Ragionamento probabilistico: rappresentazione

Marco Piastra

Ragionamento probabilistico: rappresentazione

Mondi possibili, sottoinsiemi, eventi

Variabili aleatorie

Probabilità

Marginalizzazione

Condizionalizzazione

Indipendenza, indipendenza condizionale

Modelli grafici

Eventi come sottoinsiemi di mondi possibili

- Fbf e insiemi di *mondi possibili*

  Si consideri un linguaggio logico L (p.es. del primo ordine)

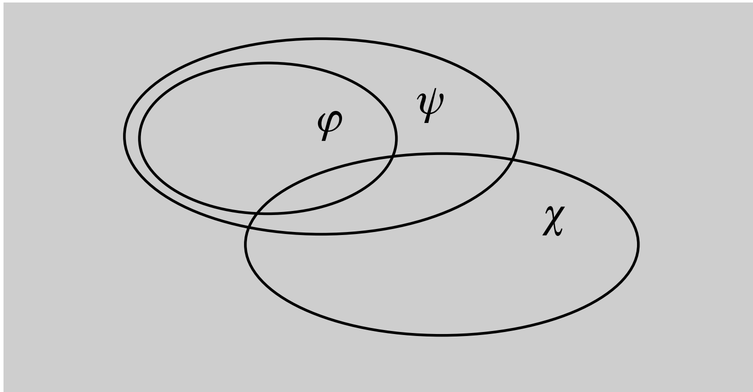
  A ciascuna fbf (chiusa) φ di L corrisponde un sottoinsieme di tutte le possibili strutture semantiche che soddisfano φ

  Vale a dire, a ciascuna fbf (chiusa) φ corrisponde $\{<\mathrm{U}, v> : <\mathrm{U}, v> \models \varphi\}$

  (assumiamo per semplicità di mantenere fisso U)

  Ciascuna struttura semantica $<\mathrm{U}, v>$ rappresenta un *mondo possibile*

  Quindi a ciascuna a ciascuna fbf (chiusa) φ corrisponde un insieme di *mondi possibili*



Intuitivamente

Un **evento** può esser visto come un sottoinsieme di *mondi possibili:* un evento si **verifica** quando il *mondo attuale* appartiene al corrispondente sottoinsieme

L'agente usa le *descrizioni* (fbf) degli *eventi* e non sa qual'è il mondo attuale

Possibilità

- Conoscenze oggettive e fbf possibili

  L'agente possiede un sistema di conoscenze oggettive

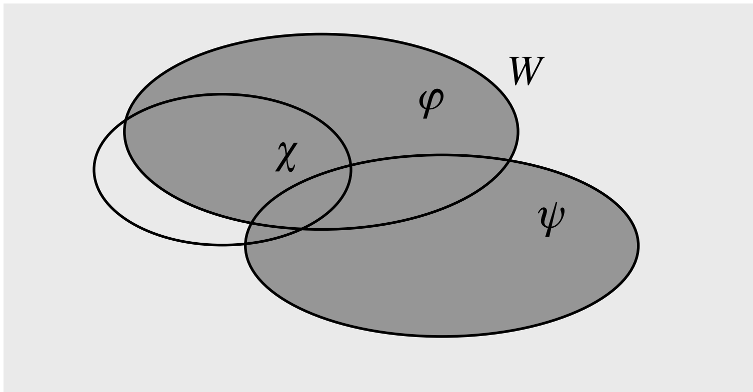
  Detta Γ la teoria che rappresenta le conoscenze dell'agente,
  l'insieme dei *mondi possibili* (per l'agente) è $W \equiv \{<\mathrm{U}, v> : <\mathrm{U}, v> \models \Gamma\}$

  P.es. l'agente sa che $\varphi \vee \psi$
  quindi solo i mondi $\{<\mathrm{U}, v> : <\mathrm{U}, v> \models \varphi \vee \psi\}$ sono *possibili* (per l'agente)
  Vuol dire che l'evento $\varphi \vee \psi$ si è già *verificato?*

