Data mining, machine learning, and uncertainty reasoning

林偉川

條件機率

- 在事件A_i發生前提下，事件B_j發生的機率稱爲條件機率，以$P(B_j|A_i)$表示，其公式爲：

$$P(B_j \mid A_i) = \frac{P(A_i \cap B_j)}{P(A_i)} \quad , \ P(A_i)\neq 0$$

- 在事件B_j 發生前提下，事件A_i發生的機率稱爲條件機率，以$P(A_i|B_j)$表示，其公式爲：

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)} \ , \ P(B_j)\neq 0$$

- 機率乘法法則：

$P(B_j \cap A_i)= P(A_i) * P(B_j|A_i)$ ，$P(B_j \cap A_i)= P(B_j) * P(A_i|B_j)$

邊際機率

- 屬性A的事件A_1與屬性B的事件B_1同時發生的機率稱爲邊際機率，以$P_{11}=P(A_1 \cap B_1)$表示，事件A_i與事件B_j同時發生的機率，以$P_{ij}=P(A_i \cap B_j)$表示，其中i=1,2,…,m且j=1,2,…,n

$$P(B_1)= P(A_1 \cap B_1)+ P(A_2 \cap B_1)+ P(A_3 \cap B_1)+\ldots+ P(A_m \cap B_1)$$

$$=P_{11}+ P_{21}+ \ldots+ P_{m1}$$

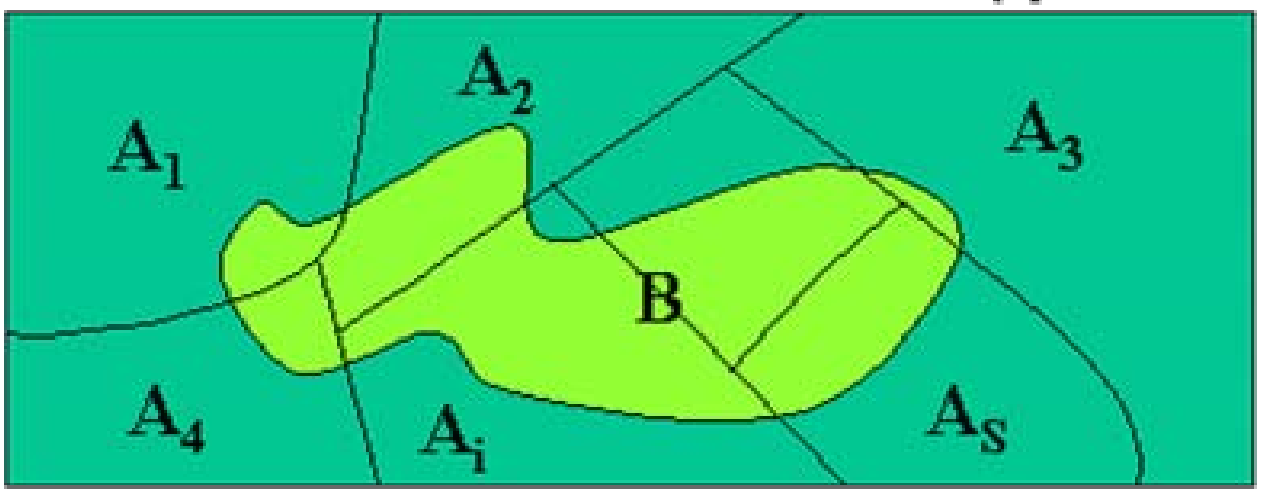
Bayes' Theorem

事前機率　條件機率(額外訊息)

事後機率

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

$$P(A_K \mid B) = \frac{P(A_K \cap B)}{P(B)} = \frac{P(A_K)P(B \mid A_K)}{\sum_{i=1}^{S} P(A_i)P(B \mid A_i)}$$



Bayes' Theorem

- $P(B_j) = P(A_1 \cap B_j) + P(A_2 \cap B_j) + P(A_3 \cap B_j) + \ldots + P(A_m \cap B_j)$
- $P(A_i \cap B_j) = P(A_i) * P(B_j|A_i)$ $P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}$
- $P(A_1 \cap B_j) = P(A_1) * P(B_j|A_1)$, so

