Πανεπιστήμιο Πειραιώς

Καθηγητής
Δρ. Φιλιππάκης Μιχαήλ

Μιχαήλ Ε. Φιλιππάκης

Κατηγοριοποίηση

Κατηγοριοποίηση είναι η τοποθέτηση ενός αντικειμένου σε μια ή περισσότερες **προκαθορισμένες κατηγορίες** (ομάδες) με βάση κάποια χαρακτηριστικά του.

Παραδείγματα

- Εντοπισμός spam email, με βάση τον header ή το περιεχόμενό τους.
- Αξιολόγηση καρκινικών κυττάρων ως καλοήθη ή κακοήθη.
- Χαρακτηρισμός συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης.
- Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού.

Κατηγοριοποίηση (classification)

Bayesian κατηγοριοποίηση

Η Bayesian κατηγοροποίηση είναι μια πιθανοτική προσέγγιση για την επίλυση προβλημάτων κατηγοριοποίησης.

Κεντρική ιδέα

- Για κάθε αντικείμενο και για κάθε κλάση υπολογίζεται (στην συνέχεια θα δούμε πως) η πιθανότητα το αντικείμενο να ανήκει στην κλάση αυτή.
- Το αντικείμενο τοποθετείται στην κλάση εκείνη για την οποία υπολογίσαμε την μεγαλύτερη πιθανότητα να ανήκει σ' αυτή.

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Αν A, B είναι (μη κενά) ενδεχόμενα ενός δειγματικού χώρου Ω τότε οι δεσμευμένες πιθανότητες $P(A|B)$ και $P(B|A)$ ορίζονται αντίστοιχα:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ και } P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Από τους ορισμούς αυτούς προκύπτει **ο τύπος του Bayes**

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

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Ο τύπος του Bayes συνδυάζεται και με το **θεώρημα ολικής πιθανότητας**:

Αν η οικογένεια $(B_i)_{i\in[n]}$ αποτελεί διαμέριση του Ω, τότε για κάθε $A \subseteq \Omega$ ισχύει

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

οπότε προκύπτει ότι

Αν η οικογένεια $(B_i)_{i\in[n]}$ αποτελεί διαμέριση του Ω, τότε για κάθε $A \subseteq \Omega$ ισχύει

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Παράδειγμα

Ένας χρήστης έχει παρατηρήσει ότι από τα email που λαμβάνει καθημερινά ότι το 60% είναι γραμμένα στα Αγγλικά και το υπόλοιπο 40% στα Ελληνικά. Επίσης, έχει παρατηρήσει ότι από τα email γραμμένα στα Αγγλικά το 90% είναι ανεπιθύμητα (spam) και από τα email γραμμένα στα Ελληνικά το 30% είναι ανεπιθύμητα.
Να βρεθεί η πιθανότητα ένα email που διαβάζει ο χρήστης να είναι spam.
Να βρεθεί η πιθανότητα ένα spam email που διαβάζει ο χρήστης να είναι γραμμένο στα Ελληνικά.

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Παράδειγμα

Ένας χρήστης έχει παρατηρήσει ότι από τα email που λαμβάνει καθημερινά ότι το 60% είναι γραμμένα στα Αγγλικά και το υπόλοιπο 40% στα Ελληνικά. Επίσης, έχει παρατηρήσει ότι από τα email γραμμένα στα Αγγλικά το 90% είναι ανεπιθύμητα (spam) και από τα email γραμμένα στα Ελληνικά το 30% είναι ανεπιθύμητα.

Να βρεθεί η πιθανότητα ένα email που διαβάζει ο χρήστης να είναι spam. $P(S)$

Να βρεθεί η πιθανότητα ένα spam email που διαβάζει ο χρήστης να είναι γραμμένο στα Ελληνικά. $P(E / S)$

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Λύση

Έστω A το ενδεχόμενο το email να είναι γραμμένο στα Αγγλικά.
Έστω E το ενδεχόμενο το email να είναι γραμμένο στα Ελληνικά.
Έστω S το ενδεχόμενο το email να είναι spam.
$P(A) = 0.6$, $P(E) = 0.4$
$P(S|A) = 0.9$, $P(S|E) = 0.3$.
Από τον τύπο της ολικής πιθανότητας έχουμε ότι

$$P(S) = P(S|A)P(A) + P(S|E)P(E) = 0.9 \cdot 0.6 + 0.3 \cdot 0.4 = 0.66.$$

Από τον τύπο του Bayes έχουμε ότι

$$P(E|S) = \frac{P(S|E)P(E)}{P(S)} = \frac{0.9 \cdot 0.4}{0.66} = 0.5454 = 54.5\%.$$

