REASONING UNDER UNCERTAINTY

Universidad Carlos III de Madrid

AI

Outline

Outline

Outline

Outline

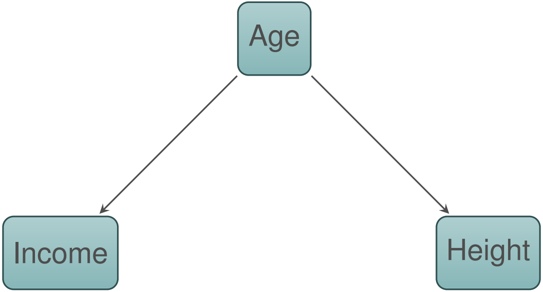
Previous class

- Representation on domains with random variables + probability distribution
- Given probability distribution for all possible events, we can solve queries $P(\text{Variables}|\text{Observation})$
- The distribution size is exponential on the number of variables
- Independence could allow us a more efficient reasoning
- Today: how can we use probabilities more efficiently?
  - Answer: Bayesian networks

Outline

Conditional independence

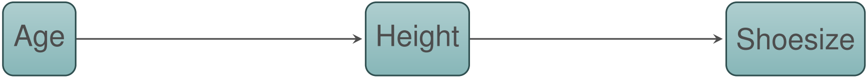
$$P(\text{Income}|\text{Height},\text{Age}) = P(\text{Income}|\text{Age})$$



Income and Height are conditionally independent "given" Age

Conditional independence

$$P(\text{Shoesize}|\text{Height},\text{Age}) = P(\text{Shoesize}|\text{Height})$$



- Age and Shoesize are conditionally independent “given” Height

Conditional independence

- X and Y are conditionally independent given Z if

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

Also:

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

Conditional independence

- X and Y are conditionally independent given Z if

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

  Also:

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

- It often reduces the number of parameters from exponential in n (number of variables) to linear in n

Conditional independence

- X and Y are conditionally independent given Z if

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

  Also:

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

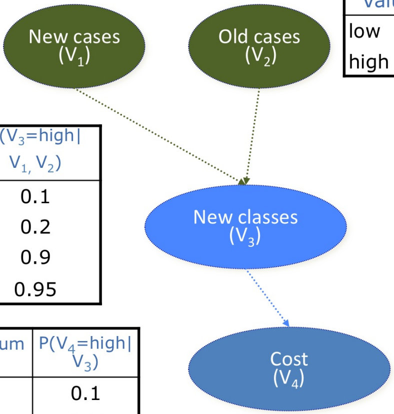
- It often reduces the number of parameters from exponential in n (number of variables) to linear in n
- Conditional independence is the tool for efficient probabilistic reasoning
  - less parameters
  - less computation
- It is represented by the missing edges

Outline

Bayesian network

Value	$P(V_1\text{=value})$
low	0.7
high	0.3

Value	$P(V_2\text{=value})$
low	0.3
high	0.7



V_1	V_2	$P(V_3\text{=low} \mid V_1, V_2)$	$P(V_3\text{=high} \mid V_1, V_2)$
low	low	0.9	0.1
low	high	0.8	0.2
high	low	0.1	0.9
high	high	0.05	0.95

V_3	$P(V_4\text{=low} \mid V_3)$	$P(V_4\text{=medium} \mid V_3)$	$P(V_4\text{=high} \mid V_3)$
low	0.7	0.2	0.1
high	0.1	0.2	0.7

Definition of Bayesian Network

- A set of nodes
  - each node represents a random variable
  - variables can be either discrete or continuous
- A set of edges
  - an edge from node X to node Y: X has a direct influence on Y
  - it is a Direct Acyclic Graph (DAG)
- Probability distributions
  - each node X has a Conditional Probability Table (CPT) that defines the effects of its parents

$$P(\text{Node}|\text{Parents}(\text{Node}))$$

  - parents of node X are the only edges directed to X
  - if a node does not have parents, it is the “a priori” probability

$$P(\text{Node})$$

Example of an alarm

- We have an anti-theft system at home with an alarm
- It detects robbers, but the alarm also fires with some earthquakes
- There are two neighbors (Juan and Maria) that will call us if they hear the alarm
- Juan always calls when he hears the alarm, but he sometimes is confused with some door bell
- Maria hears music very loud, so sometimes she cannot hear the alarm

