Μπαϋεσιανά δίκτυα

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

ΕΙΣΑΓΩΓΗ

- αναπαράσταση σύνθετων σχέσεων μεταξύ μεταβλητών
- εξαγωγή συμπερασμάτων σε συνθήκες αβεβαιότητας

- γραφικά πιθανοτικά μοντέλα, τα οποία αναπαριστούν σχέσεις με μορφή γράφων

- **Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές.**

ΕΙΣΑΓΩΓΗ

- Βάση το θεώρημα του Bayes
- Υπολογίζει την υπό συνθήκη πιθανότητα P(H|X), δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση H με δεδομένο ότι ισχύει το γεγονός X.

$$P(H|X) = \frac{P(H) * P(X|H)}{P(X)}$$

όπου P(H) είναι η εκ των προτέρων πιθανότητα να ισχύει η υπόθεση H, P(X) είναι η εκ των προτέρων πιθανότητα να συμβεί το γεγονός X και P(X|H) είναι η πιθανότητα να συμβεί το γεγονός X με δεδομένο ότι ισχύει η υπόθεση H

ΕΙΣΑΓΩΓΗ

- Υποθέτουμε ότι Χ είναι μια παρατήρηση του συνόλου δεδομένων και Η είναι η υπόθεση ότι παρατήρηση ανήκει στην κλάση C_i . Το Χ θεωρείται ότι είναι διάνυσμα τιμών $X=(x_1 , x_2 ,…, x_n)$. Υποθέτουμε ότι υπάρχουν m κλάσεις C_1 , C_2 ,…, C_m .

Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση Χ στην κλάση C_i υπολογίζεται από την σχέση

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

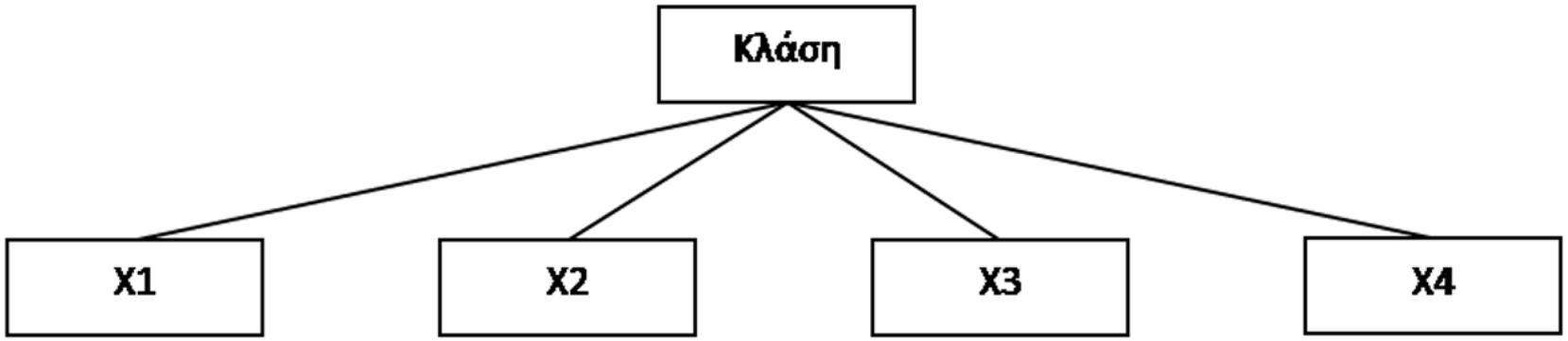
ΕΙΣΑΓΩΓΗ

- Η μέθοδο υπολογίζει τις πιθανότητες για κάθε κλάση και τοποθετεί την παρατήρηση στην κλάση με την μεγαλύτερη πιθανότητα. Εφόσον το P(X) είναι ίδιο με όλες τις κλάσεις και το P(C) μπορεί να υπολογισθεί (πλήθος παρατηρήσεων που ανήκουν στην κλάση C_i προς το πλήθος όλων των παρατηρήσεων) τότε ο υπολογισμός της υπο-συνθηκη πιθανότητας P(X| C_i) δίνεται από την σχέση