  Viceversa il valore di verità di una fbf χ
  potrebbe non essere noto (all'agente):

  $\varphi \vee \psi \not\models \chi$

  $\varphi \vee \psi \not\models \neg\chi$



  Conoscenze oggettive significa che l'agente
  è sicuro che il mondo attuale sia incluso in W

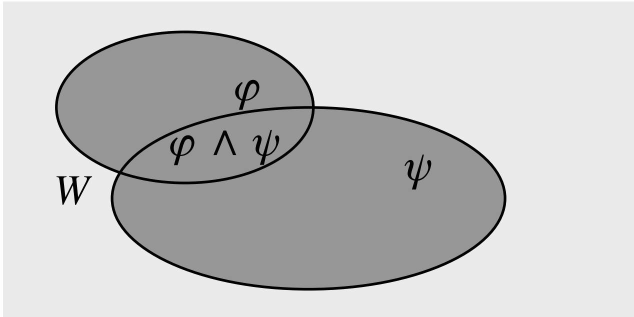
Probabilità*

(*) Nel caso di W discreto

- Una misura dei sottoinsiemi di W

  $W \equiv \{<\mathrm{U}, v> : <\mathrm{U}, v> \models \Gamma\}$ dove Γ sono le conoscenze dell'agente

  $P(.)$ è una *funzione* che assegna un numero reale agli elementi di una σ-algebra Σ formata da sottoinsiemi di W



σ-algebra

Una collezione di sottoinsiemi Σ di un insieme W per cui valgono le seguenti proprietà:

1) Σ non è vuota
2) Se $\varphi \in \Sigma$ allora $\neg\varphi \in \Sigma$
   ($\neg\varphi$ inteso come *complemento* rispetto a W)
3) Per qualsiasi collezione numerabile $\{\varphi_i\}$, $\varphi_i \in \Sigma$, si ha $\bigcup_i \varphi_i \in \Sigma$

Corollario:
Gli insiemi $\emptyset$ e W appartengono a qualsiasi σ-algebra generata su W

Gli elementi della σ-algebra sono gli **eventi**

Probabilità*

(*) Nel caso di W discreto

- Una misura dei sottoinsiemi di W

  $W \equiv \{<\mathrm{U}, v> : <\mathrm{U}, v> \models \Gamma\}$ dove Γ sono le conoscenze dell'agente

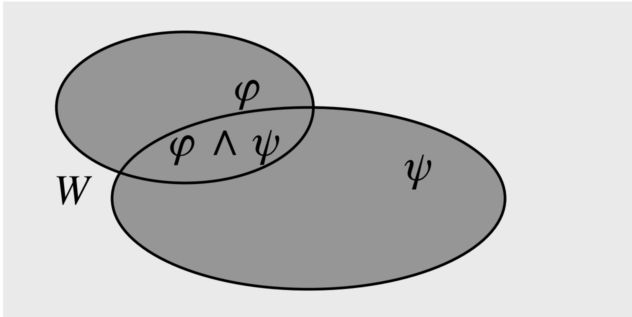
  $P(.)$ è una *funzione* che assegna un numero reale agli elementi di una σ-algebra Σ formata da sottoinsiemi di W

  Caso particolare:

  σ-algebra generata su W dalle fbf di L

  I sottoinsiemi $\{<\mathrm{U}, v> : <\mathrm{U}, v> \models \varphi\} \cap W$ che corrispondono alle fbf φ di L formano un'algebra di Boole su W tramite le operazioni di unione e complemento (vedi IA1)

