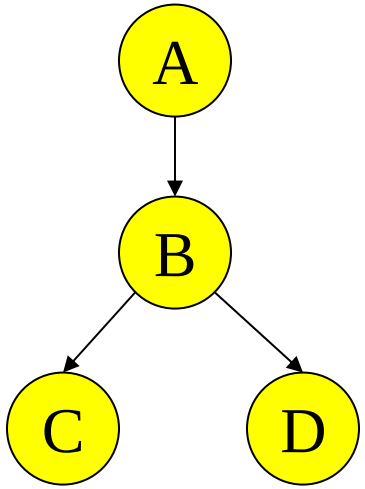
Bayesian Belief Networks (BBNs)

Εισαγωγή

- Ένα Bayesian δίκτυο αποτελείται από:
  - A Directed Acyclic Graph



  - Ένα σύνολο πινάκων για κάθε κόμβο στο γράφο

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

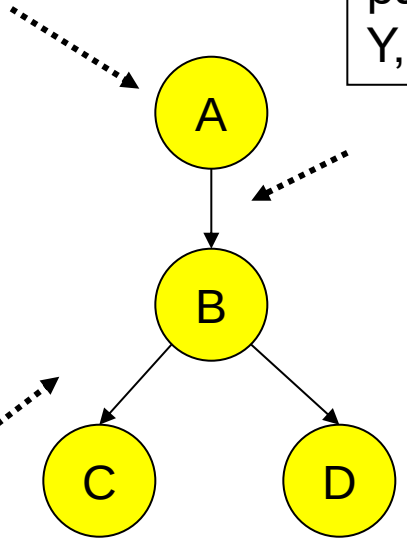
B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Εισαγωγή

Κάθε κόμβος στο γράφο είναι μια τυχαία μεταβλητή

Ένας κόμβος Χ είναι γονέας ενός άλλου κόμβου Υ εάν υπάρχει ένα βέλος από τον κόμβο Χ στον κόμβο Υ, π.χ. Ο Α είναι γονέας του Β



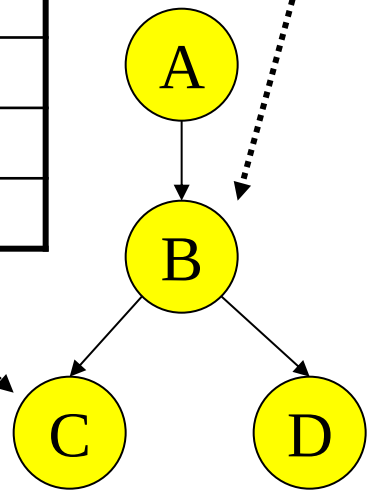
Ένα βέλος από τον κόμβο Χ στον κόμβο Υ σημαίνει ότι το Χ έχει άμεση επιρροή στο Υ

Εισαγωγή

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

- Κάθε κόμβος X_i έχει μια υπό όρους κατανομή πιθανότητας $P(X_i \mid Parents(X_i))$ που ποσοτικοποιεί την επίδραση των γονέων στον κόμβο
- Οι παράμετροι είναι οι πιθανότητες σε αυτούς τους πίνακες πιθανοτήτων υπό όρους (conditional probability tables - CPT)

Πιθανότητες

Conditional Probability Distribution for C given B

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

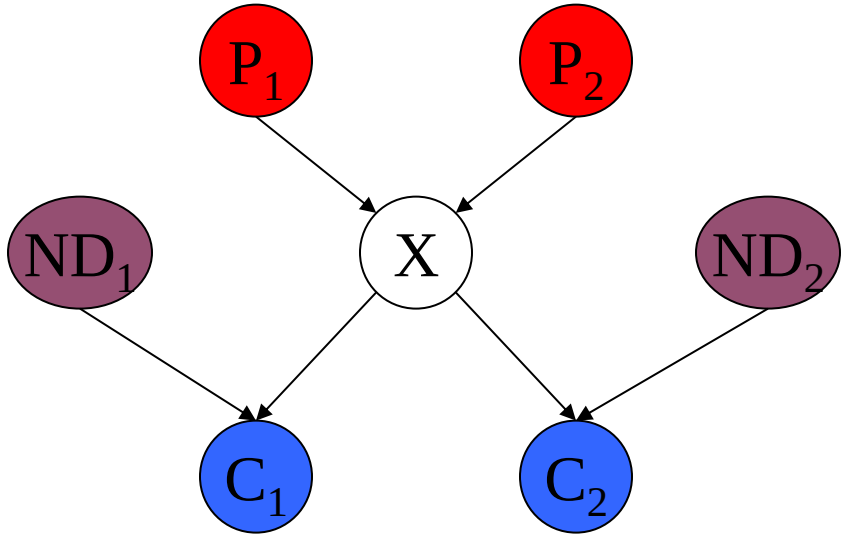
Για έναν δεδομένο συνδυασμό τιμών των γονέων (B σε αυτό το παράδειγμα), οι εγγραφές για P(C=true | B) και P(C=false | B) πρέπει να αθροίζονται σε 1π.χ. P(C=true | B=false) + P(C=false |B=false)=1