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Οι ίδιοι τύποι ισχύουν και τυχαίες μεταβλητές X, Y αντί για ενδεχόμενα A, B

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

και

Τύπος του Bayes

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

Έστω μια οικογένεια αντικειμένων της οποίας κάθε αντικείμενο κωδικοποιείται από ορισμένα χαρακτηριστικά $x_1, x_2, \ldots, x_n, y$, όπου

- $x_1, x_2, \ldots, x_n$ είναι τα γνωρίσματα/ανεξάρτητες μεταβλητές/είσοδος
- y είναι η κατηγορία/εξαρτημένη μεταβλητή/έξοδος

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Στο παράδειγμα, τα γνωρίσματα του αντικείμενου είναι τα πεδία age, income, student, credit rating, buys computer.

Εδώ θέλουμε να προσδιορίσουμε την τιμή του πεδίου buys computer (Yes ή No) με βάση τις τιμές των πεδίων age, income, student, credit rating.

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Παρατήρηση: Σε κάθε αντικείμενο οι τιμές των γνωρισμάτων του μπορούν να θεωρηθούν ως τυχαίες μεταβλητές (τ.μ.).

Για το πεδίο age θεωρούμε την τ.μ. X_1 με τιμές ≤ 30, 31..40, > 40.

Για το πεδίο income θεωρούμε την τ.μ. X_2 με τιμές high, medium, low.

Για το πεδίο student θεωρούμε την τ.μ. X_3 με τιμές yes, no.

Για το πεδίο credit rating θεωρούμε την τ.μ. X_4 με τιμές fair, excellent.

Για το πεδίο buys computer θεωρούμε την τ.μ. Y με τιμές Yes, No.

Στην Bayesian κατηγοριοποίηση, προκειμένου να κατηγοριοποιήσουμε ένα αντικείμενο για το οποίο έχουμε ότι $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$, $X_4 = x_4$ όπου x_1, x_2, x_3, x_4 είναι συγκεκριμένες τιμές των τ.μ. X_1, X_2, X_3, X_4 (π.χ. $x_1 = 31..40$, $x_2 = \text{high}$, $x_3 = \text{no}$, $x_4 = \text{fair}$) υπολογίζουμε (με τη βοήθεια των δεδομένων εκπαίδευσης) τις δεσμευμένες πιθανότητες

$$P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

και

$$P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

Επιλέγουμε ως τιμή της τ.μ. Y (Yes, No) αυτή που έχει την μεγαλύτερη πιθανότητα.

Οι δεσμευμένες πιθανότητες

$$P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

και

$$P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

υπολογίζονται με την βοήθεια του τύπου του Bayes απ' όπου προκύπτει

$$\begin{aligned} &P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\ &\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)} \end{aligned}$$

και

$$\begin{aligned} &P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\ &\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)} \end{aligned}$$

$$P(Y = \text{Yes}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$
$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$
$$P(Y = \text{No}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$
$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$

- Η πιθανότητα $P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ είναι σταθερή και μπορούμε να την αγνοήσουμε στην σύγκριση. Αρκεί να συγκρίνουμε τους αριθμητές.
- Οι πιθανότητες $P(Y = \text{Yes})$ και $P(Y = \text{No})$ υπολογίζονται εύκολα από τα δεδομένα εκπαίδευσης: Είναι το ποσοστό των αντικειμένων που έχουν τιμή Yes και No αντίστοιχα στο πεδίο defaulted borrower.

$$P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$

$$P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Υπάρχουν **δύο βασικοί τρόποι** υπολογισμού:

- **ο απλοϊκός τρόπος** (naive Bayes)
- **δίκτυο πεποίθησης** (Bayes belief network (BBN))

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Βασική παραδοχή:
Για να απλοποίησουμε τους υπολογισμούς θεωρούμε ότι οι τ.μ. X_1, X_2, X_3, X_4 είναι **υπο συνθήκη ανεξάρτητες δοθέντος του** Y δηλαδή αν γνωρίζουμε π.χ. την τιμή της τ.μ. X_1 με δεδομένο ότι γνωρίζουμε την κλάση Y δεν μπορούμε να πούμε τίποτα για την τιμή της τ.μ. X_2. και ισχύει η σχέση

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{yes}) =$$
$$P(X_1 = x_1 | Y = \text{yes}) \cdot P(X_2 = x_2 | Y = \text{yes}) \cdot P(X_3 = x_3 | Y = \text{yes}) \cdot P(X_4 = x_4 | Y = \text{yes})$$

Γενικότερα οι τ.μ. $X_1, X_2, \ldots, X_n$ ονομάζονται **υπο συνθήκη ανεξάρτητες δοθέντος του** Y αν ισχύει η σχέση:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y = y)$$
$$= P(X_1 = x_1 | Y = y) \cdot P(X_2 = x_2 | Y = y) \cdots P(X_n = x_n | Y = y)$$

Προσοχή!

