Θεώρημα Μπέυζ (Bayes)

Λίγα λόγια σχετικά με το θεώρημα Μπέυζ.

Από τη στατιστική εξαγωγή συμπερασμάτων προκύπτει ότι οι πληροφορίες για την κατανομή των δεδομένων προέρχονται από την εξέταση των δεδομένων που ακολουθούν την κατανομή. Εάν έχουμε ένα σύνολο δεδομένων $X=\{\chi_1,_{,.....,}\chi_n\}$, το πρόβλημα για την εξόρυξη γνώσης είναι να ανακαλύψει τις ιδιότητες της κατανομής από την οποία προέρχεται το σύνολο . Ο κανόνας Bayes , είναι μια τεχνική που εκτιμά την πιθανοφάνεια μια ιδιότητας παίρνοντας το σύνολο των δεδομένων σαν απόδειξη ή σαν είσοδο .
Ο κανόνας Bayes μας επιτρέπει να προσδιορίζουμε τις πιθανότητες των υποθέσεων, με δεδομένη την τιμή κάποιου δεδομένου . Εδώ μιλάμε για πλειάδες όπου στη πραγματικότητα κάθε χ_i μπορεί να είναι τιμή ενός γνωρίσματος ή ένα χαρακτηριστικό των δεδομένων .

Ορισμός θεωρήματος

Το θεώρημα Μπέυζ ορίστηκε μαθηματικά ως η ακόλουθη εξίσωση:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

όπου *A* και *B* είναι γεγονότα .

- *P*(*A*) και *P*(*B*) είναι οι πιθανότητες των *A* και *B* που είναι ανεξάρτητα μεταξύ τους.
- *P*(*A* | *B*), η υπό συνθήκη πιθανότητα, είναι η πιθανότητα του *A* δεδομένου του *B* να είναι αληθής.
- *P*(*B* | *A*), είναι η πιθανότητα του *B* δεδομένου του *A* να είναι αληθής.

Μπέυζ και Εξόρυξη δεδομένων.

Θεώρημα Μπέυζ ως μέθοδος κατηγοριοποίησης .

- Bayesian
  Η Bayesian κατηγοριοποίηση αποτελεί μία κατηγορία μεθόδων της κατηγοριοποίησης και βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Αυτο σημαίνει ότι πραγματοποιείται μια πιθανοτική πρόβλεψη, δηλαδή προβλέπει την πιθανότητα ένα δείγμα X να ανήκει σε κάποια κατηγορία. Ο απλούστερος Bayesian κατηγοριοποιητής είναι ο Naïve Bayesian. Αυτός υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των υπόλοιπων γνωρισμάτων. Ο λόγος που γίνεται αυτό είναι για να αποφεύγονται οι πολύπλοκοι υπολογισμοί κατά τη συνθήκη ανεξαρτησίας της κατηγορίας.

Θεώρημα Μπέυζ (Bayes)

- Περιγραφή Naïve Bayesian
  Υποθέτουμε ότι έχουμε ένα σύνολο δεδομένων S και έστω ότι κάθε δείγμα δεδομένων X=(x1,x2,...,xn) με m κατηγορίες C1,C2,..,Cm. Δεδομένου ενός αγνώστου δείγματος δεδομένων X, ο κατηγοριοποιητής θα προβλέψει ότι το X ανήκει στην κατηγορία C που έχει την μέγιστη εκ των υστέρων(posterior) πιθανότητα με βάση το X. Αυτό σημαίνει ότι το X κατηγοριοποιείται στην Ci αν και μόνο αν:

  $$p(Ci|X)>p(Cj|X) \text{ για κάθε } 1\leq j\leq m \text{ και } j\neq i$$

  Ο στόχος, λοιπόν, είναι να βρούμε την μέγιστη posterior πιθανότητα, δηλαδή το μέγιστο p(Ci|X) για κάθε κλάση, με αποτέλεσμα ο Naïve Bayesian κατηγοριοποιητής να έχει υψηλή απόδοση. Η απόδοση του συγκρίνεται με αυτή των δέντρων απόφασης και κάποιους κατηγοριοποιητές που στηρίζονται σε νευρωνικά δίκτυα σε ορισμένες εφαρμογές.