  $P(A_2 \cap B_j) = P(A_2) * P(B_j|A_2)$ …

  $P(A_m \cap B_j) = P(A_m) * P(B_j|A_m)$
- $P(B_j) = P(A_1) * P(B_j|A_1) + P(A_2) * P(B_j|A_2) + \ldots + P(A_m) * P(B_j|A_m)$

Bayes' theorem

- 於是在事件B_j發生的前題下，事件A_i發生的條件機率爲

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)} = \frac{P(A_i)P(B_j \mid A_i)}{P(A_1)P(B_j \mid A_1) + P(A_2)P(B_j \mid A_2) + \ldots + P(A_m)P(B_j \mid A_m)}$$

- 在此定理中，A_1、A_2、…A_m爲m個相互互斥的事件，而$P(A_1)$、 $P(A_2)$、… $P(A_m)$稱爲事前機率，$P(B_j|A_1)$、$P(B_j|A_2)$、… $P(B_j|A_m)$稱爲條件機率， $P(A_i|B_j)$則爲事後機率

Example of Bayes' theorem

- 某工廠有三條生產線以A_1、A_2、及A_3表示，由各生產線的產能可知：A_1生產線產量的四成， $P(A_1)=0.4$， A_2及A_3生產線產量的比例爲$P(A_2)=P(A_3)=0.3$，令B_j爲生產線是不良品的事件， $P(B_j|A_1)=0.01$(由A_1生產線生產的不良品機率)、$P(B_j|A_2)$ =0.02、 $P(B_j|A_3)$ =0.03，由貝式定理可知不良品B_j是從A_1生產線生產出來的事後機率爲

Example of Bayes' theorem

$P(A_1|B_j)=(0.4*0.01)/(0.4*0.01+0.3*0.02+0.3*0.03)=0.2105$

$P(A_2|B_j)=(0.3*0.02)/(0.4*0.01+0.3*0.02+0.3*0.03)=0.3158$

$P(A_3|B_j)=(0.3*0.03)/(0.4*0.01+0.3*0.02+0.3*0.03)=0.4737$

➔ A_3生產線應該要檢討了！

Example of Bayes' theorem

- 從台北火車站前往世貿中心，假設目前有三條可行道路分別爲信義路、忠孝東路、及南京東路，各路線被選中的機率分別爲：P(A)=0.3，P(B)=0.4，P(C)=0.3，假設走信義路塞車的機率爲0.4，走忠孝東路塞車的機率爲0.5，走南京東路塞車的機率爲0.3，試問不塞車以走那條路線爲宜？
(答案➔南京東路why?)

Answer of 塞車範例

- P(A)=0.3，P(B)=0.4，P(C)=0.3，G表示塞車，P(G|A)=0.4表示走信義路塞車機率，P(G|B)=0.5表示走忠孝東路塞車機率，P(G|C)=0.3表示走南京東路塞車機率，

$$P(A|G)=(0.3*0.4)/(0.3*0.4+0.4*0.5+0.3*0.3)=12/23$$

$$P(B|G)=(0.4*0.5)/(0.3*0.4+0.4*0.5+0.3*0.3)=20/23$$

$$P(C|G)=(0.3*0.3)/(0.3*0.4+0.4*0.5+0.3*0.3)=9/23$$ (塞車機率最小!!!)

Bayesian Networks

- The concept of conditional probability is a useful one.
- There are countless real world examples where the probability of one event is conditional on the probability of a previous one.

Bayesian Networks

- While the sum and product rules of probability theory can anticipate this factor of conditionality, in many cases such calculations are NP-hard.
- The prospect of managing a scenario with 5 discrete random variables (2^5-1=31 discrete parameters) might be manageable.
- An expert system for monitoring patients with 37 variables resulting in a joint distribution of over 2^{37} parameters would not be manageable

Baye's logic

- Bayesian Theorem provided, for the first time, a mathematical method that could be used to calculate, given occurrences in prior trials, the likelihood of a target occurrence in future trials.
- According to Baye's logic, the only way to quantify a situation with an uncertain outcome is through determining its probability.

Baye's logic

- Baye's Theorem is a means of quantifying uncertainty.
- Based on probability theory, the theorem defines a rule for refining an hypothesis by factoring in additional evidence and background information, and leads to a number representing the degree of probability that the hypothesis is true.