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

- όπου x_k είναι η τιμή της διάστασης κ του διανύσματος Χ

ΕΙΣΑΓΩΓΗ



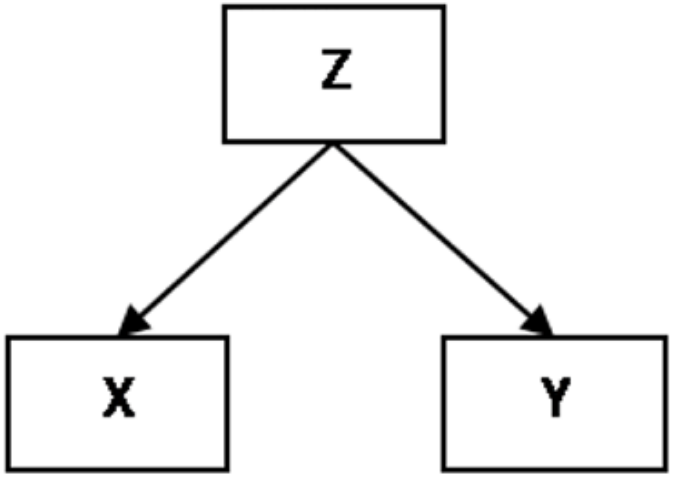
- Η μέθοδο υπολογίζει τις πιθανότητες για κάθε κλάση και τοποθετεί την παρατήρηση στην κλάση με την μεγαλύτερη πιθανότητα.

Δίκτυα

- Ένα Μπαυέσιανο Δίκτυο αναπαριστά τις εξαρτήσεις μεταξύ των μεταβλητών με την χρήση ενός κατευθυνόμενου ακυκλικού γράφου. Κάθε κόμβος του γράφου συμβολίζει μια μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης. Ένα βέλος το οποίο κατευθύνεται από την μεταβλητή Χ προς την μεταβλητή Υ τότε η Υ εξαρτάται από την Χ. Η μεταβλητή Χ καλείται γονέας της Υ και η Υ καλείται τέκνο της Χ
- Οι μεταβλητές Χ και Υ είναι υπό συνθήκη ανεξάρτητες, εάν οι τιμές της Χ, με δεδομένες τις τιμές των Υ και Z, εξαρτώνται μόνο από τις τιμές της Z.

$$P(X|Z,Y) = P(X|Z)$$

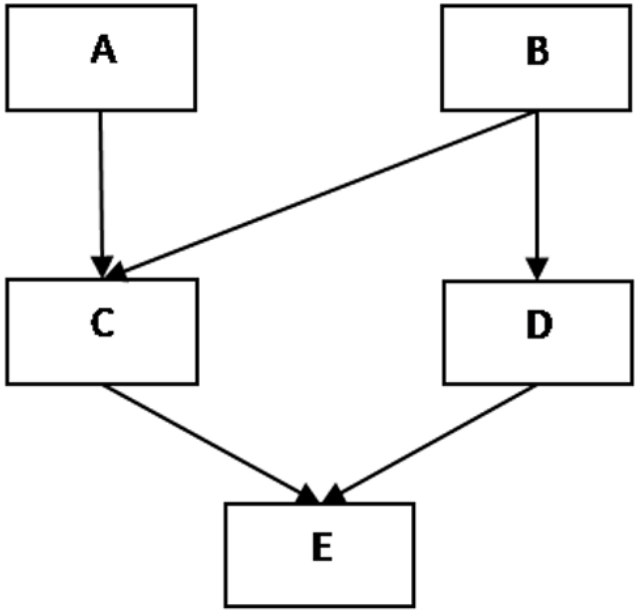
Δίκτυα



$$P(X|Z,Y) = P(X|Z)$$

Δίκτυα

- Στα Μπαϋεσιανά Δίκτυα ισχύει η τοπική ιδιότητα Markov, σύμφωνα με την οποία κάθε μεταβλητή είναι υπό συνθήκη ανεξάρτητη από τους μη απογόνους της όταν είναι δεδομένοι οι γονείς της.



Η μεταβλητή C είναι ανεξάρτητη από την D εάν είναι γνωστές οι μεταβλητές A και B. Αυτό σημαίνει ότι εάν οι τιμές των μεταβλητών A και B είναι γνωστές, τότε η μεταβλητή D δεν προφέρει πρόσθετη πληροφορία σχετικά με τη μεταβλητή C. Οι μεταβλητές μπορούν να παίρνουν τιμές διακριτές ή συνεχόμενες.

Δίκτυα

Ο γράφος των Μπαϋεσιανών Δικτύων καταγράφει τις σχέσεις μεταξύ των μεταβλητών. Οι σχέσεις αυτές ποσοτικοποιούνται με τον Πίνακα Υπό Συνθήκη Πιθανοτήτων

Στον πίνακα καταγράφεται για κάθε μεταβλητή Χ η κατανομή πιθανοτήτων P(X|Par(X)), όπου Par(X) οι γονείς της μεταβλητής Χ. Αν τα δεδομένα περιέχουν n-μεταβλητές $X_1,X_2,\ldots,X_n$, τότε η πιθανότητα εμφάνισης μιας παρατήρησης με τιμές $x_1,x_2,\ldots,x_n$ για τις αντίστοιχες μεταβλητές

$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P\left(x_i|Par(X_i)\right)$$

Δίκτυα

- Ένας από τους κόμβους αντιπροσωπεύει τη μεταβλητή της κλάσης. Για μια παρατήρηση υπολογίζονται οι πιθανότητες για κάθε δυνατή τιμή της κλάσης, και η παρατήρηση εκχωρείται στην πιο πιθανή κλάση
- Η δημιουργία ενός μοντέλου Μπαϋεσιανού Δικτύου περιλαμβάνει δύο εργασίες

1. την κατασκευή του γράφου,
2. τον υπολογισμό του πίνακα πιθανοτήτων CPT.