  Qualsiasi algebra di Boole è anche una σ-algebra



  Gli **eventi** di questa σ-algebra sono insiemi di *mondi possibili*

  Più precisamente, sono i sottoinsiemi di W che corrispondono alle fbf di L

Probabilità*

(*) Nel caso di W discreto

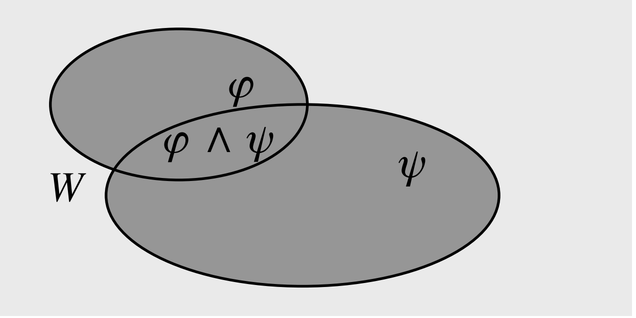
- Una misura dei sottoinsiemi di W

  $W \equiv \{<\mathrm{U}, v> : <\mathrm{U}, v> \models \Gamma\}$ dove Γ sono le conoscenze dell'agente

  $P(.)$ è una *funzione* che assegna un numero reale
  agli elementi di una σ-algebra Σ formata da sottoinsiemi di W

  $P(.)$ è una *misura* della σ-algebra Σ

  1) Per qualsiasi evento $\varphi \in \Sigma$, $P(\varphi) \geq 0$
  2) $P(W) = 1$
  3) Per qualsiasi sequenza numerabile φ_i di eventi *disgiunti* di Σ (*disgiunti* $\Leftrightarrow \varphi_i \cap \varphi_j \equiv \emptyset$ se $i \neq j$) si ha $P(\varphi_1 \vee \varphi_2 \vee \ldots \vee \varphi_n) = \sum_i P(\varphi_i)$



  Corollario:
  Per qualsiasi *evento* $\varphi \in \Sigma$, si ha $0 \leq P(\varphi) \leq 1$

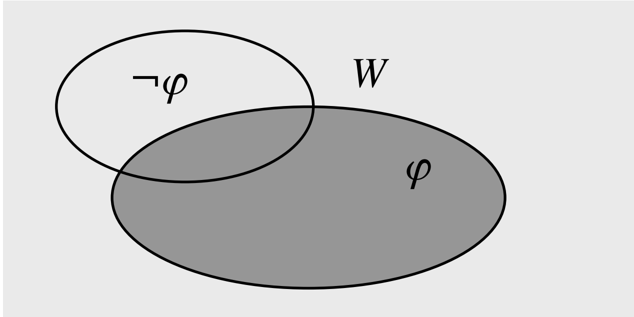
Partizioni, variabile aleatoria*

(*) Nel caso di W discreto

- Partizione

  Ciascuna fbf (chiusa) φ suddivide W
  in due sottoinsiemi disgiunti, φ e $\neg\varphi$

  (Quindi $P(\varphi) + P(\neg\varphi) = \mathrm{P}(W) = 1$, da cui $P(\neg\varphi) = 1 - P(\varphi)$)



- Variabile aleatoria

  Si consideri una variabile X che ha $\{x_1, x_2, \ldots , x_n\}$ come *dominio*

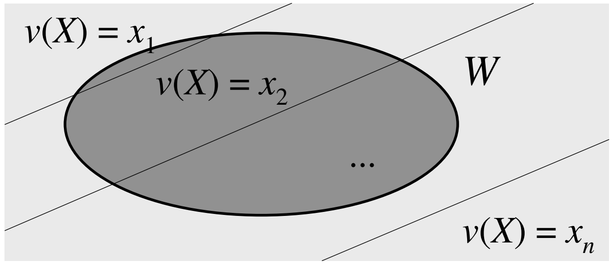
  In ciascun mondo possibile X assume un determinato valore x_i

  I possibili valori $v(X) = x_1$, $v(X) = x_2$, ..., $v(X) = x_n$ definiscono una *partizione* di W in base ad X

  - X è una variabile aleatoria
  - Ciascun $v(X) = x_i$ è un evento (un sottoinsieme di W)

  (Anche φ può essere vista come una variabile aleatoria)

  Le v.a. binarie o *binomiali* sono anche dette *bernoulliane*
  Le v.a. a più valori sono dette *multinomiali*