Example of Bayesian network

- Given a situation where it might rain today, and might rain tomorrow, what is the probability that it will rain on both days?
- Rain on two consecutive days are not independent events with isolated probabilities.
- Solving such a problem involves determining the chances that it will rain today, and then determining the chance that it will rain tomorrow conditional on the probability that it will rain today.

Example of Bayesian network

- These are known as "joint probabilities." Suppose that P(rain today) = 0.20 and P(rain tomorrow given that it rains today) = 0.70. The probability of such joint events is determined by: $P(E_1, E_2) = P(E_1)P(E_2|E_1)$ which can also be expressed as:

$$P(E_2|E_1) = \frac{P(E_1, E_2)}{P(E_1)}$$

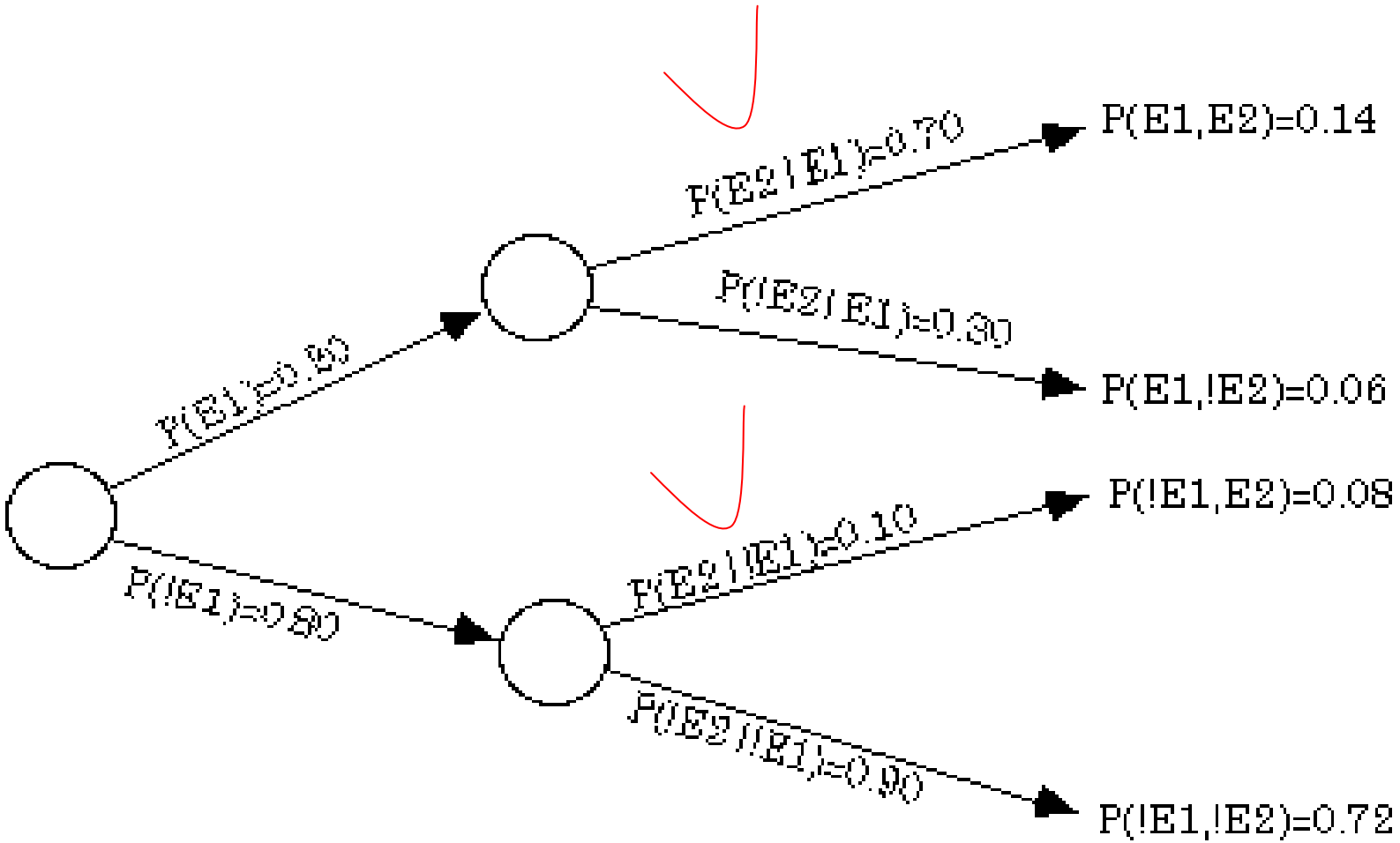
- Working out the joint probabilities, the results can be expressed in a table