Εάν έχετε μια μεταβλητή Boolean με k Boolean γονείς, αυτός ο πίνακας έχει 2^{k+1} πιθανότητες (αλλά μόνο 2^k πρέπει να αποθηκευτούν)

Στοιχεία

- Δύο σημαντικές ιδιότητες:
  - Κωδικοποιεί τις σχέσεις ανεξαρτησίας υπό όρους μεταξύ των μεταβλητών στη δομή του γράφου
  - Είναι μια συμπαγής αναπαράσταση της κοινής κατανομής πιθανοτήτων στις μεταβλητές

- Η συνθήκη Markov: δεδομένων των γονιών της (P_1, P_2),ένας κόμβος (X) είναι υπό όρους ανεξάρτητος από τους μη απογόνους του (ND_1, ND_2)



Στοιχεία

- Λόγω της συνθήκης Markov, μπορούμε να υπολογίσουμε την κοινή κατανομή πιθανότητας σε όλες τις μεταβλητές X_1, …, X_n στο Bayesian δίκτυο χρησιμοποιώντας το μαθηματικό τύπο:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid Parents(X_i))$$

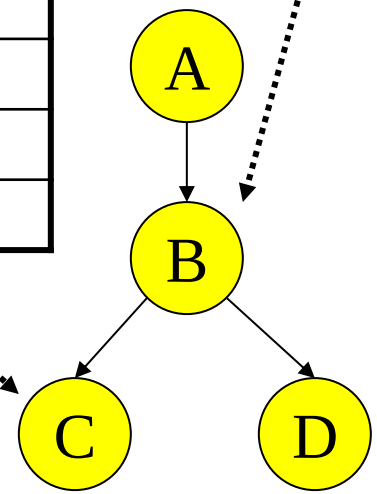
- Όπου Parents(X_i) σημαίνει τις τιμές των Parents του κόμβου X_i σε σχέση με το γράφο

Παράδειγμα

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

P(A = true, B = true, C = true, D = true) =

P(A = true) * P(B = true | A = true) * P(C = true | B = true) P(D = true | B = true) =

(0.4)*(0.3)*(0.1)*(0.95)

Παράδειγμα

P(A = true, B = true, C = true, D = true)

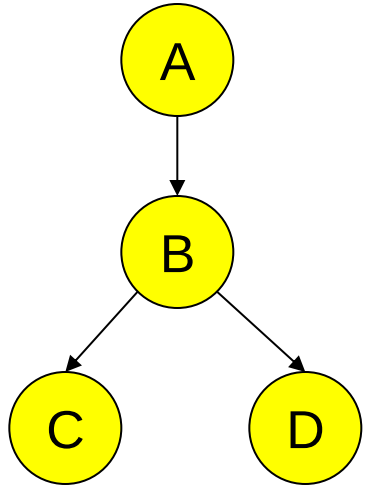
= P(A = true) * P(B = true | A = true) *

P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)

This is from the graph structure

These numbers are from the conditional probability tables



Συμπερασμός

- Τα BBN υποστηρίζουν τρία κύρια είδη συλλογιστικής:
  - Πρόβλεψη συνθηκών με δεδομένες προδιαθέσεις
  - Διάγνωση καταστάσεων με τα συμπτώματα (και προδιαθεσικές)
  - Εξήγηση μιας κατάστασης με μία ή περισσότερες προδιαθέσεις
- Στο οποίο μπορούμε να προσθέσουμε ένα τέταρτο:
  - Η απόφαση για μια ενέργεια με βάση τις πιθανότητες των συνθηκών