Modelling

- Earthquake: E
  - E=T $\rightsquigarrow$ e
  - E=F $\rightsquigarrow$ $\neg$ e

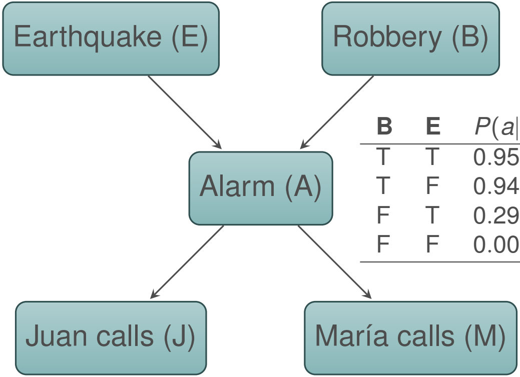
Modelling

- Earthquake: E
  - E=T $\rightsquigarrow$ e
  - E=F $\rightsquigarrow$ $\neg$ e
- Robbery: B
  - B=T $\rightsquigarrow$ b
  - B=F $\rightsquigarrow$ $\neg$ b
- Alarm: A (a, $\neg$ a)
- Juan calls: J (j, $\neg$ j)
- María calls: M (m, $\neg$ m)

Complete BN for the Alarm example

$P(e) = 0.002$

$P(b) = 0.001$



B	**E**	$P(a \vert B, E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	$P(j \vert A)$
T	0.90
F	0.05

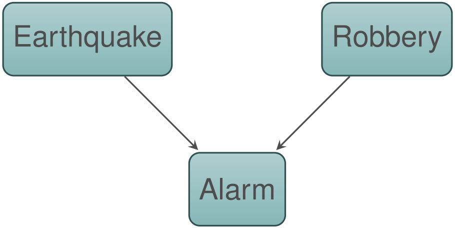
A	$P(m \vert A)$
T	0.70
F	0.001

Alarm. Causal relations

- We only provide $P(e)$ given that $P(\neg e) = 1 - P(e)$
- Also, $P(\neg a|b, \neg e) = 1 - P(a|b, \neg e)$
- The topology of this BN reflects the direct causes of its variables:
    - a robber can fire the alarm
    - an earthquake can fire the alarm
    - the alarm can cause Maria to call
    - the alarm can cause Juan to call

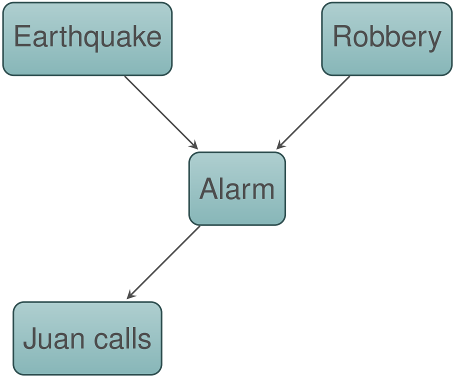
BN for the example

- There is no dependency between Earthquake and Robbery
- But, there is dependency between Alarm and the other two variables:
  - $P(\text{Alarm}|\text{Earthquake}, \text{Robbery}) \neq P(\text{Alarm}|\text{Earthquake})$
  - $P(\text{Alarm}|\text{Earthquake}, \text{Robbery}) \neq P(\text{Alarm}|\text{Robbery})$



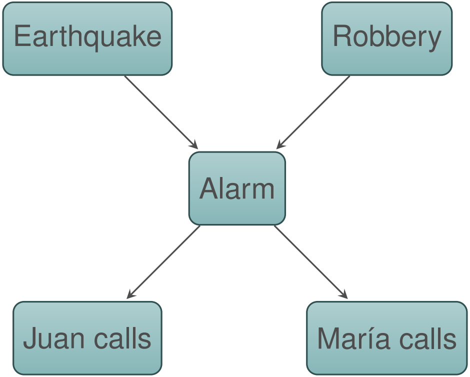
BN for the example

- There is conditional independence between Juan calling and variables Earthquake and Robbery, given the Alarm variable
  - $P(\text{Juan}|\text{Alarm}, \text{Earthquake}, \text{Robbery}) = P(\text{Juan}|\text{Alarm})$