(*) Nel caso di W discreto

Variabili aleatorie, distribuzione congiunta*

Essendo $X=x_i$ e $X=x_j$ eventi disgiunti: $P(X=x_i \vee X=x_j) = P(X=x_i) + P(X=x_j)$ se $i \neq j$

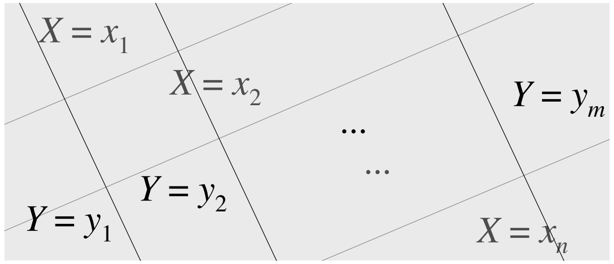
- Variabili aleatorie multiple

  Solitamente, in una rappresentazione probabilistica convivono più variabili aleatorie

  Esempi:
  X_i occorrenza in un'email di una parola i
  Y classificazione della stessa email come spam

  Ciascuna combinazione di valori delle v.a. è un *evento*

  Un'insieme di v.a. definisce una partizione di V



- Distribuzione di probabilità congiunta (*joint probability distribution*)

  Per un determinato insieme di variabili aleatorie, p.es. X, Y, Z

  E` una funzione $P(X=x_i \wedge Y=y_j \wedge Z=z_k)$ che associa un numero reale a ciascuna combinazione di valori $<x_i, y_j, z_k>$

  Si indica anche con $P(X=x_i, Y=y_j, Z=z_k)$ oppure $P(X, Y, Z)$

  Dato che X, Y e Z definiscono una partizione di V: $\sum_i \sum_j \sum_k P(X=x_i, Y=y_j, Z=z_k) = 1$

Marginalizzazione

L'eliminazione di una variabile aleatoria da una probabilità congiunta

Data una probabilità congiunta

$P(X=x_i, Y=y_j, Z=z_k)$

La *probabilità marginale* $P(X=x_i, Y=y_j)$ si ottiene per sommatoria:

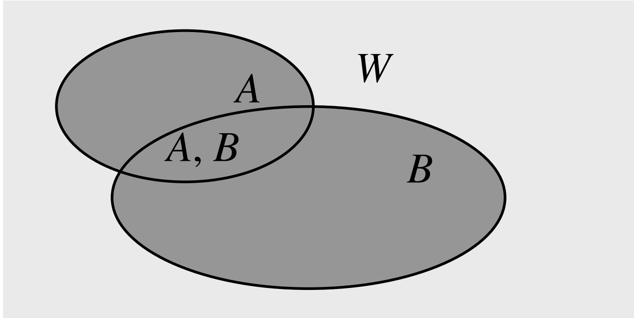
$$P(X=x_i, Y=y_j) = \sum_k P(X=x_i, Y=y_j, Z=z_k)$$

Data una probabilità congiunta su una partizione,
si può sempre ottenere una probabilità congiunta su una partizione contenuta nella prima

Probabilità condizionale

- Definizione

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



- Significato

E` una forma di *inferenza*: si passa da un'insieme di mondi possibili ad un altro

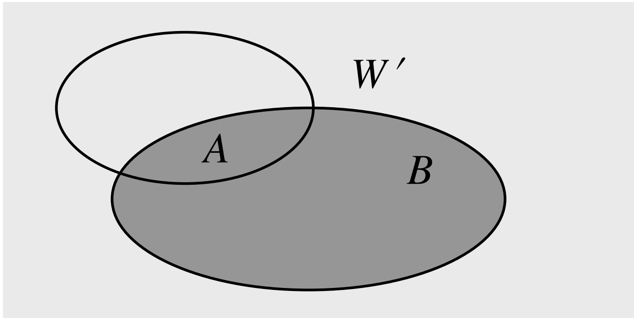
Quindi, da una misura di probabilità ad un'altra