Η παραδοχή για την ανεξαρτησία των γνωρισμάτων σχεδόν ποτέ δεν ισχύει (!!!) αλλά η υπόθεση αυτή λειτουργεί στην πράξη διότι η κατηγοριοποίηση Bayes δεν απαιτεί ακριβείς εκτιμήσεις πιθανοτήτων αλλά αρκεί η μέγιστη πιθανότητα να αντιστοιχεί στην σωστή κλάση.

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πεδίο age θεωρούμε την τ.μ. X_1 με τιμές ≤ 30, 31..40, > 40.
Για το πεδίο income θεωρούμε την τ.μ. X_2 με τιμές high, medium, low.
Για το πεδίο student θεωρούμε την τ.μ. X_3 με τιμές yes, no.
Για το πεδίο credit rating θεωρούμε την τ.μ. X_4 με τιμές fair, excellent.
Για το πεδίο buys computer θεωρούμε την τ.μ. Y με τιμές Yes, No.

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Που θα κατηγοριοποίησουμε ένα άτομο με τα εξής χαρακτηριστικά;

age	income	student	credit rating	buys computer
> 40	high	no	excellent	?

Θα συγκρίνουμε τις πιθανότητες

$$P(Y = \text{Yes} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

και

$$P(Y = \text{No} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

Ισοδύναμα, από τον τύπο του Bayes αρκεί να συγκρίνουμε τα γινόμενα

$$P(X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes})$$

και

$$P(X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No})$$

Χρησιμοποιώντας την υπόθεση της υπό συνθήκη ανεξαρτησίας των X_1, X_2, X_3, X_4 δεδομένου του Y αρκεί να συγκρίνουμε τα γινόμενα

- $P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes})$
- $P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No})$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) =$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = " > 40" | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = " > 40" | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes})$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \frac{1}{189} = 0.005$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) =$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = " > 40" | Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = " > 40" | Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14}$$

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = \frac{24}{875} = 0.027.$$

Συνοψίζοντας:

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{1}{189} = 0.005$$

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{24}{875} = 0.027.$$

Άρα, η πιθανότητα

$$P(Y = \text{Yes} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

είναι μικρότερη από την πιθανότητα

$$P(Y = \text{No} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

Επομένως, το άτομο με τα χαρακτηριστικά:

age	income	student	credit rating	buys computer
> 40	high	no	excellent	?

θα κατηγοριοποιηθεί με No στο γνώρισμα buys computer.

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Σε κάποιες περιπτώσεις μπορεί να συναντήσουμε το εξής πρόβλημα:

Που θα κατηγοριοποιήσουμε ένα άτομο με τα εξής χαρακτηριστικά;

age	income	student	credit rating	buys computer
31..40	high	no	excellent	?

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Κατά τον υπολογισμό του γινόμενου
$P(X_1 = 31..40|Y = \text{No}) \cdot P(X_2 = \text{high}|Y = \text{No}) \cdot P(X_3 = \text{no}|Y = \text{No}) \cdot P(X_4 = \text{excellent}|Y = \text{No}) \cdot P(Y = \text{No})$
παρατηρούμε ότι σύμφωνα με τα δεδομένα, η πιθανότητα
$P(X_1 = 31..40|Y = \text{No})$ είναι 0, οπότε **μηδενίζεται** ολόκληρο το γινόμενο αυτό!

Αν αρχικά

$$P(X_i = x_i | Y = y) = \frac{k}{n}$$

όπου
k ο αριθμός των αντικειμένων με $Y = y$ για τα οποία $X_i = x_i$
n ο συνολικός αριθμός των αντικειμένων με $Y = y$.
Η διορθωμένη πιθανότητα (**σύμφωνα με τον Laplace**) ορίζεται ως

$$P(X_i = x_i | Y = y) = \frac{k + 1}{n + c}$$

όπου
c το πλήθος των διαφορετικών κλάσεων (το πλήθος των διαφορετικών τιμών της Y)
Το αποτέλεσμα τώρα είναι ότι οι πιθανότητες δεν είναι ποτέ μηδέν!