BN for the example

- The same applies to María calling
  - $P(\text{Maria}|\text{Alarm}, \text{Earthquake}, \text{Robbery}) = P(\text{Maria}|\text{Alarm})$

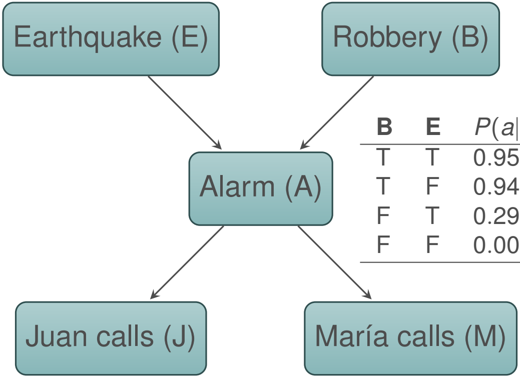


BN are compact

- The explicit joint distribution would require $2^5 - 1 = 31$ parameters
- The BN uses $1 + 1 + 4 + 2 + 2 = 10$ parameters

$P(e) = 0.002$

$P(b) = 0.001$



B	**E**	$P(a \mid B, E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	$P(j \mid A)$
T	0.90
F	0.05

A	$P(m \mid A)$
T	0.70
F	0.001

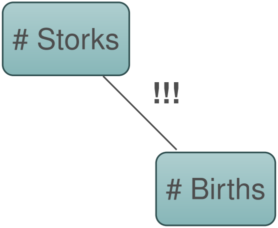
BN are compact

- A CPT for a propositional variable X with k propositional parents has 2^k rows, one for each combination of parent values
- Each row has a value p for $X = true$ ($X = false$ would be $1 - p$)
- If each variable has a maximum of k parents, the complete BN requires $O(n \cdot 2^k)$ parameters
  - that grows linearly in n, vs. the explicit joint probability distribution that requires $O(2^n)$
- The ordering of variables to add to the BN can greatly influence the resulting BN

Causality and correlation

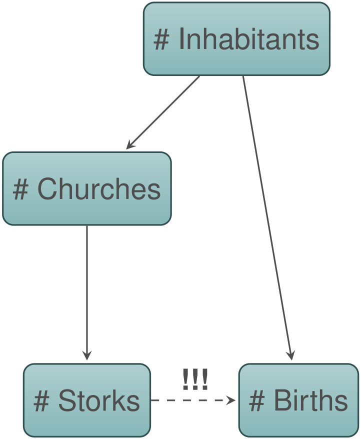
How do we build a BN?

- A study showed that there is a strong correlation between the number of storks in a city and the number of births



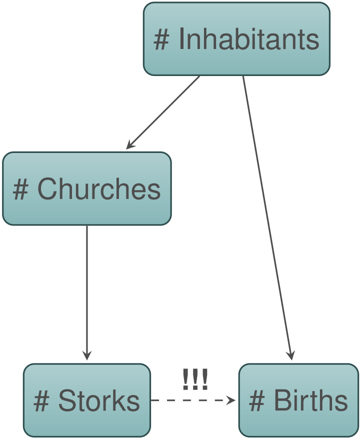
Causality and correlation

- A study showed that there is a strong correlation between the number of storks in a city and the number of births



Causality and correlation

- A study showed that there is a strong correlation between the number of storks in a city and the number of births



- Causality implies correlation
- But, correlation does not imply causality

Semantics of BNs

Global semantics: the joint probability distribution is the product of local distributions

$$\begin{aligned} P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \\ &= \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)) \end{aligned}$$

