

Τεχνολογίες Ευφυούς Διαχείρισης Ανθρωπιστικών Δεδομένων

Ταξινόμηση

Instance-based Learning

Μάθηση Βασισμένη στα

Παραδείγματα

Instance-based Learning

- Ονομάζεται και
 - Memory-based Learning
 - Case-based Learning
 - Lazy Learning
- Τα δεδομένα εκπαίδευσης διατηρούνται αυτούσια...
 - ...σε αντίθεση με τις άλλες μεθόδους μηχανικής μάθησης οι οποίες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή.
- Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την κατηγορία μιας νέας περίπτωσης, εξετάζει εκείνη τη στιγμή τη σχέση της με τα ήδη αποθηκευμένα παραδείγματα.
- Κάνει την παραδοχή ότι τα διάφορα παραδείγματα μπορεί να αναπαρασταθούν ως σημεία σε κάποιον n -διάστατο Ευκλείδειο χώρο R^n όπου n ο αριθμός των χαρακτηριστικών (ανεξάρτητων μεταβλητών).
- Κάθε νέα περίπτωση τοποθετείται στο χώρο αυτό ως νέο σημείο και η τιμή του n προσδιορίζεται με βάση το χαρακτηρισμό των k γειτονικών σημείων.

Ο αλγόριθμος των Πλησιέστερων Γειτόνων

The Nearest Neighbors Algorithm

- Μια συνάρτηση απόστασης καθορίζει την μάθηση
- Συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση (Euclidean distance)
- Αν έχω δυο παραδείγματα, το (1) και το (2), και k χαρακτηριστικά, η ΕΑ μεταξύ τους είναι

$$- EA(1,2) = \sqrt{(a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + \dots + (a_k^{(1)} - a_k^{(2)})^2}$$

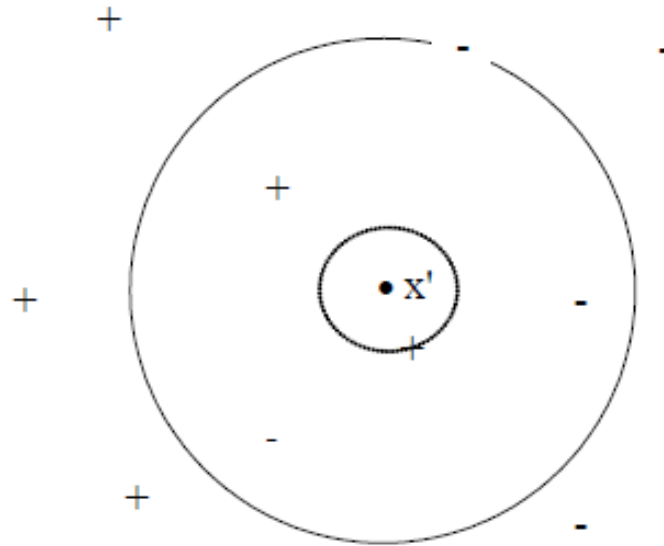
- $a_k^{(1)}$: η τιμή του χαρα/κού k παράδειγμα (1)
 - $a_k^{(2)}$: η τιμή του χαρα/κού k παράδειγμα (2)
- Άλλη συνάρτηση απόστασης: Manhattan ή City-block distance
 - $M(1,2) = |a_1^{(1)} - a_1^{(2)}| + |a_2^{(1)} - a_2^{(2)}| + \dots + |a_k^{(1)} - a_k^{(2)}|$

Ο αλγόριθμος των Πλησιέστερων Γειτόνων

The Nearest Neighbors Algorithm

- Η πιο απλή υλοποίηση του αλγορίθμου:
- Τα παραδείγματα εκπαίδευσης σαρώνονται γραμμικά, και υπολογίζεται η απόσταση του καθενός από το παράδειγμα ελέγχου.
- 1-NN: Όποιο παράδειγμα εκπαίδευσης έχει την μικρότερη απόσταση από το παράδειγμα ελέγχου, είναι ο πιο κοντινός γείτονας.
 - Το παράδειγμα ελέγχου παίρνει σαν τιμή στην κλάση ταξινόμησης ό,τι τιμή έχει στην κλάση ταξινόμησης ο πιο κοντινός γείτονας
- k-NN: Τα k παραδείγματα εκπαίδευσης που έχουν την μικρότερη απόσταση από το παράδειγμα ελέγχου, είναι οι k πιο κοντινοί γείτονες.
 - Το παράδειγμα ελέγχου παίρνει σαν τιμή στην κλάση ταξινόμησης ό,τι τιμή έχει στην κλάση ταξινόμησης η πλειοψηφία των k πιο κοντινών γειτόνων (**majority voting**)