Αν αρχικά

$$P(X_i = x_i | Y = y) = \frac{k}{n}$$

όπου

k ο αριθμός των αντικειμένων με $Y = y$ για τα οποία $X_i = x_i$

n ο συνολικός αριθμός των αντικειμένων με $Y = y$.

Η διορθωμένη πιθανότητα (**σύμφωνα με την m-εκτίμηση**) ορίζεται ως

$$P(X_i = x_i | Y = y) = \frac{k + pm}{n + m}$$

όπου

m το πλήθος των εικονικών εγγραφών με $Y = y$ που εισάγαμε στο δείγμα εκπαίδευσης

p το ποσοστό των εικονικών εγγραφών με $Y = y$ και $X_i = x_i$.

Πρέπει $m > 1$, $p > 0$. Το αποτέλεσμα τώρα είναι ότι οι πιθανότητες δεν είναι ποτέ μηδέν!

Μια άλλη δυσκολία είναι ο χειρισμός των χαρακτηριστικών που λαμβάνουν **συνεχείς τιμές**. Για παράδειγμα, αν έχουμε ως δεδομένα εκπαίδευσης

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Που θα κατηγοριοποίησουμε ένα αντικείμενο με τα εξής χαρακτηριστικά;

Home owner	Marital status	Annual income	Defaulted borrower
No	Single	82K	?

Πως θα υπολογίσουμε τις αντίστοιχες δεσμευμένες πιθανότητες

$$P(X_i = x_i | Y = y)$$

όταν η τ.μ. X_i λαμβάνει τιμές σε ένα συνεχές διάστημα;

Για παράδειγμα, πως θα εκτιμηθεί η πιθανότητα

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No})?$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Πως θα υπολογίσουμε τις αντίστοιχες δεσμευμένες πιθανότητες

$$P(X_i = x_i | Y = y)$$

όταν η τ.μ. X_i λαμβάνει τιμές σε ένα συνεχές διάστημα;

Υπάρχουν δύο προσεγγίσεις για τον χειρισμό **συνεχών** γνωρισμάτων:

- Διακριτοποίηση των τιμών
- Χρήση κάποιας συνεχούς κατανομής

Χειρισμός **συνεχών** γνωρισμάτων:
Στην περίπτωση της **διακριτοποίησης**

- διαμερίζουμε το σύνολο τιμών της τ.μ. σε διαστήματα
- και ο υπολογισμός της πιθανότητας γίνεται με βάση το ποσοστό των αντικειμένων που έχουν γνώρισμα με τιμή στο αντίστοιχο διάστημα.

Εδώ πρέπει να έχουμε υπόψη ότι

- πολλά διαστήματα θα έχουν ως συνέπεια λίγα αντικείμενα σε κάθε διάστημα
- λίγα διαστήματα θα έχουν ως συνέπεια πολλά αντικείμενα σε κάθε διάστημα τα οποία πιθανόν να ανήκουν σε διαφορετικές κατηγορίες

Χειρισμός **συνεχών** γνωρισμάτων:

- Υποθέτουμε κάποια συγκεκριμένη μορφή κατανομής πιθανοτήτων. Συνήθως κανονική (Gaussian) κατανομή.
- Μια κατανομή χαρακτηρίζεται από την συνάρτηση πυκνότητας πιθανότητας $f(x) = P(X = x)$ αυτής. Εδώ μας ενδιαφέρει η συνάρτηση πυκνότητας πιθανότητας (ς.π.π.)

$$f(x|y) = P(X = x|Y = y).$$

- Αν $Y = y$, για την κανονική κατανομή η ς.π.π. $f(x|y)$ έχει την μορφή

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma^2_{x|y}}} \exp(-\frac{(x - \mu_{x|y})^2}{2\sigma^2_{x|y}})$$

όπου $\mu_{x|y}$ είναι η αναμενόμενη τιμή της τ.μ. X δεδομένου ότι $Y = y$ και $\sigma^2_{x|y}$ είναι η διακύμανση της τ.μ. X δεδομένου ότι $Y = y$.

