Αποθήκες Δεδομένων και Εξόρυξη Γνώσης

Naïve Bayes – Δίκτυα Bayes

Βασικές Έννοιες: Γεγονότα, Ανεξαρτησία και Πιθανότητες

- Γεγονότα (Events)
  - π.χ. «ένα email είναι spam»
  - αναπαρίσταται σαν μια μεταβλητή (π.χ. A, B)
  - εμφανίζεται με μια τυχαία πιθανότητα, π.χ. P(A)
- Συνεμφάνιση των A και B (joint distribution)
  - Η πιθανότητα τα A και B να εμφανίζονται ταυτόχρονα
  - P(A^B) ή P(A,B)
- Ανεξαρτησία των A και B (conditional independence)
  - Η εμφάνιση ενός γεγονότος δεν επηρεάζει την εμφάνιση του άλλου
  - Τότε: P(A,B)=P(A)*P(B)

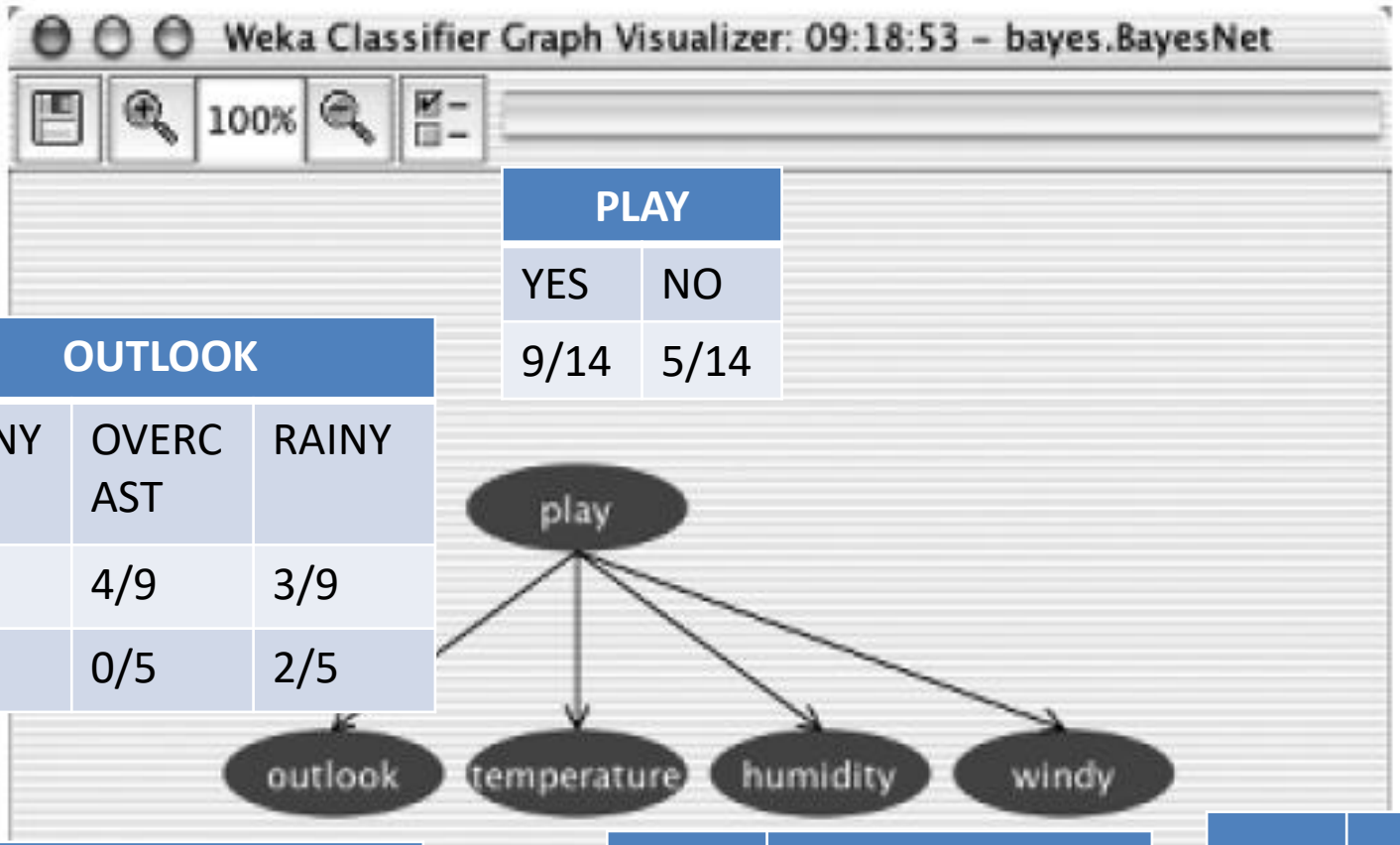
Το θεώρημα του Bayes

- Η πιθανότητα του γεγονότος Α, δεδομένου του Β: P(A|B)
- P(A): η εκ των προτέρων (a priori) πιθανότητα του Α
- Αν τα Α και Β είναι ανεξάρτητα: P(A|B)=P(A)
- Γενικά: P(A|B)=P(A,B)/P(B)
- Θεώρημα του Bayes: P(A|B)=P(B|A)*P(A)/P(B)

Table 1.2 The weather data.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Naïve Bayes (όλα τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους)



PLAY	
YES	NO
9/14	5/14

	OUTLOOK		
PLAY	SUNNY	OVERCAST	RAINY
YES	2/9	4/9	3/9
NO	3/5	0/5	2/5

	TEMPERATURE		