Πλεονεκτήματα

- Δημιουργούν ένα μοντέλο για την κατανομή πιθανοτήτων για ένα πρόβλημα. Υπό αυτήν την έννοια είναι ιδιαίτερα κατάλληλα για περιπτώσεις όπου υπάρχουν σύνθετες εξαρτήσεις μεταξύ της μεταβλητής της κλάσης και των μεταβλητών εισόδου ή και ακόμα μεταξύ των μεταβλητών εισόδου.
- Ο γράφος που δημιουργείται οπτικοποιεί τις σχέσεις μεταξύ της κλάσης και των μεταβλητών εισόδου
- Μπορούν να χειριστούν και αριθμητικές και ονομαστικές μεταβλητές

Μειονεκτήματα

- Δεν υπάρχει ένας καθιερωμένος και γενικά αποδεκτός τρόπος εξαγωγής του γράφου από τα δεδομένα.
- Για τον υπολογισμό των πιθανοτήτων ενός κλάδου του δικτύου απαιτείται ο υπολογισμός όλων των άλλων κλάδων επιφέροντας σημαντικό υπολογιστικό κόστος

Παραδείγματα

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

προβλέψετε τις πιθανότητες έγκρισης και απόρριψης της αίτησης ενός υποψηφίου με μέσο εισόδημα και μικρή ηλικία

Παραδείγματα

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

Ο πίνακας περιέχει 14 περιπτώσεις και στις τρεις από αυτές οι υποψήφιοι έχουν μέσο εισόδημα και μικρή ηλικία.

$$P(X) = 3/14 = 0,21$$

Για τις οκτώ περιπτώσεις του συνόλου το δάνειο εγκρίνεται, ενώ για τις τέσσερεις απορρίπτεται.

$$P(Yes) = 8/14 = 0,57$$
$$P(No) = 6/14 = 0,429$$

Από τις οκτώ περιπτώσεις όπου το δάνειο εγκρίνεται, στις δύο το εισόδημα είναι μέσο και η ηλικία μικρή.
Από τις έξι περιπτώσεις όπου το δάνειο δεν εγκρίνεται, στη μια περίπτωση το εισόδημα είναι μέσο και η μικρή.

$$P(X|Yes) = 2/8 = 0,25$$
$$P(X|No) = 1/6 = 0,167$$

Παραδείγματα

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

οι πιθανότητες έγκρισης και απόρριψης του δανείου για μέσο εισόδημα και μικρή ηλικία είναι:

$$P(\text{Yes}|X) = (0{,}25*0{,}57)/0{,}21 = 0{,}667$$
$$P(\text{No}|X) = (0{,}167*0{,}429)/0{,}21 = 0{,}333$$

Η πιθανότητα έγκρισης του δανείου είναι διπλάσια από την πιθανότητα απόρριψης του

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

Παραδείγματα

Tid	Refund	Marital Status	Taxable Income	Evade
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

Κλάση: $P(C) = N_c/N$

- Π.χ., P(No) = 7/10,
  P(Yes) = 3/10

Για διακριτές ιδιότητες:

$$P(A_i \mid C_k) = |A_{ik}| / N_c \quad k$$

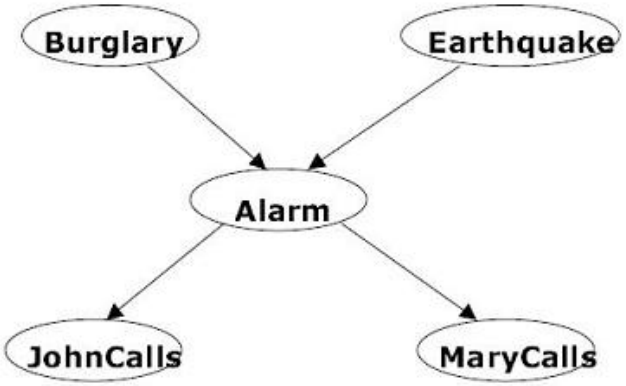
- όπου $|A_{ik}|$ είναι ο αριθμός των παραδειγμάτων που έχουν την ιδιότητα A_i και ανήκει στην κλάση C_k
- Παράδειγμα:

  P(Status=Married|No) = 4/7
  P(Refund=Yes|Yes)=0

Παραδείγματα

- Ενα σπίτι που βρίσκεται σε σεισμογενή περιοχή έχει σύστημα συναγερμού. Ο συναγερμός μπορεί να ενεργοποιηθεί εξαιτίας ενός σεισμού. Υπάρχουν δυο γείτονες, η Mary και ο John, που δε γνωρίζονται μεταξύ τους. Αν ακούσουν το συναγερμό θα τηλεφωνήσουν στον ιδιοκτήτη του σπιτιού, αλλά αυτό δεν είναι σίγουρο.
- Θέλουμε να αναπαραστήσουμε την πιθανοτική κατανομή των γεγονότων:

Ληστεία (Burglary), Σεισμός (Earthquake), Συναγερμός (Alarm), τηλεφωνεί η Mary (MaryCalls) και τηλεφωνεί ο John (JohnCalls).



Παραδείγματα

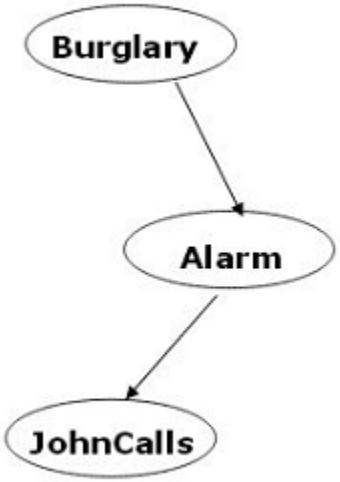
- Τοπικές δεσμευμένες πιθανοτικές κατανομές.
- Για κάθε σχήμα μεταβλητή - γονέας, $P(X_i | pa(X_i))$, όπου το $pa(X_i)$ συμβολίζει το γονέα του X_i

B	E	$P(A\|B,E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Η πλήρης από κοινού κατανομή ορίζεται σύμφωνα με τις τοπικές δεσμευμένες κατανομές (μέσω του κανόνα της αλυσίδας): $P(X_1, X_2, \ldots, X_n) = \prod_{i=1,\ldots,n} P(X_i | pa(X_i))$

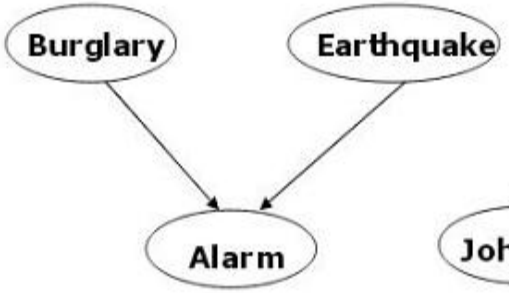
Παραδείγματα

1.



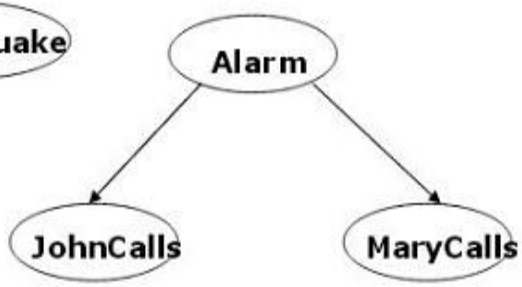
Το ότι ο John τηλεφωνεί (JohnCalls) είναι ανεξάρτητο απ΄ τη ληστεία (Burglar) δοθέντος του συναγερμού (Alarm).

2.



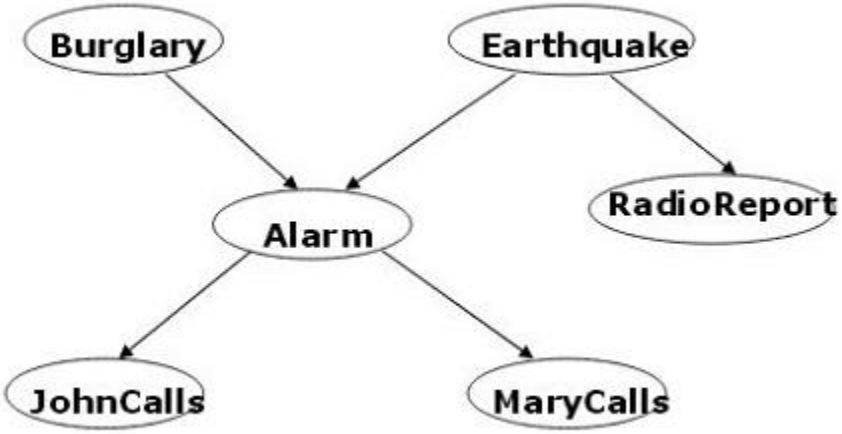
Η ληστεία είναι ανεξάρτητη απ΄ το σεισμό (Earthquake) (χωρίς να ξέρουμε κάτι για το συναγερμό).
Η ληστεία και ο σεισμός δεν είναι ανεξάρτητα δοθέντος του συναγερμού.