	Owner	Status	Income	Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No}) = ?$$

Οι τιμές της τ.μ. Annual Income όταν Defaulted Borrower = No έχουν μέσο όρο

$$\mu = \frac{125 + 100 + 70 + 120 + 60 + 220 + 75}{7} = 110.$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No}) = ?$$

$\mu = 110$ και διακύμανση σ^2.

$$\sigma^2 = \frac{(125-110)^2 + (100-110)^2 + (70-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2}{7}$$

$$= 2550$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Επομένως, μπορούμε να θεωρήσουμε ότι

$$P(\text{Annual Income} = x | \text{Defaulted Borrower} = \text{No})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi \cdot 2550}} \exp\left(-\frac{(x-110)^2}{2 \cdot 2550}\right)$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Οπότε

$$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{No})$$

$$= \frac{1}{\sqrt{2\pi \cdot 2550}} \exp\left(-\frac{(82-110)^2}{2 \cdot 2550}\right)$$

$$= 0.0067$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Αντίστοιχα, υπολογίζουμε την συνάρτηση πυκνότητας πιθανότητας

$$P(\text{Annual Income} = x | \text{Defaulted Borrower} = \text{Yes})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi \cdot 16.67}} \exp(-\frac{(x-90)^2}{2 \cdot 16.67})$$

όπου εδώ $\mu = 90$ και $\sigma^2 = \frac{50}{3} = 16.67$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Επομένως,

$$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{Yes})$$

$$= \frac{1}{\sqrt{2\pi \cdot 16.67}} \exp(-\frac{(82-90)^2}{2 \cdot 16.67}) = 0.014.$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Που θα κατηγοριοποίησουμε ένα άτομο με τα εξής χαρακτηριστικά;

Home owner	Marital status	Annual income	Defaulted borrower
No	Single	82K	?

$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{No}) = 0.0067$

$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{Yes}) = 0.014.$

Τα λεγόμενα **Bayesian belief networks** βασίζονται στην αναπαράσταση των σχέσεων εξάρτησης μεταξύ των γνωρισμάτων χρησιμοποιώντας ένα προσανατολισμένο γράφημα.

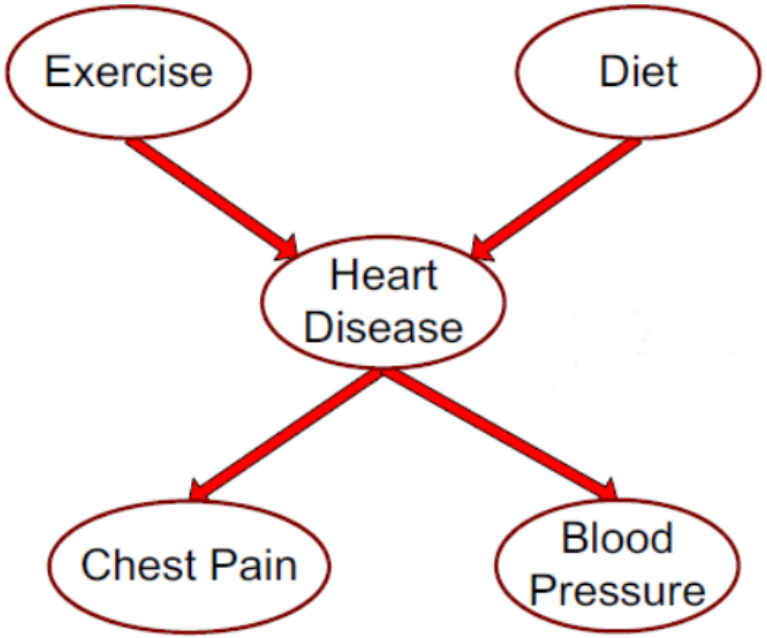
Παράδειγμα

Έστω ότι μας ενδιαφέρει να δούμε να ένα άτομο έχει κάποιο καρδιακό νόσημα (heart disease).
Έχουμε μια συλλογή δεδομένων εκπαίδευσης που αφορούν n άτομα και για κάθε άτομο έχουμε τα παρακάτω γνωρίσματα:

- Exercise (Yes, No)
- Diet (Healthy, Unhealthy)
- Chest pain (Yes, No)
- Blood pressure (Yes, No)
- Heart disease (Yes, No)

- Οι τιμές των γνωρίσματων Exercise και Diet επηρρεάζουν τις τιμές του γνωρίσματος Heart disease (π.χ. η ανθυγεινή διατροφή και η απουσία άσκησης μπορεί να δημιουργήσουν προβλήματα στην καρδιά). Υπάρχει μεταξύ τους μια **σχέση αιτίου – αποτελέσματος**.
- Από την άλλη, οι τιμές του γνωρίσματος Heart disease επηρρεάζουν τις τιμές των γνωρισμάτων Blood pressure και Heart disease (π.χ. προβλήματα στην καρδιά μπορεί να έχουν ως συνέπεια υψηλή πίεση και πόνο στο στήθος). Υπάρχει μεταξύ τους μια **σχέση αιτίου – αποτελέσματος**.