Συμπερασμός

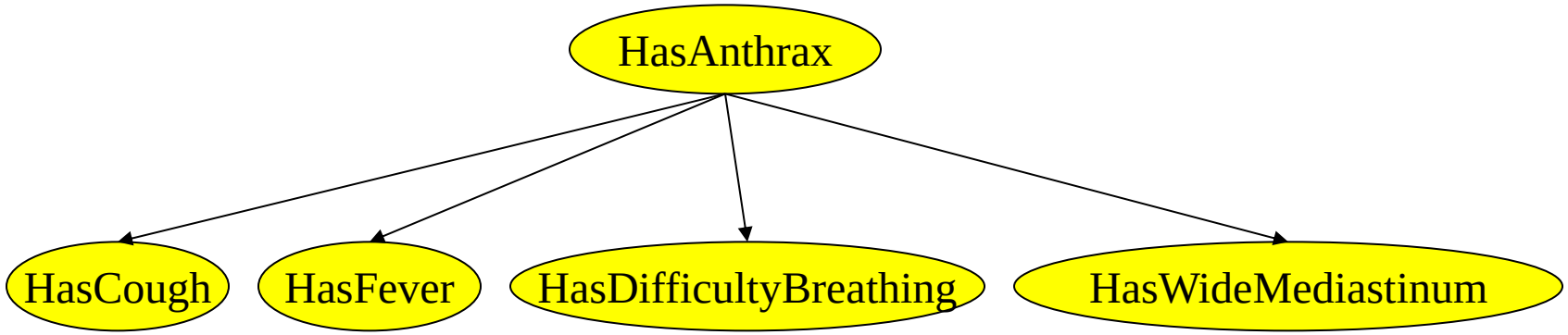
- Η χρήση ενός Bayesian δικτύου για τον υπολογισμό των πιθανοτήτων ονομάζεται συμπέρασμα
- Γενικά, το συμπέρασμα περιλαμβάνει ερωτήματα της μορφής:

$$P(X \mid E)$$

E = The evidence variable(s)

X = The query variable(s)

Συμπερασμός



- Ένα παράδειγμα ερωτήματος θα ήταν:

$$P(\text{HasAnthrax} = \text{αληθές} \mid \text{HasFever} = \text{αληθές}, \text{HasCough} = \text{αληθές})$$

- Σημείωση: Παρόλο που το HasDifficultyBreathing και το HasWideMediastinum βρίσκονται στο δίκτυο Bayes, δεν τους δίνονται τιμές στο ερώτημα (δηλ. δεν εμφανίζονται ούτε ως μεταβλητές ερωτήματος ούτε ως μεταβλητές απόδειξης)
- Αντιμετωπίζονται ως μη παρατηρούμενες μεταβλητές και συνοψίζονται.

Λήψη Αποφάσεων

- Η πρόγνωση του καιρού για σήμερα μπορεί να είναι είτε ηλιόλουστος, συννεφιασμένος ή βροχερός

  Πρέπει να πάρετε μια ομπρέλα όταν φεύγετε;

- Η απόφασή σας εξαρτάται μόνο από την πρόβλεψη
- Η πρόγνωση «εξαρτάται» από τον πραγματικό καιρό
- Η ικανοποίησή σας εξαρτάται από την απόφασή σας και τον καιρό
- Αντιστοιχίστε ένα βοηθητικό πρόγραμμα σε καθεμία από τις τέσσερις καταστάσεις: (βροχή, χωρίς βροχή) x (ομπρέλα, χωρίς ομπρέλα)

Λήψη Αποφάσεων

- Επεκτείνετε το πλαίσιο BBN ώστε να περιλαμβάνει δύο νέα είδη κόμβων:

  Απόφαση και Όφελος

- Ένας κόμβος απόφασης υπολογίζει την αναμενόμενη χρησιμότητα μιας απόφασης δεδομένης των γονέων της, π.χ., πρόβλεψη, μια αποτίμηση
- Ένας κόμβος υπολογίζει μια τιμή χρησιμότητας δεδομένων των γονέων του, π.χ. απόφαση και καιρός
- Μπορούμε να αντιστοιχίσουμε ένα όφελος σε καθεμία από τις τέσσερις καταστάσεις:

  (βροχή, χωρίς βροχή) x (ομπρέλα, χωρίς ομπρέλα)