Marginal and Joint Probabilities for rain both today and tomorrow			
	Rain Tomorrow	No Rain Tomorrow	Marginal Probability of Rain Today
Rain Today	0.14	0.06	0.2
No Rain Today	0.08	0.72	0.8
Marginal Probability of Rain Tomorrow	0.22	0.78	

Example of Bayesian network

- From the table, it is evident that the joint probability of rain over both days is 0.14, but there is a great deal of other information that had to be brought into the calculations before such a determination was possible. With only two discrete, binary variables, four calculations were required.
- This same scenario can be expressed using a Bayesian Network Diagram as shown ("!" is used to denote "not").

Bayesian Network Diagram



Bayesian Network Diagram

- While the probability of rain today and the probability of rain tomorrow are two discrete events (it cannot rain both today and tomorrow at the same time), there is a conditional relationship between them (if it rains today, the lingering weather systems and residual moisture are more likely to result in rain tomorrow).
- For this reason, the directed edges of the graph are connected to show this dependency.

Bayesian Network Diagram

- One attraction of Bayesian Networks is the efficiency that only one branch of the tree needs to be traversed. We are really only concerned with P(E1), P(E2|E1) and P(E2,E1).
- We can also utilize the graph both visually and algorithmically to determine which parameters are independent of each other

Bayesian Network Diagram

- Instead of calculating four joint probabilities, we can use the independence of the parameters to limit our calculations to two.
- It is self-evident that the probabilities of rain on the second day having rained on the first are completely autonomous (獨立自主) from the probabilities of rain on the second day having not rained on the first.

Bayesian Network Diagram

- At the same time as emphasizing parametric indifference, Bayesian Networks also provide a parsimonious (儉省) representation of conditionality among parametric relationships.

Example of Bayesian network

- Friedman and Goldszmidt suggest looking at Bayesian Networks as an example containing 5 random variables: "Burglary (B)", "Earthquake (E)", "Alarm (A)", "Neighbor Call (C)", and "Radio Announcement (R)".
- In such a story, "Burglary" and "Earthquake" are independent, and "Burglary" and "Radio Announcement" are independent given "Earthquake."

Advantages of Bayesian network

- Using a Bayesian Network offers many advantages over traditional methods of determining causal relationships.
- Independence among variables is easy to recognize and isolate while conditional relationships are clearly delimited by a directed graph edge: two variables are independent if all the paths between them are blocked (given the edges are directional).

Example of Bayesian network

- This is to say that there is no event which effects both burglaries and earthquakes.
- As well, “Burglary” and “Radio Announcements” are independent given “Earthquake”--meaning that while a radio announcement might result from an earthquake, it will not result as a repercussion (影響) from a burglary.

Example of Bayesian network

- Because of the independence among these variables, the probability of P(A,R,E,B) (The joint probability of an alarm, radio announcement, earthquake and burglary) can be reduced from:

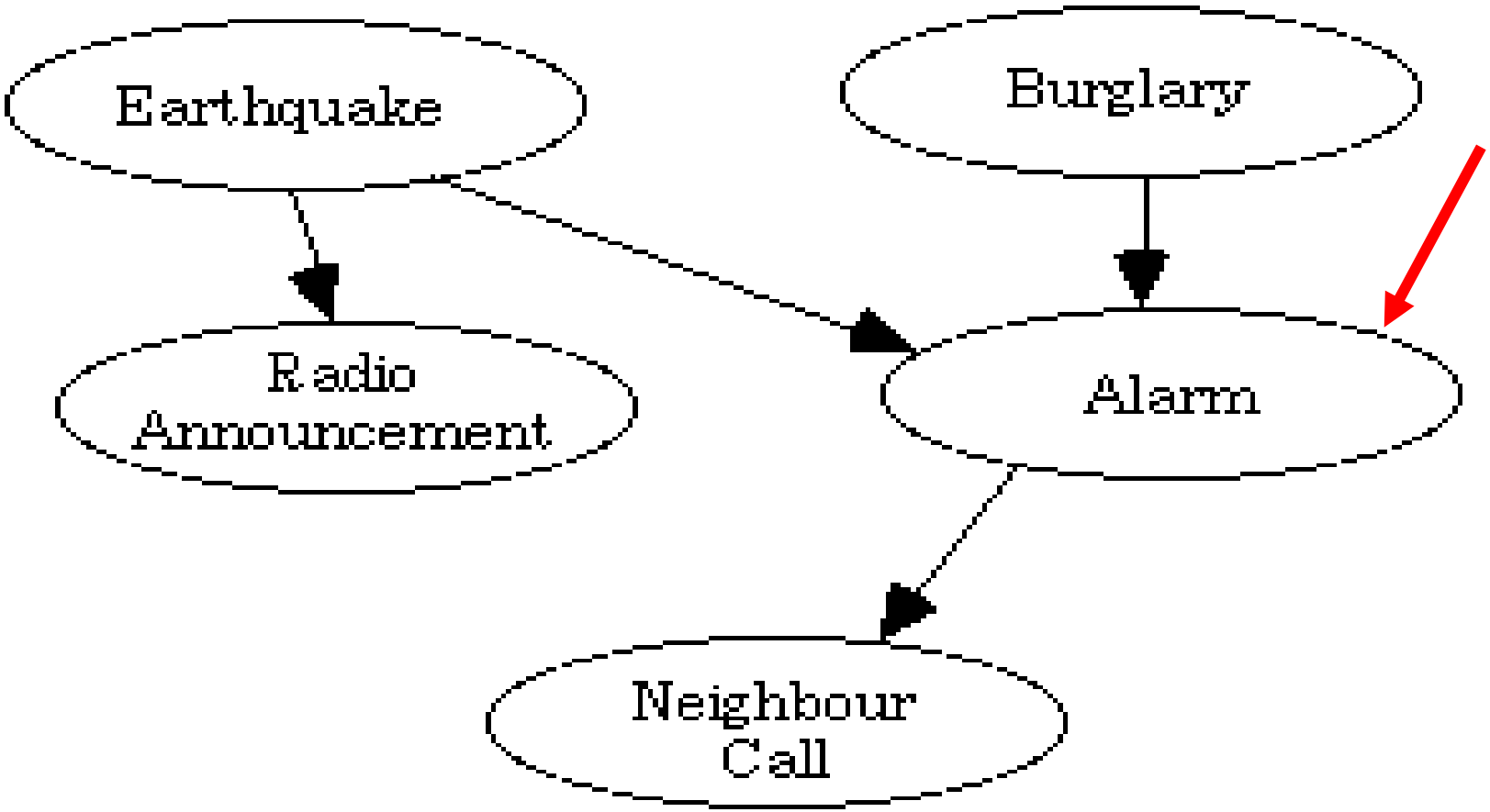
$$P(A,R,E,B)=P(A|R,E,B)*P(R|E,B)*P(E|B)*P(B)$$

  involving 15 parameters to 8:

$$P(A,R,E,B) = P(A|E,B)*P(R|E)*P(E)*P(B)$$

- This significantly reduced the number of joint probabilities involved.

The corresponding of Bayesian network



Bayesian network

- Not all the joint probabilities need to be calculated to make a decision; extraneous branches and relationships can be ignored (One can make a prediction of a radio announcement regardless of whether an alarm sounds).
- By optimizing the graph, every node can be shown to have at most *k* parents.

Bayesian network

- The algorithmic routines required can then be run in $O(2n)$ instead of $O(2^k n)$ time. In essence, the algorithm can run in linear time (based on the number of edges) instead of exponential time (based on the number of parameters).
- Associated with each node is a set of conditional probability distributions. For example, the "Alarm" node might have the following probability distribution:

Bayesian network table

Probability Distribution for the Alarm Node given the events of "Earthquakes" and "Burglaries" ("!" denotes "not")			
E	B	P(A | E,B)	P(!A|E,B)
E	B	0.9	0.1
E	!B	0.2	0.8
!E	B	0.9	0.1
!E	!B	0.01	0.99

Example conclusion

- Should there be both an earthquake and a burglary, the alarm has a 90% chance of sounding.
- With only an earthquake and no burglary, it would only sound in 20% of the cases.
- A burglary unaccompanied by an earthquake would set on the alarm 90% of the time, and the chance of a false alarm given no antecedent event should only have a probability of 0.1% of the time.