Si assuma un agente consideri W come insieme di mondi possibili

$P(A)$ è la probabilità che A si verifichi

Si supponga che l'agente venga a sapere che l'evento B si è verificato

L'evento complementare $\neg B$ è quindi *impossibile*



$W' \equiv B$ è il nuovo insieme dei mondi possibili

$P(A \mid B)$ è la nuova probabilità che l'evento A si verifichi

Esempio: distribuzione congiunta

(*Vedi anche DutchBook.xls)

A	B	C	$P(A, B, C)$
0	0	0	0.10
0	0	1	0.12
0	1	0	0.35
0	1	1	0.08
1	0	0	0.01
1	0	1	0.02
1	1	0	0.23
1	1	1	0.09

La conoscenza della distribuzione di probabilità congiunta permette di stabilire la probabilità di qualsiasi combinazione logica di eventi

Esempi:

$$P(A \vee C) = \sum_{B} P(A \vee C,\ B) = 0.55$$

$(0.55 \cdot x)$ dovrebbe essere la somma che siete disposti a scommettere per una vincita x

$$P(\neg A \wedge \neg B) = \sum_{C} P(\neg A \wedge \neg B,\ C) = 0.22$$

Esempio: probabilità condizionale

La conoscenza della distribuzione
di probabilità congiunta
permette di stabilire
qualsiasi probabilità condizionale

A	B	C	$P(A, B, C)$
0	0	0	0.10
0	0	1	0.12
0	1	0	0.35
0	1	1	0.08
1	0	0	0.01
1	0	1	0.02
1	1	0	0.23
1	1	1	0.09

Esempio:

$$P(A \vee C \mid B=1) = \frac{P(A \vee C,\ B=1)}{P(B=1)} = \frac{0.40}{0.75} = 0.53$$

$P(A \vee B)$ era 0.55:
la conoscenza $B=1$ diminuisce, in questo caso, il valore della scommessa al totalizzatore

Teorema di Bayes (T. Bayes, 1764)

- Definizione

  Una relazione tra probabilità condizionali e marginali

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

Nelle applicazioni pratiche, $P(B \mid A)$
viene anche detta verosimiglianza *(likelihood)* $L(A \mid B)$

$$P(A \mid B) \propto L(A \mid B)\, P(A)$$

Corollario della definizione di probabilità condizionale *(chain rule)*

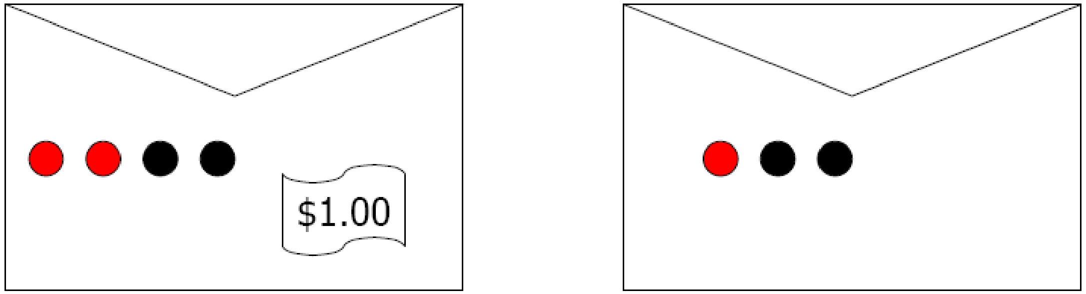
$$P(A, B) = P(B \mid A)\, P(A)$$

Per la definizione di marginalizzazione: $P(B) = \sum_A P(A, B) = \sum_A P(B \mid A)\, P(A)$

Da cui (formulazione alternativa del teorema di Bayes):

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

Esercizio: informazioni e scommesse



- Due buste, una viene estratta

  Una busta contiene due gettoni rossi e due neri, vale 1 euro
  Una busta contiene un gettone rosso e due neri, non vale nulla

  La busta è stata estratta.