3.



Το ότι τηλεφωνεί η Mary (MaryCalls) είναι ανεξάρτητο απ΄ το ότι ο John τηλεφωνεί δοθέντος του συναγερμού.

Παραδείγματα



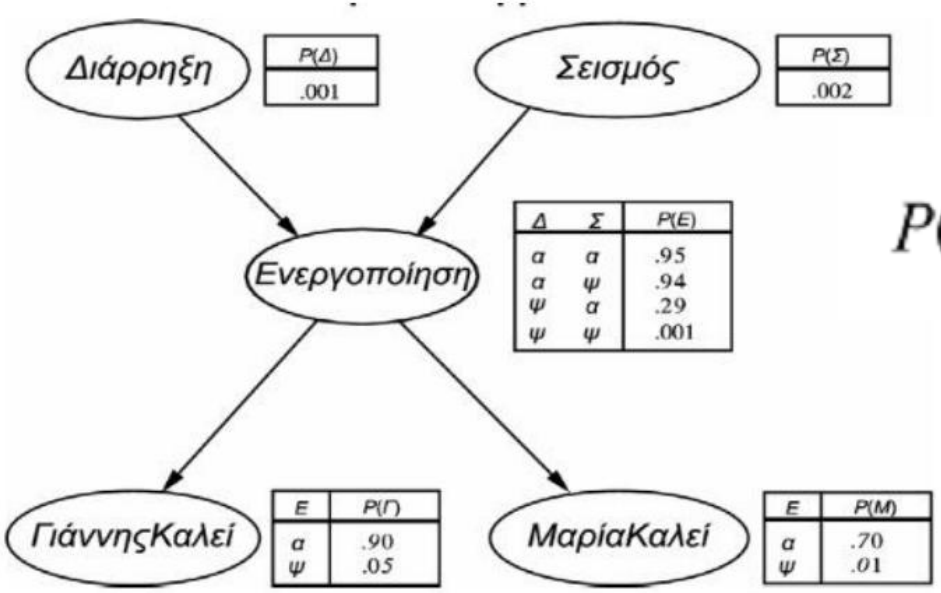
Ο σεισμός και η ληστεία δεν είναι ανεξάρτητα δοθέντος του τηλεφωνήματος της Mary.

Η ληστεία και το τηλεφώνημα της Mary δεν είναι ανεξάρτητα όταν δεν ξέρουμε κάτι για το συναγερμό.

Η ληστεία και η ανακοίνωση στο ραδιόφωνο RadioReport είναι ανεξάρτητα δοθέντος του σεισμού.

Η ληστεία και η ανακοίνωση στο ραδιόφωνο δεν είναι ανεξάρτητα δοθέντος του τηλεφωνήματος της Mary.

Παραδείγματα



$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{γονείς}(x_i))$$

Κανόνας αλυσίδας:

$$P(x_1, x_2, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \cdot P(x_{n-1} \mid x_{n-2}, \ldots, x_1)$$

$$\cdot \ldots \cdot P(x_2 \mid x_1) \cdot P(x_1) =$$

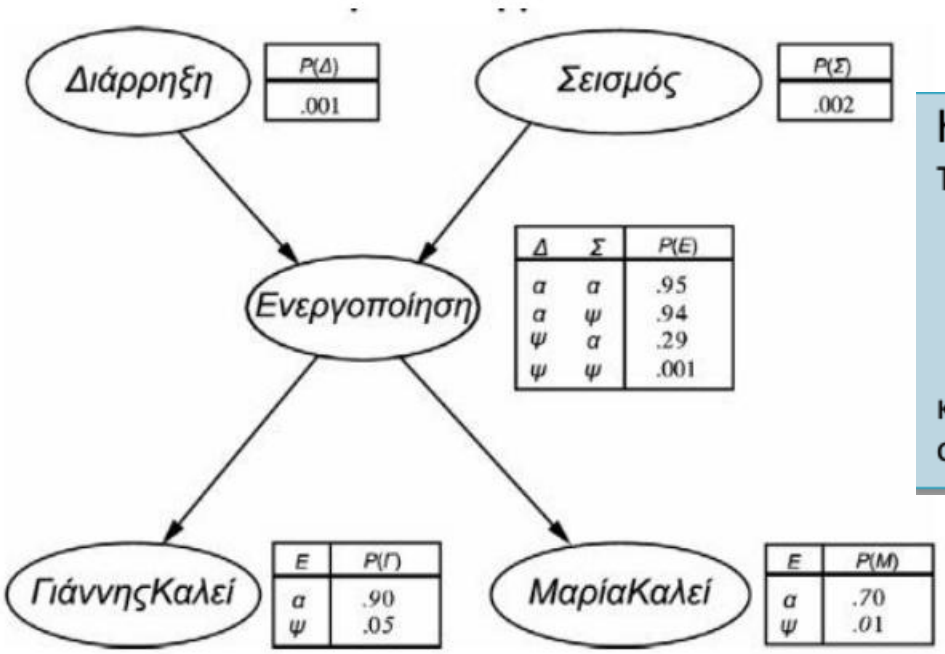
$$\prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1)$$