PLAY	HOT	MILD	COOL
YES	2/9	4/9	3/9
NO	2/5	2/5	1/5

	HUMIDITY	
PLAY	HIGH	NORMAL
YES	3/9	6/9
NO	4/5	1/5

	WINDY	
PLAY	YES	NO
YES	3/9	6/9
NO	3/5	2/5

Ταξινόμηση με Naïve Bayes

- P(C|att1, att2, att3, …attn)=P(att1|C)* P(att2|C) *…*P(attn|C)*P(C)/P(att1, att2, att3, …attn)
- Μια μέρα με βροχή, κρύο, υψηλή υγρασία και αέρα θα παίξει τένις ο παίκτης, ή όχι;
- P(PLAY|RAINY, COOL, HIGH, TRUE) = P(RAINY|PLAY)*P(COOL|PLAY)*P(HIGH|PLAY)* P(TRUE|PLAY)*P(PLAY)/P(RAINY, COOL, HIGH, TRUE) = (3/9 * 3/9 * 3/9 * 3/9 * 9/14) / P(RAINY, COOL, HIGH, TRUE)
- P(NOPLAY|RAINY, COOL, HIGH, TRUE) = P(RAINY|NOPLAY)*P(COOL|NOPLAY)*P(HIGH|NOPLAY)* P(TRUE|NOPLAY)*P(NOPLAY) /P(RAINY, COOL, HIGH, TRUE) = (2/5 * 1/5 * 4/5 * 3/5 * 5/14) / P(RAINY, COOL, HIGH, TRUE)

Naïve Bayes με αριθμητικά χαρακτηριστικά

Weather	Temperature	Humidity	Wind	Play
Rainy	71	91	Yes	No
Sunny	69	70	No	Yes
Sunny	80	90	Yes	No
Overcast	83	86	No	Yes
Rainy	70	96	No	Yes
Rainy	65	70	Yes	No
Overcast	64	65	Yes	Yes
Overcast	72	90	Yes	Yes
Sunny	75	70	Yes	Yes
Rainy	68	80	No	Yes
Overcast	81	75	No	Yes
Sunny	85	85	No	No
Sunny	72	95	No	No
Rainy	75	80	No	Yes