- Η τιμή που αποδίδεται σε καθένα είναι πιθανώς υποκειμενική

Εκπαίδευση

- **Σενάριο 1**: Δεδομένης τόσο της δομής του δικτύου όσο και όλων των παρατηρήσιμων μεταβλητών: υπολογίστε μόνο τις εγγραφές CPT
- **Σενάριο 2**: Γνωστή δομή δικτύου, ορισμένες μεταβλητές κρυφές: μέθοδος gradient descent (greedy hill-climbing), δηλ. αναζήτηση λύσης κατά μήκος της πιο απότομης κατάβασης μιας συνάρτησης κριτηρίου
  - Τα βάρη αρχικοποιούνται σε τυχαίες τιμές πιθανότητας
  - Σε κάθε επανάληψη, κινείται προς αυτό που φαίνεται να είναι η καλύτερη λύση αυτή τη στιγμή, π.χ. οπισθοδρόμηση
  - Τα βάρη ενημερώνονται σε κάθε επανάληψη και συγκλίνουν στο τοπικό βέλτιστο
- **Σενάριο 3**: Άγνωστη δομή δικτύου, παρατηρήσιμες όλες οι μεταβλητές: αναζήτηση στο χώρο του μοντέλου για ανακατασκευή της τοπολογίας δικτύου
- **Σενάριο 4**: Άγνωστη δομή, όλες οι κρυφές μεταβλητές: Δεν υπάρχουν καλοί αλγόριθμοι γνωστοί για αυτόν τον σκοπό

Δημιουργία

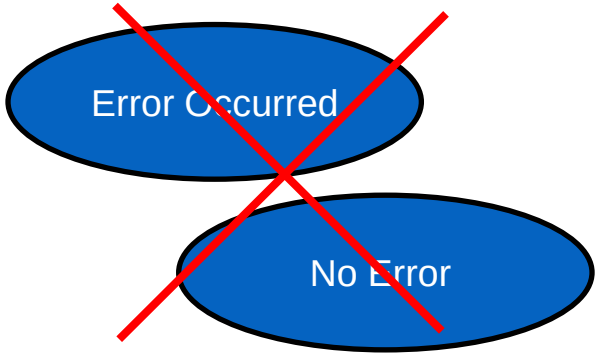
- Η διαδικασία απόκτησης γνώσης για ένα BBN περιλαμβάνει τρία βήματα
  - Επιλογή κατάλληλων μεταβλητών
  - Λήψη απόφασης για τη δομή του δικτύου
  - Λήψη δεδομένων για τους πίνακες πιθανοτήτων υπό όρους

Δημιουργία

- Οι μεταβλητές πρέπει να είναι συλλογικά εξαντλητικές, αμοιβαία αποκλειστικές τιμές

$$x_1 \vee x_2 \vee x_3 \vee x_4$$

$$\neg(x_i \wedge x_j) \quad i \neq j$$



- Πρέπει να είναι τιμές, όχι πιθανότητες

Δημιουργία

- Example of good variables
  - Weather {Sunny, Cloudy, Rain, Snow}
  - Gasoline: Cents per gallon
  - Temperature { >=100F , < 100F}
  - User needs help on Excel Charting {Yes, No}
  - User’s personality {dominant, submissive}

Association Rules

Εισαγωγή

- Είναι ένα σημαντικό μοντέλο εξόρυξης δεδομένων που έχει μελετηθεί εκτενώς από τις κοινότητες των βάσεων δεδομένων και της εξόρυξης δεδομένων.
- Ας υποθέσουμε ότι όλα τα δεδομένα είναι κατηγορηματικά.
- Δεν υπάρχει καλός αλγόριθμος για αριθμητικά δεδομένα.
- Χρησιμοποιήθηκε αρχικά για την Ανάλυση Καλαθιού Αγοράς για να βρείτε πώς σχετίζονται τα στοιχεία που αγοράζουν οι πελάτες.

Bread → Milk [sup = 5%, conf = 100%]

Εισαγωγή

- **Συχνό μοτίβο/πρότυπο (frequent pattern)**: ένα μοτίβο (ένα σύνολο στοιχείων, υποακολουθίες, πρότυπα, κ.λπ.) που εμφανίζεται συχνά σε ένα σύνολο δεδομένων
- Προτάθηκε για πρώτη φορά από τους Agrawal, Imielinski και Swami στο πλαίσιο των συχνών συνόλων στοιχείων και της εξόρυξης κανόνων συσχέτισης
- **Κίνητρο**: Εύρεση εγγενών κανονικοτήτων στα δεδομένα
  - Ποια προϊόντα αγοράζονταν συχνά μαζί;— Μπύρα και πάνες;
  - Ποιες είναι οι επόμενες αγορές μετά την αγορά ενός υπολογιστή;
  - Ποια είδη DNA είναι ευαίσθητα σε αυτό το νέο φάρμακο;
  - Μπορούμε να ταξινομήσουμε αυτόματα έγγραφα web;
- **Εφαρμογές**
  - Ανάλυση δεδομένων καλαθιού, διασταυρούμενο μάρκετινγκ, σχεδιασμός καταλόγου, ανάλυση εκστρατειών εκπτώσεων, ανάλυση καταγραφής Ιστού (ροή κλικ) και ανάλυση αλληλουχίας DNA.