Μπορούμε να απεικονίσουμε τις προηγούμενες εξαρτήσεις με ένα προσανατολισμένο γράφημα



- Τα γνωρίσματα Exercise και Diet έχουν ανεξάρτητες τιμές.
- Το γνώρισμα Heart Disease εξαρτάται από τα Exercise και Diet
- Τα γνωρίσματα Chest Pain και Blood Pressure εξαρτώνται άμεσα από τις τιμές του Heart Disease (π.χ. είναι συμπτώματα). Επίσης, είναι ανεξάρτητα μεταξύ τους.

Έστω ότι θέλουμε να κατηγοροποιήσουμε ένα άτομο στην κλάση Yes ή No του γνωρίσματος Heart Disease όταν γνωρίζουμε ότι έχει τα παρακάτω χαρακτηριστικά:

Exercise	Diet	Chest pain	Blood Pressure
Yes	Unhealthy	No	Yes

Θεωρούμε τις τ.μ. E (Exercise), D (Diet), C (Chest Pain), B (Blood Pressure) και H (Heart Disease)

Επίσης, θεωρούμε τις συντομογραφίες $Y = \text{Yes}, \text{Healthy}$ και $N = \text{No}, \text{Unhealthy}$.

Όπως και στον Naive Bayes, για να αποφασίσουμε, θα συγκρίνουμε τις πιθανότητες

- $P(H = Y | E = Y, D = N, C = N, B = Y)$
- $P(H = N | E = Y, D = N, C = N, B = Y)$

Εδώ, λόγω των εξαρτήσεων δεν μπορούμε να χρησιμοποιήσουμε την παραδοχή της υπό συνθήκη ανεξαρτησίας. Γι᾽ αυτό θα ακολουθήσουμε την εξής μέθοδο για την σύγκριση:
Από τον ορισμό της δεσμευμένης πιθανότητας
$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$ έχουμε αντίστοιχα να συγκρίνουμε τις πιθανότητες

$$\frac{P(E = Y, D = N, C = N, B = Y, H = Y)}{P(E = Y, D = N, C = N, B = Y)}$$

και

$$\frac{P(E = Y, D = N, C = N, B = Y, H = N)}{P(E = Y, D = N, C = N, B = Y)}$$

Επειδή τα κλάσματα έχουν τους ίδιους παρονομαστές αρκεί να συγκρίνουμε τους αριθμητές.

Για να συγκρίνουμε τους αριθμητές

- $P(E = Y, D = N, C = N, B = Y, H = Y)$
- $P(E = Y, D = N, C = N, B = Y, H = N)$

μπορούμε να χρησιμοποιήσουμε τον τύπο

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

(ο οποίος προκύπτει με διαδοχική εφαρμογή του τύπου του ορισμού της δεσμευμένης πιθανότητας

$$P(X = x, Y = y) = P(Y = y)P(X = x | Y = y).)$$

Παρατηρήστε ότι στον τύπο

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

κάθε όρος του γινομένου είναι της μορφής

$$P(X_k = x_k | X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1})$$

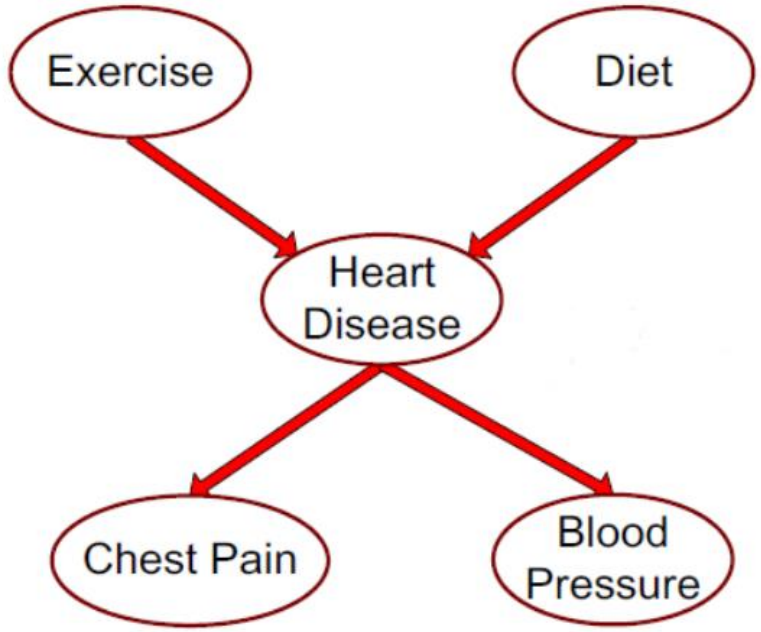
όπου $X_1, X_2, \ldots, X_{k-1}$ είναι οι τ.μ. των όρων του γινομένου που προηγούνται (από τα αριστερά προς τα δεξιά).
Επομένως, όταν εφαρμόζουμε τον τύπο ορίζουμε έμμεσα και κάποια σειρά στις τ.μ. $X_1, X_2, \ldots, X_n$.