Outline

Basic Bayesian inference

Bayes theorem

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

Law of total probability

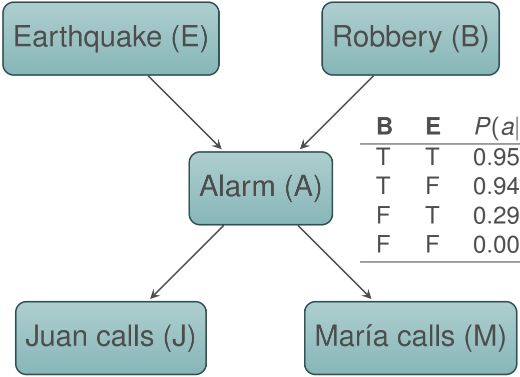
If we have variables $A_1, \ldots, A_n$ and an event B:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Alarm example

$P(e) = 0.002$

$P(b) = 0.001$



B	**E**	$P(a \mid B, E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	$P(j \mid A)$
T	0.90
F	0.05

A	$P(m \mid A)$
T	0.70
F	0.001

Exact inference. Enumeration

- Events: b= robbery, e = earthquake, a = alarm, j = juan calls, m = maría calls
- Query: probability of robbery given that both Juan and María called: $P(b|j, m)$

Exact inference. Enumeration

- Events: b= robbery, e = earthquake, a = alarm, j = juan calls, m = maría calls
- Query: probability of robbery given that both Juan and María called: $P(b|j, m)$
- We compute the query by using the definition of conditional probability:

$$P(b|j, m) = \frac{P(b, j, m)}{P(j, m)} = \alpha P(b, j, m)$$

Exact inference. Enumeration

- Events: b= robbery, e = earthquake, a = alarm,
  j = juan calls, m = maría calls
- Query: probability of robbery given that both Juan and María called: $P(b|j, m)$
- We compute the query by using the definition of conditional probability:

$$P(b|j, m) = \frac{P(b, j, m)}{P(j, m)} = \alpha P(b, j, m)$$

- We do not have $P(b, j, m)$, but we can use the joint distribution of all variables in the BN: $P(B, E, A, J, M)$
  - the value of output variable is fixed, $B = T(b)$
  - the values of evidences are fixed, $J = T(j), M = T(m)$
  - the values of all hidden variables (E, A) are summed

Exact inference. Enumeration. Algorithm

- Using the law of total probability:

$$
\begin{aligned}
P(b|j,m) &= \alpha P(b,j,m) \\
&= \alpha \sum_{e'=\{v,f\}} \sum_{a'=\{v,f\}} P(b, E=e', A=a', j, m)
\end{aligned}
$$

- Then, the joint distribution is decomposed by using BN:

$$
\begin{aligned}
P(b|j,m) &= \alpha \sum_{e'} \sum_{a'} P(b)P(e')P(a'|b,e')P(j|a')P(m|a') \\
&= \alpha P(b) \sum_{e'} P(e') \sum_{a'} P(a'|b,e')P(j|a')P(m|a')
\end{aligned}
$$

- And all these probabilities can be obtained from the CPTs of the BN

Exact inference for the Alarm example

$$
\begin{aligned}
P(b|j,m) = \alpha P(b)[&P(e)(P(a|b,e)P(j|a)P(m|a)+\\
&\qquad\qquad P(\neg a|b,e)P(j|\neg a)P(m|\neg a))+\\
&P(\neg e)(P(a|b,\neg e)P(j|a)P(m|a)+\\
&\qquad\qquad P(\neg a|b,\neg e)P(j|\neg a)P(m|\neg a))]
\end{aligned}
$$

Exact inference for the Alarm example

$$
\begin{aligned}
P(b|j,m) = \alpha 0.001[0.002(&0.95 \times 0.90 \times 0.70+\\
&0.05 \times 0.05 \times 0.001)+\\
0.998(&0.94 \times 0.90 \times 0.70+\\
&0.06 \times 0.05 \times 0.001)]
\end{aligned}
$$