Εισαγωγή

- Συχνό μοτίβο/πρότυπο: Μια εγγενής και σημαντική ιδιότητα των συνόλων δεδομένων
- Θεμέλιο για πολλές βασικές εργασίες εξόρυξης δεδομένων
  - Ανάλυση συσχέτισης, συσχέτισης και αιτιότητας
  - Διαδοχικά, δομικά (π.χ. υπογραφικά) μοτίβα
  - Ανάλυση προτύπων σε χωροχρονικά, πολυμέσα, χρονοσειρές και δεδομένα ροής
  - Ταξινόμηση: διακριτική, συχνή ανάλυση προτύπων
  - Ανάλυση συστάδων: συχνή ομαδοποίηση βάσει προτύπων
  - Αποθήκευση δεδομένων: κύβος ‘παγόβουνου’ και κλίση κύβων
  - Συμπίεση σημασιολογικών δεδομένων
  - Ευρείες εφαρμογές

Δεδομένα

$I = \{i_1, i_2, \ldots, i_m\}$: a set of *items*.

Transaction t :
t a set of items, and $t \subseteq I$.

Transaction Database T: a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$.

Παράδειγμα

- Συναλλαγές καλαθιού αγοράς:
  - t_1: {bread, cheese, milk}
  - t_2: {apple, eggs, salt, yogurt}
  - … …
  - t_n: {biscuit, eggs, milk}
- Έννοιες:
  - Αντικείμενο: ένα αντικείμενο σε ένα καλάθι
  - I: το σύνολο όλων των αντικειμένων που πωλούνται στο κατάστημα
  - Μια συναλλαγή: είδη που αγοράζονται σε ένα καλάθι - μπορεί να έχει TID (αναγνωριστικό συναλλαγής)
  - Ένα σύνολο δεδομένων συναλλαγών: Ένα σύνολο συναλλαγών

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Παράδειγμα

A text document data set. Each document is treated as a “bag” of keywords

doc1: Student, Teach, School
doc2: Student, School
doc3: Teach, School, City, Game
doc4: Baseball, Basketball
doc5: Basketball, Player, Spectator
doc6: Baseball, Coach, Game, Team
doc7: Basketball, Team, City, Game

Κανόνες

- Μια συναλλαγή t περιέχει X, ένα σύνολο στοιχείων (itemset) στο I, αν $X \subseteq t$.
- Ένας κανόνας συσχέτισης είναι μια συνέπεια της μορφής:

$$X \rightarrow Y, \text{ όπου } X, Y \subset I, \text{ και } X \cap Y = \varnothing$$

- Ένα itemset είναι ένα σύνολο στοιχείων.
  - Π.χ., το X = {γάλα, ψωμί, δημητριακά} είναι ένα σύνολο στοιχείων.
- Ένα k-itemset είναι ένα σύνολο με k στοιχεία.
  - Π.χ., {γάλα, ψωμί, δημητριακά} είναι ένα σύνολο 3 ειδών

Κανόνες

Support: The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactions contain $X \cup Y$.

$sup = \Pr(X \cup Y)$.

Confidence: The rule holds in *T* with confidence *conf* if *conf*% of transactions that contain *X* also contain *Y.*

$conf = \Pr(Y \mid X)$

An association rule is a pattern that states when *X* occurs, *Y* occurs with certain probability.

Κανόνες

Support count: The support count of an itemset *X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain *X*.

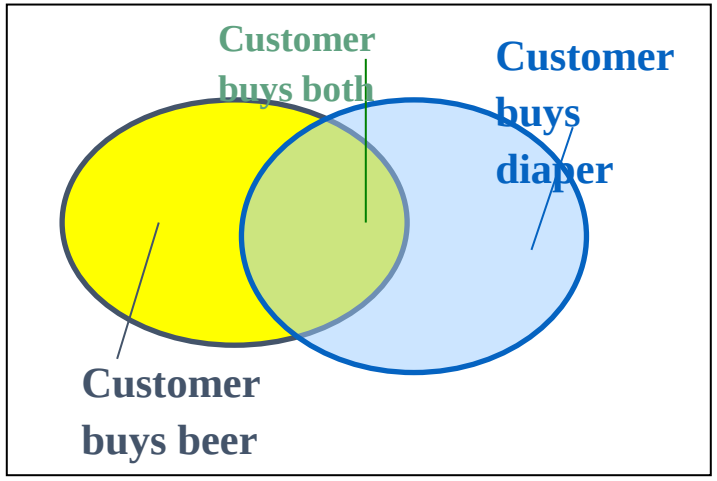
Assume *T* has *n* transactions.
Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Κανόνες