Έδω έχουμε να τις πιθανότητες

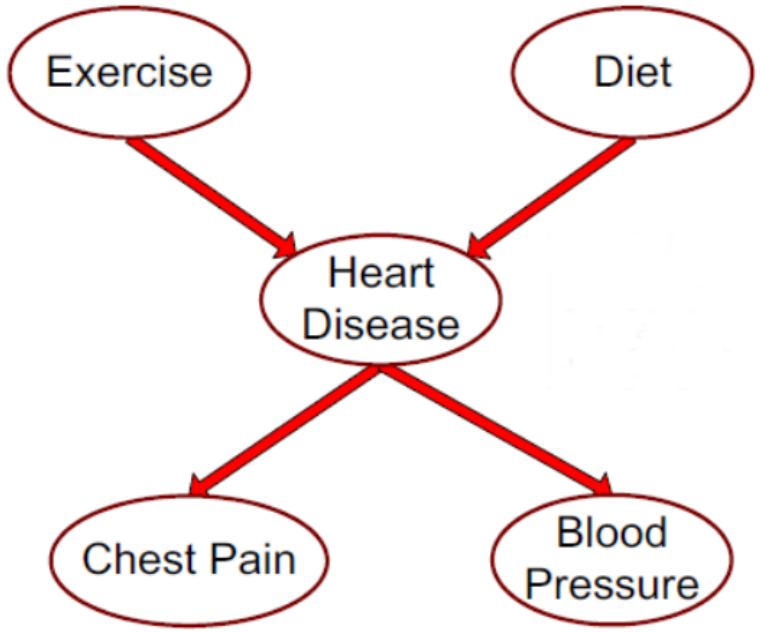
$$P(E = Y, D = N, C = N, B = Y, H = Y)$$

$$P(E = Y, D = N, C = N, B = Y, H = N)$$

Ποια σειρά μας συμφέρει να χρησιμοποιήσουμε για τις τ.μ. E, D, C, B, H;


Exercise
Diet
Heart Disease
Chest Pain
Blood Pressure

Θα βρούμε μια σειρά με την βοήθεια του γραφήματος εξαρτήσεων που κατασκευάσαμε:



Γνωρίζουμε ότι η τ.μ. H εξαρτάται από τις τ.μ. E, D, ενώ οι τ.μ. B και C εξαρτώνται από την τ.μ. H. Δεν υπάρχουν άλλες εξαρτήσεις.
Θα εφαρμόσουμε τον τύπο πρώτα για τις E, D που δεν έχουν άλλες εξαρτήσεις. Μετά για την H που εξαρτάται μόνο από τις E, D και μετά για τις B, C που εξαρτώνται μόνο από την H.

Κανόνας

Γενικά, με βάση το γράφημα των εξαρτήσεων, εφαρμόζουμε τον τύπο

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\
&= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\
&\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1})
\end{aligned}
$$

για μια τ.μ. Y όταν έχουμε ήδη εφαρμόσει τον τύπο για όλες τις τ.μ. $Y_1, Y_2, \ldots, Y_k$ από τις οποίες η Y έχει εξαρτήσεις.

Με βάση τη σειρά E, D, H, C, B έχουμε ότι

$$
\begin{aligned}
P(E = Y, D = N, C = N, B = Y, H = Y) = \\
& P(E = Y) \\
& \cdot P(D = N | E = Y) \\
& \cdot P(H = Y | E = Y, D = N) \\
& \cdot P(C = N | E = Y, D = N, H = Y) \\
& \cdot P(B = Y | E = Y, D = N, H = Y, C = N)
\end{aligned}
$$

Όμως,

- $P(D = N|E = Y) = P(D = N)$ (D, E ανεξάρτητες)
- $P(C = N|E = Y, D = N, H = Y) = P(C = N|H = Y)$ (C ανεξάρτητη από τις E, D, Εξαρτάται μόνο από την H)
- $P(B = Y|E = Y, D = N, H = Y, C = N) = P(B = Y|H = Y)$ (B ανεξάρτητη από τις E, D, C. Εξαρτάται μόνο από την H)

Επομένως, με βάση τις προηγούμενες απλοποίησεις

$$\begin{aligned} &P(E = Y, D = N, C = N, B = Y, H = Y) \\ &\quad = P(E = Y) \cdot P(D = N) \cdot P(H = Y | E = Y, D = N) \\ &\qquad \cdot P(C = N | H = Y) \cdot P(B = Y | H = Y) \end{aligned}$$