Exact inference. Enumeration

- To compute the normalization factor $\alpha = \frac{1}{P(j,m)}$
  - we perform the same steps for the other case

$$P(\neg b, j, m)$$

  - and apply the Law of total probability

$$P(j, m) = P(b, j, m) + P(\neg b, j, m)$$

- $P(B|j, m) = \langle 0.284, 0.716 \rangle$
- Meaning: $P(B = T|j, m) = 28.4\%$, $P(B = F|j, m) = 71.6\%$

Exact inference. Enumeration

- To compute the normalization factor $\alpha = \frac{1}{P(j,m)}$
  - we perform the same steps for the other case

$$P(\neg b, j, m)$$

  - and apply the Law of total probability

$$P(j, m) = P(b, j, m) + P(\neg b, j, m)$$

- $P(B|j, m) = \langle 0.284, 0.716 \rangle$
- Meaning: $P(B = T|j, m) = 28.4\%$, $P(B = F|j, m) = 71.6\%$
- Problem of enumeration: exponential complexity (in the worst case)
- If we have N hidden variables, each with M values, we have to compute M^N values

Exact inference. Caching

$$P(b|j,m) = \alpha P(b)[P(e)(P(a|b,e)P(j|a)P(m|a)+$$

$$P(\neg a|b,e)P(j|\neg a)P(m|\neg a))+$$

$$P(\neg e)(P(a|b,\neg e)P(j|a)P(m|a)+$$

$$P(\neg a|b,\neg e)P(j|\neg a)P(m|\neg a))]$$

- It can be cumbersome, specially with bigger BN
- It can be alleviated by noticing repetitions: $P(m|a)$, $P(m|\neg a)$, $P(j|a)$, $P(j|\neg a)$
- But, more importantly: $P(j|a)P(m|a)$, $P(j|\neg a)P(m|\neg a)$
- Can be arbitrarily big formulae
- Caching: remember the first computation

Approximate inference

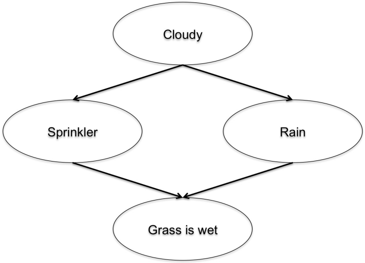
- Exact methods can require a long time in some problems
- Often, the relation among values is more important than the exact values: $P(e) > P(e')$?

Approximate inference

- Exact methods can require a long time in some problems
- Often, the relation among values is more important than the exact values: $P(e) > P(e')$?
- Approximate inference
  - goal: obtain an estimate $\hat{P}(X, e)$ of $P(X, e)$
  - idea: we can progressively improve results while we still have time

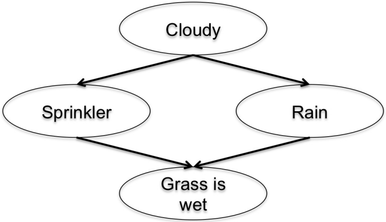
Approximate inference

- Exact methods can require a long time in some problems
- Often, the relation among values is more important than the exact values: $P(e) > P(e')$?
- Approximate inference
  - goal: obtain an estimate $\hat{P}(X, e)$ of $P(X, e)$
  - idea: we can progressively improve results while we still have time
- Direct sampling



Approximate inference. Direct sampling[1]

P(C=V)=0.5
P(C=F)=0.5



C	P(S|C)
V	.10
F	.50

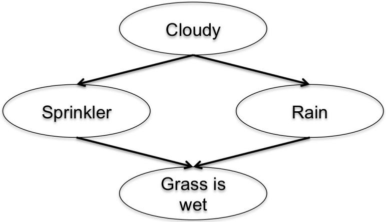
N	P(R|C)
V	.80
F	.20

S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[1] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling[2]

P(C=V)=0.5
P(C=F)=0.5



C	P(S|C)
V	.10
F	.50

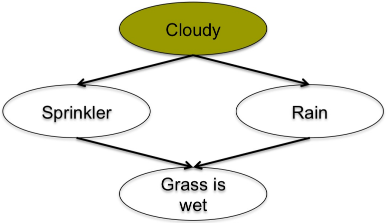
N	P(R|C)
V	.80
F	.20

S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[2] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling[3]