Εάν *Γονείς*(X_i) $\subseteq \{X_{i-1}, \ldots, X_1\}$ τότε:

- $\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid \textit{Γονείς}(X_i))$

- **P**(*ΜαρίαΚαλεί* | *ΓιάννηςΚαλεί*, *Ενεργοποίηση*, *Σεισμός*, *Διάρρηξη*) = **P**(*ΜαρίαΚαλεί* | *Ενεργοποίηση*)

Παραδείγματα



Συμπερασμός

Η πιθανότητα του ερωτήματος ***X*** δοθέντος του συμβάντος ***e*** είναι:

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X,\mathbf{e}) = \alpha \cdot \sum_{y} \mathbf{P}(X,\mathbf{e},\mathbf{y})$$

όπου *α* παράγοντας κανονικοποίησης της πιθανότητας [0-1], Y οι μη συσχετιζόμενες μεταβλητές

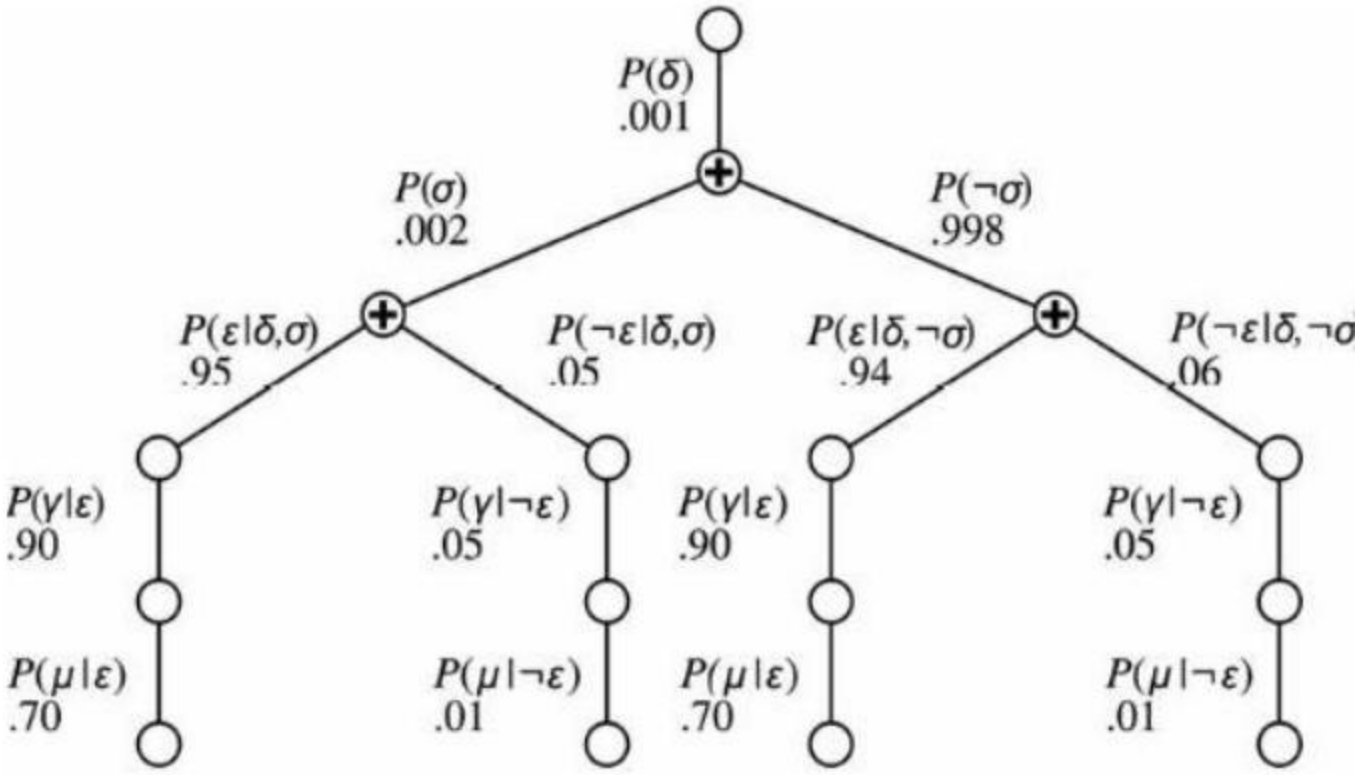
- Για *Διάρρηξη = αληθές* έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot \sum_{\sigma}\sum_{\varepsilon} P(\delta)P(\sigma)P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$