Weather			Temperature			Humidity			Wind		
	Yes	No		Yes	No		Yes	No		Yes	No
Sunny	2/9	3/5	Mean	73	74.6	Mean	79.1	86.2	False	6/9	2/5
Overcast	4/9	0/5	SD	6.2	7.9	SD	10.2	9.3	True	3/9	3/5
Rainy	3/9	2/5									

$$N(x) = \frac{1}{\sqrt{2\pi\sigma}} e - \frac{(x-\mu)^2}{2\sigma^2}$$

μ: μέση τιμή των τιμών της μεταβλητής

μ= (τιμη1+τιμη2+τιμη3+…+τιμηN)/N

σ: τυπική απόκλιση των τιμών της μεταβλητής

$$\sigma = \sqrt{\frac{(\tau\iota\mu\eta 1-\mu)^2 + (\tau\iota\mu\eta 2-\mu)^2 + (\tau\iota\mu\eta 3-\mu)^2 + \ldots + (\tau\iota\mu\eta \mathrm{N}-\mu)^2}{\mathrm{N}-1}}$$

$$P(temp = 80 \mid PLAY) = \frac{1}{\sqrt{2\pi} * 6.2} e^{\frac{(80-73)^2}{2*6.2^2}}$$

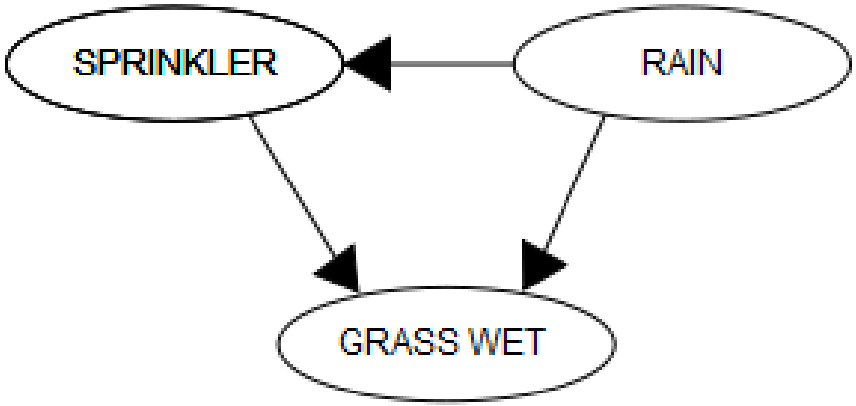
$$P(temp = 80 \mid NOPLAY) = \frac{1}{\sqrt{2\pi} * 7.9} e^{\frac{(80-74.6)^2}{2*7.9^2}}$$

Δίκτυο Bayes

- Φορμαλισμός που μοντελοποιεί εξαρτήσεις αιτιατότητας με γραφική αναπαράσταση
- Είναι ένας κατευθυνόμενος ακυκλικός γράφος με
  - Ένα σύνολο κόμβων, κάθε κόμβος αναπαριστά μια μεταβλητή
  - Μια δεσμευμένη κατανομή για κάθε κόμβο, δεδομένων των μητρικών του κόμβων (γονέων)
    - P(Xi| γονείς(Xi))
  - Έναν πίνακα δεσμευμένων πιθανοτήτων (Conditional probability Table - CPT) που δίνει την κατανομή ως προς το Xi κάθε συνδυασμού τιμών των μητρικών κόμβων

Δίκτυα Bayes: Υπάρχουν αλληλεξαρτήσεις ανάμεσα στα χαρακτηριστικά - Ένα απλό παράδειγμα