Παράδειγμα



$k=1$ το x' χαρακτηρίζεται θετικό
 $k=5$ το x' χαρακτηρίζεται αρνητικό

Κανονικοποίηση - Normalization

- Για να μην έχω ανισορροπία στις κλίμακες των αριθμητικών μεταβλητών πρέπει να τις κανονικοποιήσω ώστε να έχουν εύρος τιμών μέσα στο διάστημα $[0,1]$.
- Έτσι, η τιμή που βάζω στον τύπο του μέτρου απόστασης είναι
- $$a_i = (v_i - \min_{v_i}) / (\max_{v_i} - \min_{v_i})$$



πραγματική (αρχική)
τιμή του χαρ/κού i στο
παράδειγμα



μέγιστη τιμή που
παίρνει το χαρ/κο
 i στα δεδομένα



ελάχιστη τιμή που
παίρνει το χαρ/κο i
στα δεδομένα

Πλεονεκτήματα-Μειονεκτήματα

- Ο ταξινομητής μπορεί να επικαιροποιηθεί (επεκταθεί), δηλ. να εκπαιδευτεί σε περισσότερα δεδομένα, με την απλή προσθήκη καινούριων παραδειγμάτων εκπαίδευσης.
- Σε πολλές περιπτώσεις είναι πολύ ακριβής η ταξινόμηση
 - Εξαιρέσεις
- Δεν πραγματοποιείται εκπαίδευση (δεν επάγεται κάποιο γενικευμένο μοντέλο)
- Αφιερώνει πολύ χρόνο για τον υπολογισμό των αποστάσεων
- Θεωρεί ότι όλα τα χαρακτηριστικά είναι ισοβαρή (έχουν την ίδια σημασία για την μάθηση της έννοιας)
 - Αυτό μπορεί να αντιμετωπιστεί με απόδοση βαρών στα χαρακτηριστικά (weighting)

Παράδειγμα

- Έστω ένα σύστημα που ταξινομεί μια ελιά αυτόματα αν είναι μελάνωμα ή όχι.
- Τα χαρακτηριστικά που παίζουν ρόλο είναι το χρώμα, το σχήμα, η υφή και το μέγεθος της ελιάς
- Έστω τα παρακάτω παραδείγματα εκπαίδευσης

	ΧΡΩΜΑ	ΣΧΗΜΑ	ΥΦΗ	ΜΕΓΕΘΟΣ	ΜΕΛΑΝΩΜΑ
1	ΜΑΥΡΟ	ΣΤΡΟΓΓΥΛΟ	ΛΕΙΑ	ΜΕΓΑΛΗ	ΝΑΙ
2	ΜΑΥΡΟ	ΑΚΑΝΟΝΙΣΤΟ	ΌΧΙ ΛΕΙΑ	ΜΙΚΡΗ	ΝΑΙ
3	ΚΑΦΕ	ΑΚΑΝΟΝΙΣΤΟ	ΛΕΙΑ	ΜΙΚΡΗ	ΌΧΙ
4	ΚΑΦΕ	ΣΤΡΟΓΓΥΛΟ	ΌΧΙ ΛΕΙΑ	ΜΕΣΑΙΑ	ΟΧΙ

- Έστω πως έρχεται προς εξέταση μια ελιά μικρή, λεία, στρογγυλή, καφέ.
- Πώς θα ταξινομήσει την ελιά ο 1-NN;
- Πώς θα ταξινομήσει την ελιά ο 3-NN;
- $EA(1,5)=((M-K)^2+(\Sigma-\Sigma)^2+(\Lambda-\Lambda)^2+(ME-MI)^2)^{1/2}=2^{1/2}$
- $EA(2,5)=((M-K)^2+(A-\Sigma)^2+(O\Lambda-\Lambda)^2+(MI-MI)^2)^{1/2}=3^{1/2}$
- $EA(3,5)=((K-K)^2+(A-\Sigma)^2+(\Lambda-\Lambda)^2+(MI-MI)^2)^{1/2}=1$
- $EA(4,5)=((K-K)^2+(\Sigma-\Sigma)^2+(O\Lambda-\Lambda)^2+(ME\Sigma-MI)^2)^{1/2}=2^{1/2}$