Παραδείγματα

Για *Διάρρηξη* = *αληθές* έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot \sum_{\sigma} \sum_{\varepsilon} P(\delta)P(\sigma)P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$



Υπολογισμός πιθανότητας από δεδομένα

Εκτίμηση της πυκνότητα πιθανότητας:

- Υποθέτουμε ότι οι ιδιότητες ακολουθούν την κανονική κατανομή
- Χρησιμοποιούμε τα δεδομένα για να εκτιμήσουμε τις παραμέτρους της κατανομής (π.χ., μέση τιμή και διασπορά)
- όταν γίνει γνωστή η κατανομή, μπορεί να χρησιμοποιηθεί για να υπολογισθεί η πιθανότητα $P(A_i|c)$

Κανονική κατανομή

Tid	Refund	Marital Status	Taxable Income	Evade
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

Υπολογισμός πιθανότητας από δεδομένα

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

Για (Income=120K, Class=No):

- αν Class=No
  - Μέση τιμή = 110
  - Διασπορά= 2975

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

Tid	Refund	Marital Status	Taxable Income	Evade
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

$$P(Income=120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Υπολογισμός πιθανότητας από δεδομένα

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
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(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

Υπολογισμός πιθανότητας από δεδομένα

Tid	Refund	Marital Status	Taxable Income	Evade
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(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Αφού P(X|No)P(No) > P(X|Yes)P(Yes)

Τότε P(No|X) > P(Yes|X)

=> Class = No

Υπολογισμός πιθανότητας από δεδομένα

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: ιδιότητες θηλαστικά

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Παραδείνματα

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

Παραδείγματα

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$P(x \mid c) = P(Sunny \mid No) = 2/5 = 0.4$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

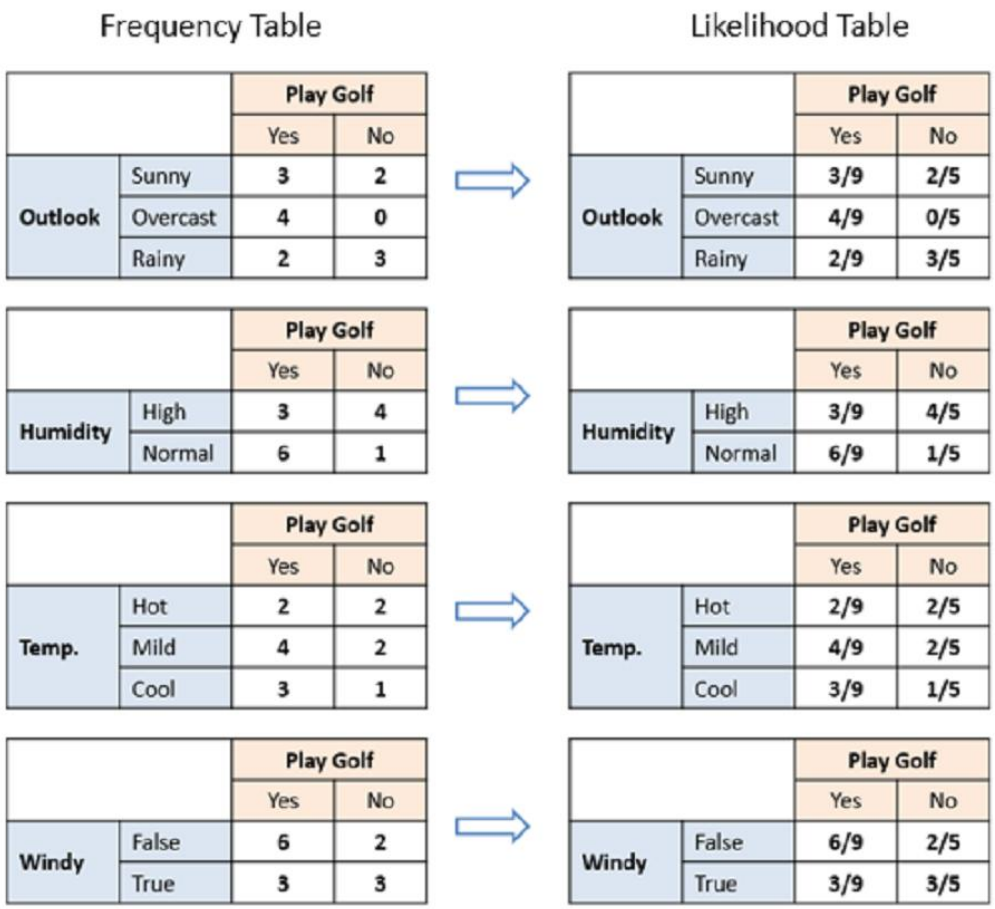
		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
		9	5	14