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



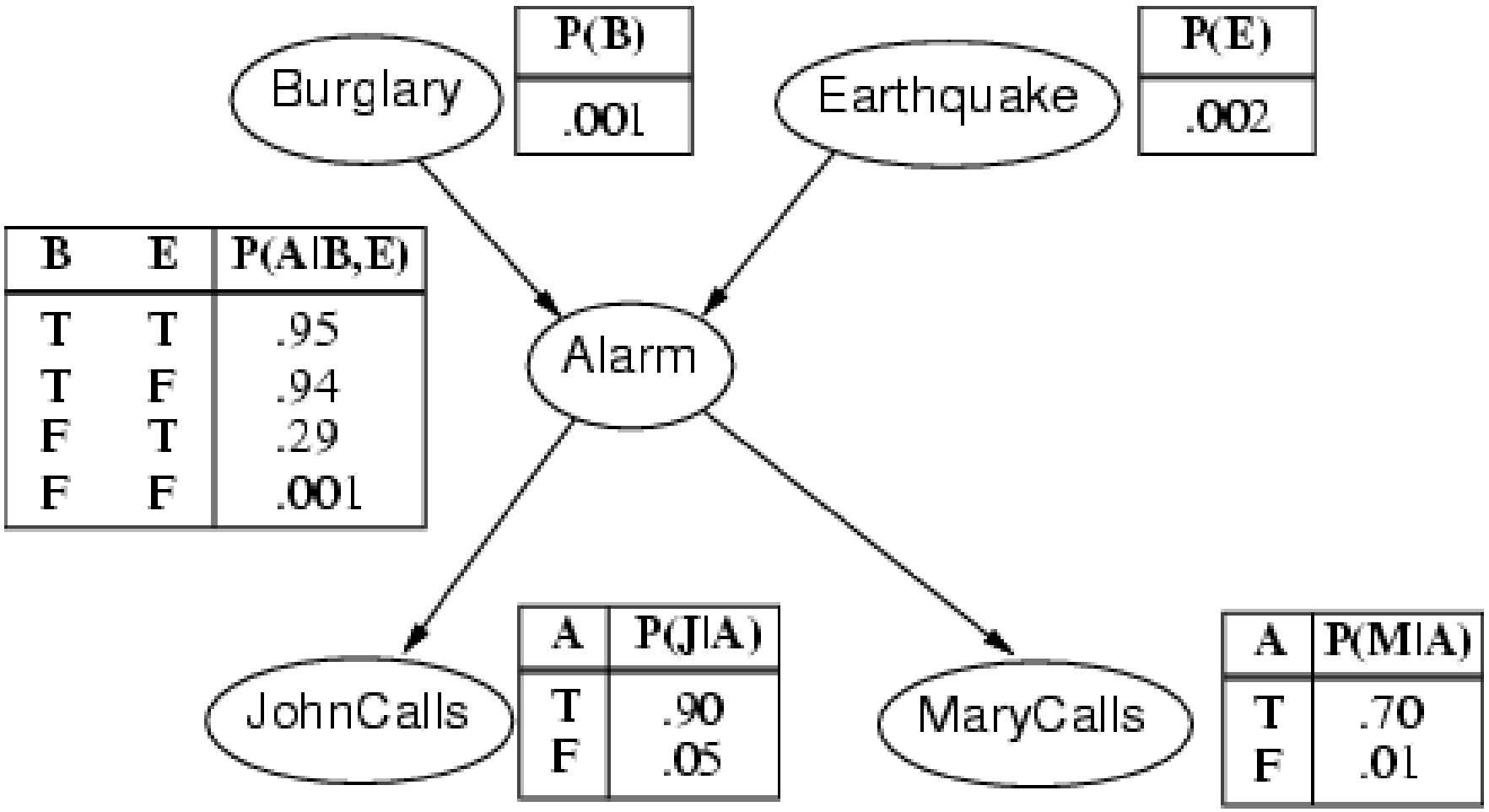
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Άλλο παράδειγμα

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- **Μεταβλητές**
  - Burglary, Earthquake, Alarm, JohnCalls, MaryCalls
- **Η τοπολογία του δικτύου αναπαριστά την γνώση αιτιότητας (τι συνδέεται με τι):**
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Παράδειγμα του δικτύου



Ολική από κοινού κατανομή (Full Joint Distribution)

- Η ολική από κοινού κατανομή ορίζεται σαν το γινόμενο των τοπικών δεσμευμένων κατανομών (local conditional distributions)
- $P(X_1, \ldots, X_n) = \pi \; P(X_i \mid \text{γονείς}(X_i))$
- Λαμβάνονται υπόψην μόνο οι άμεσοι γονείς.