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



Βρες όλους τους κανόνες *X → Y με το ελάχιστο* support και confidence

support, *s*, η πιθανότητα (probability) μια συναλλαγή να περιέχει το X ∪ Y

confidence, *c,* η υπό συνθήκη πιθανότητα (conditional probability) μια συναλλαγή που έχει το X θα έχει και το *Y*

Έστω minsup = 50%, minconf = 50%

Frequent Patterns: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
  - *Beer → Diaper* (60%, 100%)
  - *Diaper → Beer* (60%, 75%)

Στόχος

- **Στόχος**: Βρείτε όλους τους κανόνες που ικανοποιούν το ελάχιστο support που καθορίζεται από τον χρήστη (**minsup**) και το ελάχιστο confidence (**minconf**).
- Βασικά Χαρακτηριστικά
  - Πληρότητα: βρείτε όλους τους κανόνες.
  - Δεν υπάρχει αντικείμενο(α) στη δεξιά πλευρά
  - Εξόρυξη με δεδομένα σε σκληρό δίσκο (όχι στη μνήμη)

Πρότυπα

- Ένα μεγάλο πρότυπο περιέχει έναν συνδυαστικό αριθμό υπο-προτύπων, π.χ., για το {a1, …, a100} περιέχει $\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 = 1.27*10^{30}$ υπο-πρότυπα!
- Λύση: Αντ' αυτού, χρησιμοποιούμε κλειστά πρότυπα και μέγιστα πρότυπα
- Ένα σύνολο στοιχείων X είναι κλειστό πρότυπο εάν το X είναι συχνό και δεν υπάρχει υπερ-πρότυπο Y ↄX, με το ίδιο support με το X
- Ένα σύνολο στοιχείων X είναι ένα μέγιστο πρότυπο εάν το X είναι συχνό και δεν υπάρχει συχνό υπερ- πρότυπο Y ↄX
- Το κλειστό πρότυπο είναι μια συμπίεση χωρίς απώλειες της συχνότητας των συχνών προτύπων
  - Μείωση του αριθμού των προτύπων και κανόνων

Πρότυπα

- DB = {<a1, …, a100>, < a1, …, a50>}

  Min_sup = 1.

- What is the set of closed itemset?

