχου $T_T = \{x_i, y_i\}, \ i = 1, \ldots, N$ έχει δεδομένα που δεν βρίσκονται στο σύνολο εκπαίδευσης
2. Αν πρόκειται για μάθηση συναρτήσεων μας νοιάζει το μέσο τετραγωνικό σφάλμα:

$$E\left(M(x)\right) = \frac{1}{N}\sum_{i=1}^{N}\left(M\left(x_i\right) - y_i\right)^2$$

3. Αν πρόκειται για μάθηση κατηγοριών μας νοιάζει το σφάλμα κατηγοριοποίησης:

$$E(M(x)) = \sum_{i=1}^{N}\left(C\left(M\left(x_i\right)\right) \neq y_i\right)$$

όπου $C(M(x))$ η κατηγορία που προβλέπει το μοντέλο για την είσοδο x.

Πρόβλημα μάθησης σε Imbalanced Data

1. Πιθανόν το μοντέλο να μαθαίνει μόνο τις κατηγορίες με τα περισσότερα πρότυπα και να αγνοεί τις υπόλοιπες
2. Πιθανή λύση να υπάρχει μέσο σφάλμα ανα κατηγορία:

$$E(M(x)) = \frac{1}{N_c} \sum_{i=1}^{N_c} E_i$$

όπου N_c είναι το σύνολο των κατηγοριών και E_i το σφάλμα για την κατηγορία i.

Θετικές και αρνητικές κατηγορίες

1. Για προβλήματα δύο κατηγοριών, y_i η εκτιμώμενη κατηγορία και t_i η πραγματική.
2. Θετική κατηγορία: $t_i = 1$ Αρνητική κατηγορία: $t_i = 0$
   1. TRUE NEGATIVE: $y_i = 0, t_i = 0$ (ανήκει όντως στην αρνητική κατηγορία).
   2. FALSE NEGATIVE $y_i = 0, t_i = 1$ (ανήκει στην θετική και μπήκε στην αρνητική κατηγορία).
   3. FALSE POSITIVE: $y_i = 1, t_i = 0$ (ανήκει στην αρνητική και μπήκε στην θετική κατηγορία)
   4. TRUE POSITIVE $y_i = 1, t_i = 1$ (ανήκει όντως στην θετική κατηγορία

Πίνακας σύγχυσης (confusion matrix)

1. accuracy $= \frac{TN+TP}{TN+TP+FN+FP}$
2. Πίνακας σύγχυσης με αριθμό προτύπων που ανήκουν στις παραπάνω κατηγορίες

TN	FP
FN	TP

3. Ιδανικά θέλουμε FN=FP=0, άρα accuracy=1

Πρόβλημα accuracy

1. Έστω ένα πρόβλημα πχ γρίπης με 10 άτομα με γρίπη και 990 χωρίς γρίπη.
2. Έστω TP=1, FN=9, FP=1, TN=989
3. accuracy $= \frac{TN+TP}{TN+TP+FN+FP} = \frac{989+1}{989+1+9+1} = 0.99(99\%)$
4. Δεν μπορεί να αναγνωρίσει σωστά την κατηγορία της γρίπης
5. Χρειάζονται πιο εύστοχες μετρικές

Precision και Recall

1. Precision (Πόσες από τις προβλέψεις που το μοντέλο έκανε ως "θετικές" είναι πράγματι σωστές) :
   $\text{precision} = \frac{TP}{POSITIVE} = \frac{TP}{TP+FP}$
2. Recall (Πόσα από τα πραγματικά θετικά δείγματα αναγνώρισε σωστά το μοντέλο): $\text{recall} = \frac{TP}{\text{CLASS1}} = \frac{TP}{TP+FN}$
3. Συμβιβασμός των παραπάνω ειναι το f1 score:
   $F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Geometric mean

1. Εναλλακτική μετρική που χρησιμοποιείται σε μεθόδους βελτιστοποίησης
2. $G_{\text{mean}} = \sqrt{\text{Precision} \times \text{Recall}}$
3. Χρησιμοποιείται κυρίως όταν τα δεδομένα που διατίθενται δεν είναι ισορροπημένα μεταξύ των κατηγοριών.

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.

Η τεχνική SMOTE

1. Χρησιμοποιείται για να εξισορροπήσει δεδομένα, όταν οι κατηγορίες είναι άνισα κατανεμημένες (το πιο σύνηθες).
2. Εφαρμόζεται μόνον στην φάση της εκπαίδευσης
3. Για κάθε δείγμα x από την κατηγορία με τα λιγότερα πρότυπα
   1. Εντοπίζονται οι k κοντινότεροι γείτονές του.
   2. Επιλέγεται τυχαία ένας από τους γείτονες x_N
   3. Δημιουργείται ένα νέο πρότυπο ανάμεσα στο αρχικό δείγμα και τον γείτονα αυτό: $x' = x + r \times (x - x_N)$, με $r \in [0, 1]$.

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

1. 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

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

2. Πρέπει να ειναι γνωστά τα άκρα
3. Τα άκρα θα πρέπει να είναι γνωστά και στο train και στο test set.

Κανονικοποίηση z-score

1. 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

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

2. Θα πρέπει να ειναι γνωστή η κατανομή κάθε χαρακτηριστικού.
3. Προσφέρει πιο ομαλή κλιμάκωση από ότι η προηγούμενη κλιμάκωση.

Κανονικοποίηση δεκαδικής κλιμάκωσης

1. Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε εξαιρετικά μεγάλες τιμές, όπου γίνεται διαίρεση των τιμών των μεταβλητών με δυνάμεις του 10.

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

2. Δεν απαιτεί γνώση ορίων
3. Είναι πιο δίκαιη κλιμάκωση
4. Μπορεί να οδηγήσει σε πολύ χαμηλές τιμές τα χαρακτηριστικά.

Κανονικοποίηση εξόδων

1. Τα μοντέλα μηχανικής μάθησης συνήθως αποδίδουν καλύτερα αν οι έξοδοι είναι σε κάποιο δεδομένο διάστημα
2. Πολλές φορές η κανονικοποίηση των εξόδων, πχ σε [0,1] προσδίδει καλύτερη απόδοση σε μηχανισμούς μάθησης και μειώνει τα αριθμητικά σφάλματα

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
- Παρουσιάτηκαν τρόποι κανονικοποίησης δεδομένων.

Βιβλιογραφία I

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