Παράδειγμα υπολογισμού

- Μια μέρα παίρνει τηλέφωνο ο John και η Mary, χτυπάει το alarm, δεν έχει γίνει ληστεία. Έχει γίνει σεισμός;

$$P(E|J, M, A, \sim B) = \frac{P(E, J, M, A, \sim B)}{P(J, M, A, \sim B)} = \frac{P(E, J, M, A, \sim B)}{P(E, J, M, A, \sim B) + P(\sim E, J, M, A, \sim B)}$$

P(E, J,M,A,¬B)=

P(E) · P(J|A) · P(M|A) · P(A|¬B,E) · P(¬B)
= 0,002 · 0,9 · 0,7 · 0,29 · 0,999

P(¬E,J,M,A,¬B)=

P(¬E) · P(J|A) · P(M|A) · P(A|¬B, ¬E) · P(¬B)
= 0,998 · 0,9 · 0,7 · 0,001 · 0,999

Μάθηση Δικτύου Bayes από δεδομένα Αλγόριθμος k2

- Στόχος: δεδομένου ενός σετ παραδειγμάτων η εύρεση της πιο πιθανής δομής δικτύου Bayes
- Σαν αρχική κατάσταση εισάγεται μια αρχική δομή γράφου (κατάταξη κόμβων).
- Καθορίζεται μια μέγιστη τιμή μητρικών κόμβων για έναν κόμβο.
- Στην κατάταξη των κόμβων, κόμβοι που εμφανίζονται κάτω από έναν κόμβο δεν μπορούν να είναι μητρικοί του κόμβοι
- Αρχικά το σετ PA_i των μητρικών κόμβων του κόμβου i είναι κενό.
- Υπολογίζεται το σκορ του δικτύου
- Οι κόμβοι στην κατάταξη ελέγχονται σειριακά
- Για κάθε κόμβο υπολογίζεται η αύξηση του σκορ του δικτύου αν αυτός προστεθεί στο PA_i
- Ο κόμβος που μεγιστοποιεί το σκορ του δικτύου προστίθεται στο PA_i
- Η προσθήκη γονέων στο PA_i συνεχίζεται
  - Μέχρι να φτάσουν τον μέγιστο αριθμό γονέων για τον κόμβο
  - Μέχρι να μην υπάρχουν άλλοι κόμβοι για προσθήκη
  - Μέχρι η προσθήκη γονέα να μην αυξάνει το σκορ του δικτύου
- Ο αλγόριθμος τερματίζει όταν όλοι οι κόμβοι στην κατάταξη έχουν δεχτεί επίσκεψη μια φορά

Το Σκορ του Δικτύου

$$P(B_s, D) = P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}N_{ij}\prod_{k=1}^{r_i}N_{ijk}!$$

D - dataset, it has m cases(records)

Z - a set of n discrete variables: (x_1, …, x_n)

r_i - a variable x_i in Z has r_i possible value assignments: $(v_{i1},...v_{ir_i})$

B_s - a bayes network structure containing just the variables in Z

π_i - each variable x_i in B_s has a set of parents which we represent with a list of variables π_i

q_i - there are unique instantiations of π_i (το καρτεσιανό γινόμενο όλων των τιμών των γονέων του x_i)

w_{ij} - πόσες φορές στα δεδομένα οι γονείς εμφανίζονται με τον jth συνδυασμό τιμών (unique instantiation of π_i relative to D).

N_{ijk} –Σε πόσες από τις φορές που οι γονείς έχουν στα δεδομένα συνδυασμό τιμών w_{ij} και η μεταβλητή x_i έχει τιμή v_{ik}

N_{ij} - $$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

Αλγόριθμος TAN (Tree-augmented Naïve Bayes)

- Προσθέτει ακμές σε ένα δίκτυο Naïve Bayes
- Συγκεκριμένα, ελέγχεται η προσθήκη ενός δεύτερου γονέα (εκτός της κλάσης ταξινόμησης) σε κάθε κόμβο.
- Η κλάση δεν έχει γονείς
- Οι κόμβοι επιτρέπεται να πάρουν μόνο δομή δένδρου
- Το βάρος μιας ακμής είναι η αμοιβαία πληροφορία των δύο κόμβων της
- Βρίσκει το δέντρο με το μέγιστο βάρος
- Αλγόριθμος
  - Αρχικά σχηματίζεται ένας πλήρης γράφος ανάμεσα στους κόμβους εκτός της κλάσης
  - Υπολογίζεται το δέντρο με το μέγιστο βάρος που εκτείνεται στον γράφο
  - Διαλέγει ρίζα για το δέντρο και δίνει κατεύθυνση στις ακμές του
  - Πρόσθεσε τις ακμές από το δέντρο στο δίκτυο