  <a1, …, a100>: 1

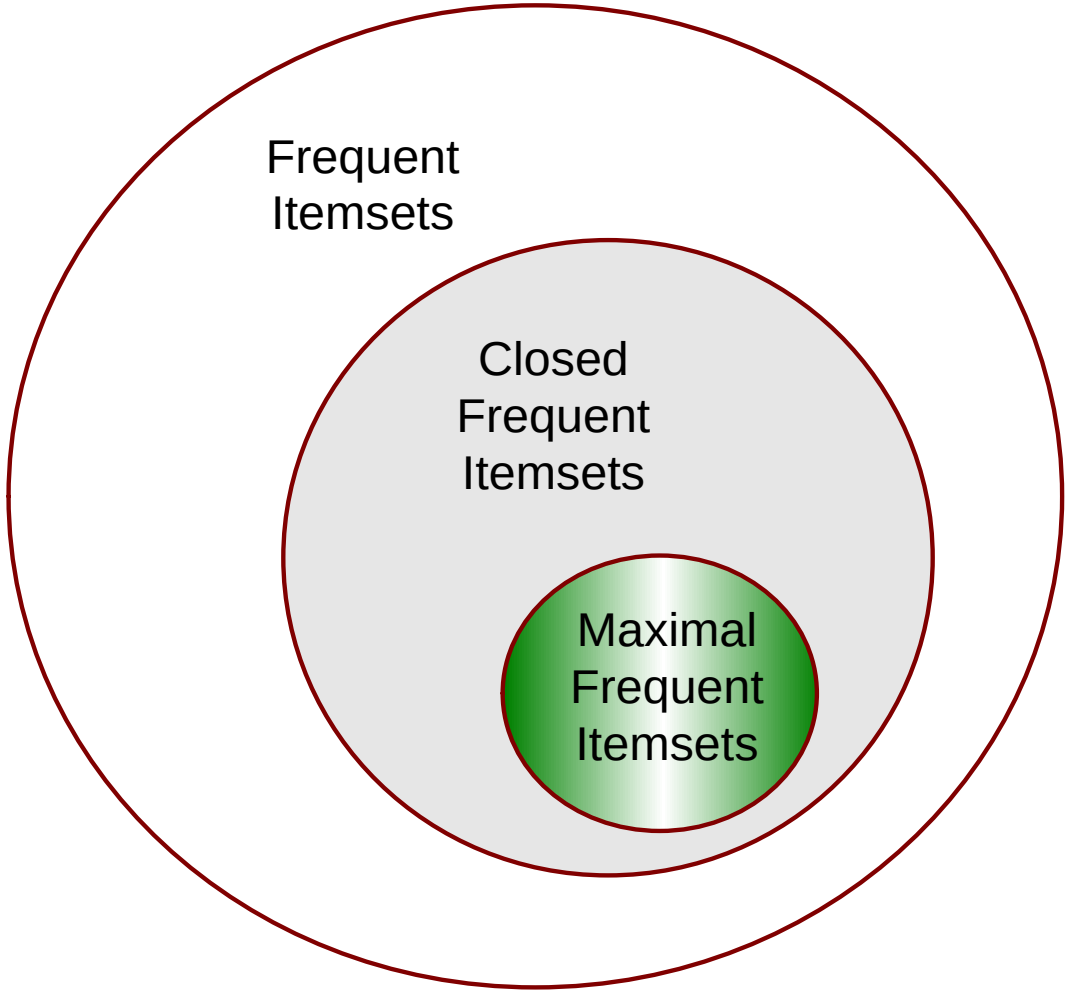
  < a1, …, a50>: 2

- What is the set of max-pattern?

  <a1, …, a100>: 1

- What is the set of all patterns?
- !!

Πρότυπα


Frequent
Itemsets
Closed
Frequent
Itemsets
Maximal
Frequent
Itemsets

Apriori Algorithm

- Η προς τα κάτω ιδιότητα κλεισίματος των συχνών προτύπων
- Οποιοδήποτε υποσύνολο ενός συνόλου συχνών στοιχείων πρέπει να είναι συχνό
- Εάν το {beer, diaper, nuts} είναι συχνό, το ίδιο ισχύει και για το {beer, diaper}, δηλαδή, κάθε συναλλαγή με {beer, diaper, nuts} περιέχει επίσης {beer, diaper}

Apriori Algorithm

- Βασική ιδέα: Η ιδιότητα apriori (ιδιότητα κλεισίματος προς τα κάτω): οποιαδήποτε υποσύνολα ενός συνόλου συχνών στοιχείων είναι επίσης συχνά σύνολα στοιχείων
- Αρχή κλαδέματος (pruning) Apriori: Εάν υπάρχει κάποιο σύνολο στοιχείων που είναι σπάνιο, το υπερσύνολο του δεν πρέπει να δημιουργηθεί/δοκιμαστεί!
- Ένα συχνό σύνολο στοιχείων είναι ένα σύνολο στοιχείων του οποίου το support είναι ≥ minsup.
- Μέθοδος:
  - Αρχικά, σαρώστε τη DB μία φορά για να έχετε συχνό σύνολο 1 στοιχείου
  - Δημιουργήστε υποψήφια σύνολα στοιχείων μήκους (k+1) από συχνά σύνολα στοιχείων μήκους k
  - Δοκιμάστε τα υποψηφία σύνολα έναντι του DB
  - Τερματισμός όταν δεν μπορεί να δημιουργηθεί συχνό ή υποψήφιο σύνολο

Apriori Algorithm

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) **do begin**
    C_{k+1} = candidates generated from L_k;
    for each transaction t in database do
        increment the count of all candidates in C_{k+1} that are contained in t
    L_{k+1} = candidates in C_{k+1} with min_support
    end
return $\cup_k L_k$;

Apriori Algorithm

Function candidate-gen(F_{k-1})

$C_k \leftarrow \emptyset$;

forall $f_1, f_2 \in F_{k-1}$

with $f_1 = \{i_1, \ldots, i_{k-2}, i_{k-1}\}$

and $f_2 = \{i_1, \ldots, i_{k-2}, i'_{k-1}\}$

and $i_{k-1} < i'_{k-1}$ **do**

$c \leftarrow \{i_1, \ldots, i_{k-1}, i'_{k-1}\}$; // join f_1 and f_2

$C_k \leftarrow C_k \cup \{c\}$;

for each (k-1)-subset s of c **do**

if ($s \notin F_{k-1}$) **then**

delete c from C_k; // prune

end

end

return C_k;

