k Πλησιέστεροι γείτονες (kNN)

k Πλησιέστεροι γείτονες

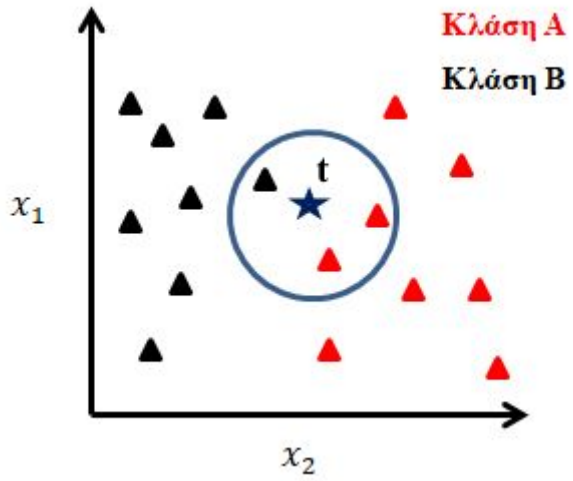
Μια ευρέως χρησιμοποιούμενη τεχνική κατηγοριοποίησης είναι ο αλγόριθμος kNN (K Nearest Neighbors – KNN) ο οποίος βασίζεται στη χρήση μέτρων βασισμένων στην απόσταση. Σύμφωνα με τον αλγόριθμο αυτό, τα διάφορα δείγματα του συνόλου δεδομένων μπορούν να αναπαρασταθούν ως σημεία σε κάποιο n-διαστατο Ευκλείδιο χωρο Rn (όπου n ο αριθμός των χαρακτηριστικών ή αλλιώς των ανεξάρτητων μεταβλητών). Κάθε νέο δείγμα τοποθετείται στο χώρο ως νέο σημείο και η κλάση στην οποία κατηγοριοποιείται προσδιορίζεται με βάση την κλάση στην οποία ανήκουν τα k πλησιέστερα σε αυτό γειτονικά σημεία. Αν k=1, τότε το δείγμα θα ανατεθεί στην κατηγορία που ανήκει ο κοντινότερος γείτονας του.

Βήματα κατηγοριοποίησης

1ο ΒΗΜΑ Αρχικά καθορίζεται η τιμή της σταθερής παραμέτρου k.

2ο ΒΗΜΑ Ο αλγόριθμος αναζητά τα k σημεία που βρίσκονται πλησιέστερα στη νέα παρατήρηση.

3ο ΒΗΜΑ Το νέο στοιχείο τοποθετείται στην κατηγορία που περιέχει τα περισσότερα στοιχεία από το σύνολο των k κοντινότερων στοιχείων.



ατηγοριοποίηση με χρήση kNN

Στο Σχήμα παρουσιάζονται τα τρία κοντινότερα στοιχεία στο σύνολο εκπαίδευσης. Το t θα τοποθετηθεί στην κατηγορία στην οποία ανήκουν τα περισσότερα από αυτά τα k στοιχεία.

Όπως όλοι οι κατηγοριοποιητές έτσι και ο kNN έχει κάποια πλεονεκτήματα και αντίστοιχα μειονεκτήματα.

Πλεονεκτήματα

- Είναι αποτελεσματικός όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών
- Ο αλγόριθμος του είναι απλός
- Έχει πετύχει υψηλές επιδόσεις κατηγοριοποιήσης σε πολλές περιπτώσεις.

Μειονεκτήματα

- Επειδή γίνονται πολλές συγκρίσεις μεταξύ παρατηρήσεων, απαιτεί πολύ αποτελεσματικές τεχνικές καταλαγοποιήσης (indexing).
- Σε περιπτώσεις νέων παρατηρήσεων και ειδικότερα όταν ο αριθμός των εν δυνάμει γειτόνων είναι μεγάλος, η κατηγοριοποιήση διαρκεί πολύ περισσότερο.
- Είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων
- Είνα ευαίσθητοι στην ύπαρξη μη σημαντικών μεταβλητών εισόδου.
- Το πλήθος των γειτόνων k μπορεί να επηρεάσει σημαντικά τα αποτελέσματα