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(No) = 5/14 = 0.36$

Posterior Probability: $P(c \mid x) = P(No \mid Sunny) = 0.40 \times 0.36 \div 0.36 = 0.40$

Παραδείγματα



Παραδείγματα

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Μια νέα ημέρα:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Πιθανοφάνειες για τις δύο κλάσεις

"yes" : $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

"no" : $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Πιθανότητες (μετά την κανονικοποίηση):

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Παραδείγματα

- Η πιθανότητα να συμβεί ένα γεγονός *H* δοθείσης μιας μαρτυρίας *E*:

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]}$$

  - *A priori* πιθανότητα του *H*: $\Pr[H]$
    - Η πιθανότητα του γεγονότος *χωρίς* την επίκληση της μαρτυρίας
  - *A posteriori* πιθανότητα του *H*: $\Pr[H \mid E]$
    - Η πιθανότητα του γεγονότος *με* την επίκληση της μαρτυρίας

- Εκμάθηση κατηγοριοποίησης: ποια η πιθανότητα μιας κλάσης δοθείσης μιας μαρτυρίας;
  - Η μαρτυρία *E* είναι η εγγραφή στη ΒΔ
  - Το γεγονός *H* είναι η κλάση της εγγραφής
- Απλοϊκή (naïve) παραδοχή: η μαρτυρία διαιρείται σε μέρη (όσο και τα γνωρίσματα) που είναι ανεξάρτητα μεταξύ τους

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Παραδείγματα

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

μαρτυρία E

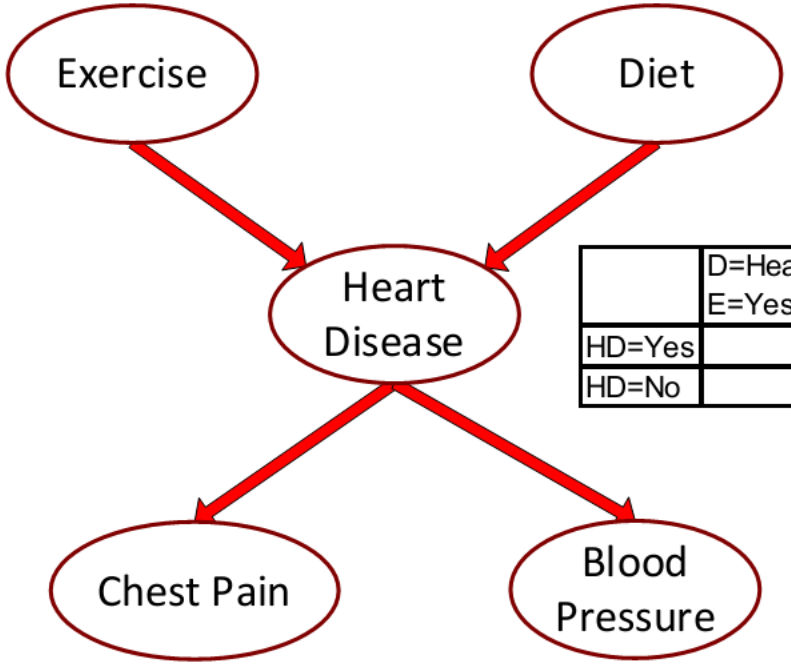
$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes]$$
$$\times \Pr[Temperature = Cool \mid yes]$$
$$\times \Pr[Humidity = High \mid yes]$$
$$\times \Pr[Windy = True \mid yes]$$
$$\times \frac{\Pr[yes]}{\Pr[E]} = \frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{0/14}$$

Πιθανότητα να συμβεί ένα γεγονός (δηλ. Play = "yes") δοθείσης της μαρτυρίας E (δηλ. Outlook = "Sunny" and Temp. = "Cool" and Humidity = "High" and Windy = "True")

Πίνακες Πιθανοτήτων

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	D=Healthy E=Yes	D=Healthy E=No	D=Unhealthy E=Yes	D=Unhealthy E=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

Πίνακες Πιθανοτήτων

X = (E=No, D=Yes, CP=Yes, BP=High)

P(HD|E,D,CP,BP)?

P(HD=Yes| E=No,D=Yes) = 0.55
P(CP=Yes| HD=Yes) = 0.8
P(BP=High| HD=Yes) = 0.85

- P(HD=Yes|E=No,D=Yes,CP=Yes,BP=High) $\propto 0.55 \times 0.8 \times 0.85 =$ 0.374

P(HD=No| E=No,D=Yes) = 0.45
P(CP=Yes| HD=No) = 0.01
P(BP=High| HD=No) = 0.2

- P(HD=No|E=No,D=Yes,CP=Yes,BP=High) $\propto 0.45 \times 0.01 \times 0.2 =$ 0.0009

Κλασικοποιηση Χ ως ναι