Apriori Algorithm

Level-wise approach

C_k = candidate itemsets of size k
L_k = frequent itemsets of size k

1. k = 1, C_1 = all items
2. While C_k not empty

Frequent itemset generation

3. Scan the database to find which itemsets in C_k are frequent and put them into L_k

Candidate generation

4. Use L_k to generate a collection of candidate itemsets C_{k+1} of size k+1

5. k = k+1

Apriori Algorithm

- Πως παράγουμε τα υποψήφια σύνολα;
  - Step 1: self-joining L_k
  - Step 2: pruning
- Παράδειγμα
  - L_3={*abc, abd, acd, ace, bcd*}
  - Self-joining: L_3*L_3
    - *abcd* from *abc* and *abd*
    - *acde* from *acd* and *ace*
  - Pruning:
    - *acde* is removed because *ade* is not in L_3
  - C_4 = {*abcd*}

Apriori Algorithm

F_3 = {{1, 2, 3}, {1, 2, 4}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}}

Μετά το join

C_4 = {{1, 2, 3, 4}, {1, 3, 4, 5}}

Μετά το pruning:

C_4 = {{1, 2, 3, 4}}

because {1, 4, 5} is not in F_3 ({1, 3, 4, 5} is removed)

Apriori Algorithm

$Sup_{min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Apriori Algorithm

- Frequent itemsets ≠ association rules
- Απαιτείται ένα ακόμα βήμα για να εξάγουμε τους κανόνες

- For each frequent itemset *X*,
- For each proper nonempty subset *A* of *X*,
  - Let *B* = X - *A*
  - A $\rightarrow$ B is an association rule if
    - Confidence(A $\rightarrow$ B) ≥ minconf,
    - support(A $\rightarrow$ B) = support(A$\cup$B) = support(X)
    - **confidence(A $\rightarrow$ B) = support(A $\cup$ B) / support(A)**

Apriori Algorithm

- Υποθέτουμε ότι το {2,3,4} είναι συχνό με sup=50%
  - Κατάλληλα μη κενά σύνολα: {2,3}, {2,4}, {3,4}, {2}, {3}, {4}, με sup=50%, 50%, 75%, 75%, 75%, 75% respectively
  - Δημιουργούνται τα ακόλουθα association rules:
    - 2,3 → 4, confidence=100%
    - 2,4 → 3, confidence=100%
    - 3,4 → 2, confidence=67%
    - 2 → 3,4, confidence=67%
    - 3 → 2,4, confidence=67%
    - 4 → 2,3, confidence=67%
    - Όλοι οι κανόνες έχουν support = 50%

Apriori Algorithm

- **Κανόνας Συσχέτισης (Association Rule)**
  - Μια έκφραση της μορφής X → Y, όπου X και Y είναι σύνολα (itemsets)
    - {Milk, Diaper} → {Beer}
- **Μετρικές Αποτίμησης (Rule Evaluation Metrics)**
  - Support (s)
    - Τμήμα των συναλλαγών που περιέχουν τα X & Y = η πιθανότητα P(X,Y) όπου X & Y συναντώνται μαζί
  - Confidence (c)
    - Πόσο συχνά το Y εμφανίζεται σε συναλλαγές που περιέχουν το X = η υπό συνθήκη πιθανότητα P(Y|X) το Y να συναντάται δεδομένου ότι συναντούμε το X.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

- **Problem Definition**
  - Input A set of transactions T, over a set of items I, minsup, minconf values
  - Output: All rules with items in I having s ≥ minsup and c≥ minconf

Apriori Algorithm

- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support ≥ minsup
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a partitioning of a frequent itemset into Left-Hand-Side (LHS) and Right-Hand-Side (RHS)

- Frequent itemset: {A,B,C,D}
- Rule: $AB \rightarrow CD$

Apriori Algorithm

What is wrong with confidence?

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Number of people that drink tea

Number of people that drink coffee and tea

Number of people that drink coffee but not tea

Number of people that drink coffee

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = $\frac{15}{20} = 0.75$

but P(Coffee) = $\frac{90}{100} = 0.9$

- Although confidence is high, rule is misleading
- P(Coffee|~~Tea~~) = 0.9375

Παράδειγμα

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3