Εντελώς, ανάλογα

$$\begin{aligned} &P(E = Y, D = N, C = N, B = Y, H = N) \\ &\quad = P(E = Y) \cdot P(D = N) \cdot P(H = N | E = Y, D = N) \\ &\qquad \cdot P(C = N | H = N) \cdot P(B = Y | H = N) \end{aligned}$$

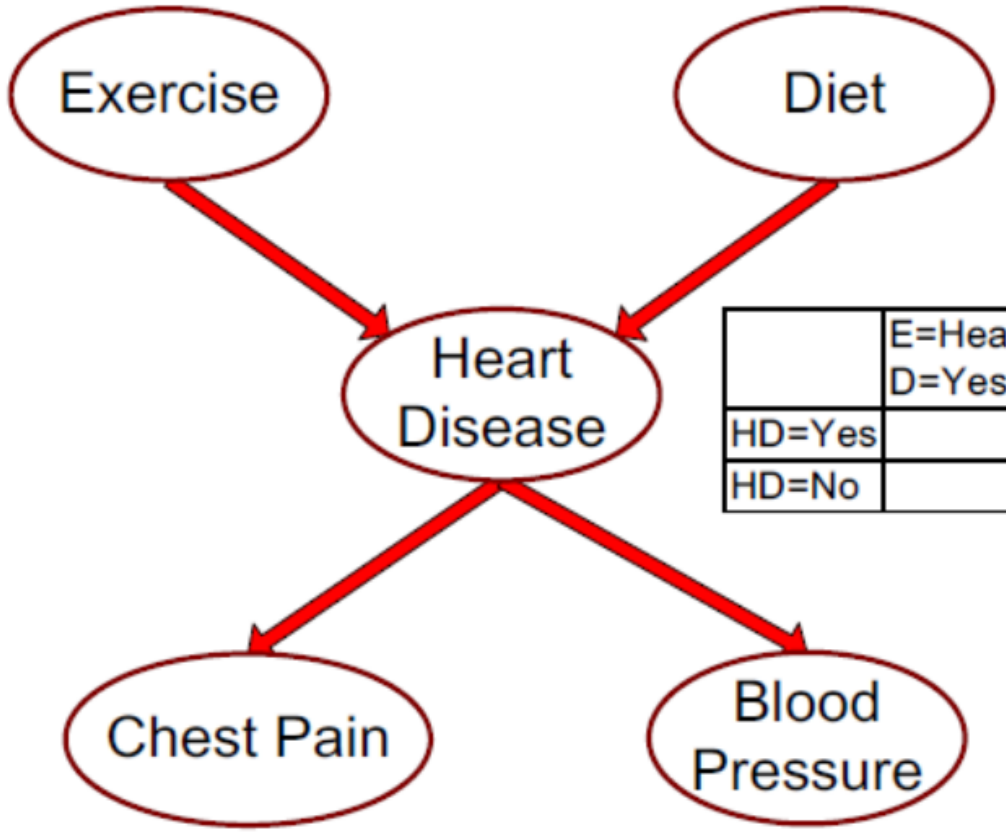
Τελικά, λόγω των κοινών παραγόντων $P(E = Y) \cdot P(D = N)$, αρκεί να συγκρίνουμε τα γινόμενα

- $P(H = Y | E = Y, D = N) \cdot P(C = N | H = Y) \cdot P(B = Y | H = Y)$
- $P(H = N | E = Y, D = N) \cdot P(C = N | H = N) \cdot P(B = Y | H = N)$.

Έστω ότι από τα δεδομένα υπολογίσαμε τις παρακάτω πιθανότητες:

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	E=Healthy D=Yes	E=Healthy D=No	E=Unhealthy D=Yes	E=Unhealthy D=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Τότε, έχουμε ότι

- $P(H = Y|E = Y, D = N) \cdot P(C = N|H = Y) \cdot P(B = Y|H = Y) = 0.55 \cdot 0.2 \cdot 0.85 = 0.0935$
- $P(H = N|E = Y, D = N) \cdot P(C = N|H = N) \cdot P(B = Y|H = N) = 0.44 \cdot 0.01 \cdot 0.2 = 0.00088.$

Άρα, το άτομο με τα παρακάτω χαρακτηριστικά

Exercise	Diet	Chest pain	Blood Pressure
Yes	Unhealty	No	Yes

θα το κατηγοριοποίησουμε με Yes στο γνώρισμα Heart disease.