Bayes

Bayesian Κατηγοριοποιήση

Η μέθοδος αυτή είναι μια από τις πιο γνωστές και πρακτικές μεθόδους και συγκρίνεται επάξια με όλες τους αλγορίθμους που ανήκουν στις παραπάνω κατηγορίες. Η κατηγοριοποιήση αυτή βασίζεται στα θεωρήματα του Bayes και υποθέτει ότι υπάρχει ανεξαρτησία ανάμεσα στις παραμέτρους των γνωρισμάτων. Σύμφωνα με τη μέθοδο αυτή στόχος είναι να κατηγοριοποιηθεί ένα δείγμα X σε μια από τις δεδομένες κατηγορίες $C1, C2 \ldots Cn$. Για κάθε κλάση C δεδομένου του κάθε δείγματος x_i, είναι γνωστή η εκ των προτέρων πιθανότητα (prior probability) $P(x_i|C)$,σε ένα σύνολο εκπαίδευσης, όπως επίσης γνωστή είναι και η εκ των προτέρων πιθανότητα $P(C)$ της κάθε κλάσης C για K διαφορετικές κλάσεις. Για να προβλέψει την κλάση ενός δείγματος x_i, ο αλγόριθμος Bayes υπολογίζει τις πιθανότητες για την κάθε κλάση και εκχωρεί το γνώρισμα στην κλάση με τη μεγαλύτερη πιθανότητα $P(C|x_i)$.

Τα βήματα που ακολουθούνται για να ανατεθεί ένα νέο δείγμα άγνωστης κλάσης σε μια κλάση είναι :

1ο ΒΗΜΑ: Υπολογισμός των εκ των προτέρων πιθανοτήτων $P(C)$ και δεσμευμένων πιθανοτήτων $P(x_i|C)$ από το σύνολο δεδομένων εκπαίδευσης.

- Η πιθανότητα να ανήκει κάποιο δείγμα x_i στην κλάση C ορίζεται ως εξής :

$$P(C) = \frac{N_c}{N}$$

όπου N_c είναι το πλήθος των δειγμάτων που ανήκουν στην κλάση C και N ο συνολικός αριθμός δειγμάτων του συνόλου εκπαίδευσης για K διαφορετικές τιμές της μεταβλητής κλάσεων.

- Η υπό συνθήκη πιθανότητα, να ανήκει η τιμή f_{ij} του χαρακτηριστικού f_j ενός δείγματος x_i στην κλάση C, δεδομένης της κλάσης C, ορίζεται ως εξής :

$$P(C) = \frac{Nf_{ij}C}{N_c}$$

και υπολογίζεται για κάθε τιμή f_{ij} του χαρακτηριστικό f_j για όλες τις K κλάσεις της μεταβλητής κλάσεων C. ΄Οπου $Nf_{ij}C$ είναι το πλήθος των δειγμάτων που περιέχουν την τιμή f_{ij} του χαρακτηριστικού f_j και ανήκουν στην κλάση C. Αυτός ο ορισμός ισχύει για διακριτές τιμές f_{ij} ενώ στις συνεχείς τιμές χρησιμοποιούνται τεχνικές διακριτικοποίησης.

Αυτές οι πιθανότητες αποτελούν το μοντέλο κατηγοριοποίησης της Bayesian κατηγοριοποίησης που είναι γνωστή και ως μεθόδος κατηγοριοποίησης Naive Bayes.

2ο ΒΗΜΑ: Στο βήμα αυτό γίνεται ο υπολογισμός των πιθανοτήτων, ώστε να ανήκει το συγκεκριμένο (άγνωστο) προς κατηγοριοποίηση δείγμα x_i, στις υπάρχουσες κλάσεις C, δεδομένης της πιθανότητας εμφάνισης των τιμών των χαρακτηριστικών του στις κλάσεις αυτές. Δηλαδή γίνεται ο υπολογισμός των πιθανοτήτων $P(C|x_i)$. Η κλάση που αντιστοιχεί στην μεγαλύτερη πιθανότητα είναι η ζητούμενη, δηλ.

$$C_{target} = arg\ max_{c} P(C|x_i)$$

Με την προϋπόθεση ότι τα χαρακτηριστικά fj είναι ανεξάρτητα μεταξύ τους οπότε τελικά η εξίσωση γίνεται:

$$C_{target} = arg\ max_{c} P(C) \prod_{j=1}^{M} P(C)$$

Η πρόσεγγιση της απλής κατηγοριοποιήσης κατα Bayes έχει αρκετά πλεονεκτήματα.

- Εύκολη στη χρήση της
- Σε αντίθεση με άλλους αλγορίθμους κατηγοριοποιήσης, απαιτείται μόνο ένα πέρασμα των δεδομένων εκπαίδευσης
- Εύκολη στον χειρισμό των ελλειπών δεδομένων.

Από την άλλη πλευρά, παρόλο που η απλοική προσέγγιση του Bayes είναι αρκετά απλή στη χρήση της, δεν δίνει πάντα ικανοποιητικά αποτελέσματα.

- Τα γνωρίσματα δεν είναι ανεξάρτητα
- Δε μπορεί να χειριστεί συνεχή δεδομένα