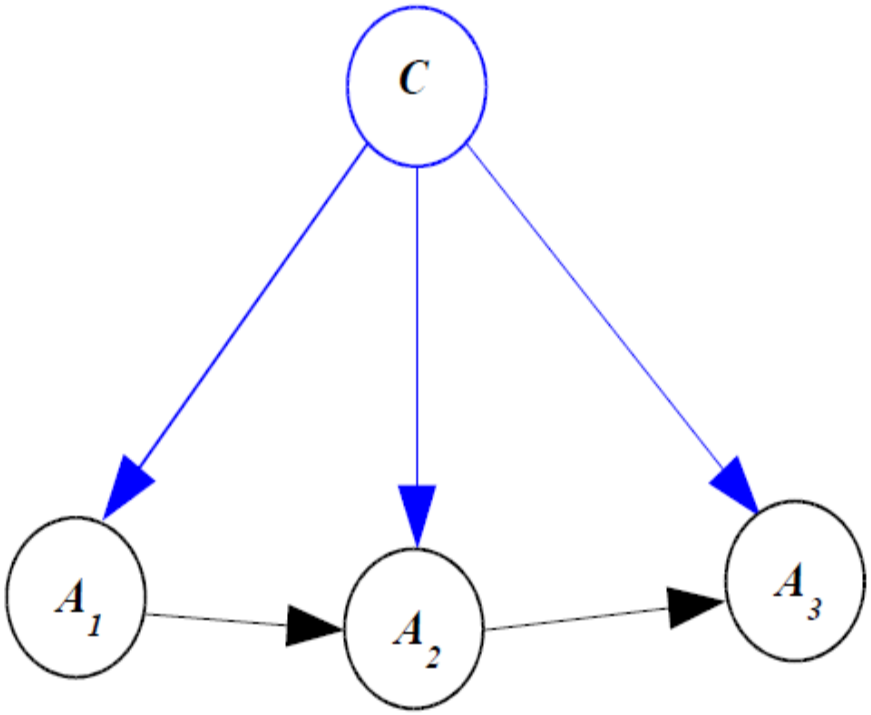


Figure 4.2 A Tree Augmented Network (TAN) with class variable *C* and three attributes nodes A_1, A_2 and A_3. *C* is connected to all the attributes. There are also directed edges from A_1 to A_2 and A_2 to A_3.

Table 4.1: The Construct-tree procedure of CL [Chow and Liu, 1968]

1. Compute $I_P(X_i ; X_j)$ between each pair of variables, i is not equal to j
2. Build a complete undirected graph in which the vertices are the variables in $\boldsymbol{X}$. Annotate the weight of an edge connecting $\boldsymbol{X_i}$ to $\boldsymbol{X_j}$ by $\boldsymbol{I_P(X_i ; X_j)}$
3. Build a maximum weighted spanning tree
4. Transform the resulting undirected tree to a directed tree one by choosing a root variable and setting the direction of all edges to be outward from it

$$I_p(X;Y)=\sum_{x\in X, y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (4.1)$$

Πλεονεκτήματα των δικτύων Bayes

- Παρέχουν μια φυσική αναπαράσταση για συσχετίσεις αιτιότητας
- Επιτρέπουν την αναπαράσταση αβέβαιων παρατηρήσεων
- Η τοπολογία και οι πίνακες CPTs αποτελούν συμπαγή αναπαράσταση της ολικής από κοινού κατανομής
- Είναι συνήθως εύκολο να κατασκευαστούν από ειδικούς της θεματικής περιοχής (domain experts)