P(C=V)=0.5
P(C=F)=0.5



C	P(S|C)
V	.10
F	.50

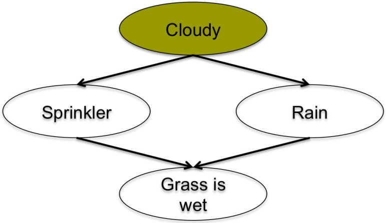
N	P(R|C)
V	.80
F	.20

S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[3] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling[4]

P(C=V)=0.5
P(C=F)=0.5



C	P(S|C)
V	.10
F	.50

N	P(R|C)
V	.80
F	.20

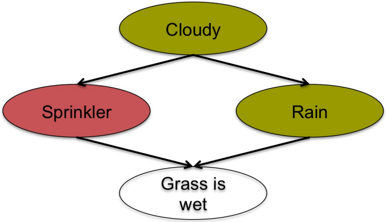
S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[4] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling[5]

P(C=V)=0.5
P(C=F)=0.5

C	P(S|C)
V	.10
F	.50



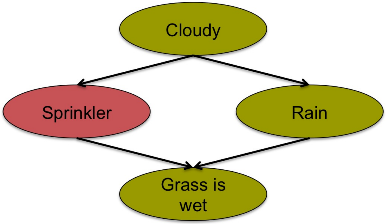
N	P(R|C)
V	.80
F	.20

S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[5] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling[6]

P(C=V)=0.5
P(C=F)=0.5



C	P(S|C)
V	.10
F	.50

N	P(R|C)
V	.80
F	.20

S	R	P(W|S,R)
V	V	.99
F	V	.90
V	F	.90
F	F	.01

[6] From [Russell and Norvig, 1995]

Approximate inference. Direct sampling

- If we repeat the previous process N times

$$\hat{P}(x_1, x_2, \ldots, x_n) = \frac{\#\text{times we saw } X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n}{N}$$

Approximate inference. Direct sampling

- If we repeat the previous process N times

$$\hat{P}(x_1, x_2, \ldots, x_n) = \frac{\#\text{times we saw } X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n}{N}$$

- And if we want to compute $P(X|e)$?

$$P(X|e) \sim \hat{P}(X|e) = \frac{\hat{P}(X, e)}{P(e)}$$

How do we generate the BN?

- We need
  - Random variables and their domain
  - Conditional independence $\Rightarrow$ Graph
  - Conditional Probability Tables

How do we generate the BN?

- We need
  - Random variables and their domain. OK
  - Conditional independence $\Rightarrow$ Graph. OK
  - Conditional Probability Tables. Not so easy

How do we generate the BN?

- We need
  - Random variables and their domain. OK
  - Conditional independence $\Rightarrow$ Graph. OK
  - Conditional Probability Tables. Not so easy
- We can
  - ask an expert
  - learn from data (machine learning)

How do we generate the BN?

- We need
  - Random variables and their domain. OK
  - Conditional independence $\Rightarrow$ Graph. OK
  - Conditional Probability Tables. Not so easy
- We can
  - ask an expert
  - learn from data (machine learning)
- We can also learn the structure

Outline

Summary Bayesian networks

- Representation of the joint probability distribution
- Show graphically the conditional (in)dependencies
- Save parameters and thus more efficient inference
- Very successful
  - decision support systems
  - documents classification, image understanding, ...
  - bioinformatics
  - man-machine interaction (e.g. Microsoft Research)
- Several commercial tools: Hugin, Netica, Agena Risk

Credits

- Material from previous years at UC3M
- Book and teaching materal of *Artificial Intelligence: A Modern Approach*. Russell&Novig. 2nd edition

References

- Jensen. An Introduction to Bayesian Networks
- Yang, X. Probabilistic Reasoning in MultiAgent Systms. A graphical models approach
- Pearl, J. Probabilistic reasoning in intelligent systems: networks of plausible inference

Outline

Outline

References

Stuart Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 1995.