  Prima di scommettere, potete estrarre un gettone

  a) Il gettone è nero. Quanto scommettete?

  b) Il gettone è rosso. Quanto scommettete?

Obiettivo: mostrare che il teorema di Bayes semplifica la rappresentazione e i calcoli

Indipendenza, indipendenza condizionale

- Indipendenza (anche detta indipendenza *marginale*)

  Due eventi sono indipendenti se la probabilità congiunta è uguale al prodotto delle probabilità marginali

  $$\langle A \perp B \rangle \quad \Rightarrow \quad P(A, B) = P(A)\, P(B)$$

- Indipendenza condizionale

  Due eventi sono condizionalmente indipendenti (dato un terzo evento) se la probabilità condizionale congiunta è uguale al prodotto delle probabilità condizionali marginali

  $$\langle A \perp B \mid C \rangle \quad \Rightarrow \quad P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

  $$\Rightarrow \quad P(A \mid B, C) = \frac{P(A, B \mid C)}{P(B \mid C)} = \frac{P(A \mid C)\, P(B \mid C)}{P(B \mid C)} = P(A \mid C)$$

  Questa è la proprietà più rilevante

Modelli grafici (anche *Bayesian Networks*)

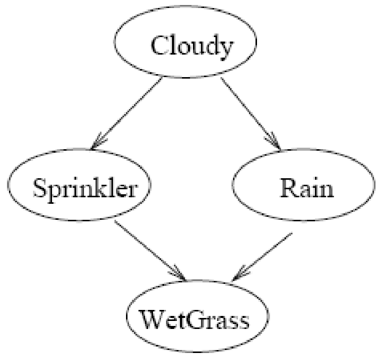
Struttura + numeri, invece di soli numeri

- Un modo per rappresentare
  una distribuzione di probabilità congiunta

  I nodi sono variabili aleatorie

  Gli archi (orientati) rappresentano *dipendenza*

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Notare che la specifica di una distribuzione congiunta di quattro v.a. richiederebbe $2^4 = 16$ valori

In figura i valori sono solo 9

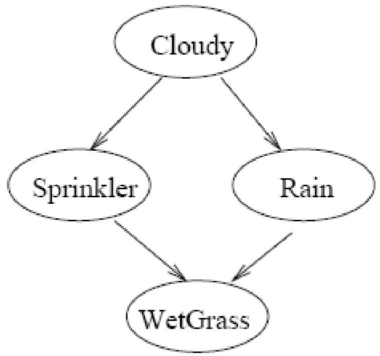
Da un modello grafico alla probabilità congiunta

- Distribuzione congiunta

  Può essere espressa come prodotto di probabilità condizionali

  (estensione della *chain rule*)

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

  Esempio:

$$P(C, S, R, W) = P(C)P(S \mid C)P(R \mid S, C)P(W \mid R, S, C)$$

  In un modello grafico, la distribuzione congiunta è un prodotto delle probabilità condizionali dei nodi

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

  Dove $parents(X_i)$ sono i nodi afferenti (diretti) del grafo orientato

  Nell'esempio:

$$P(C, S, R, W) = P(C)P(S \mid C)P(R \mid C)P(W \mid R, S)$$

  Assunzioni implicite: $<R \perp S \mid C>$, $<W \perp C \mid R, S>$

Modello grafico e indipendenze condizionali

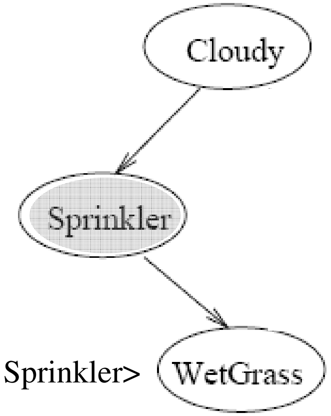
- *D-separation (Dependency-separation)*

    Come si 'legge' l'indipendenza condizionale in un modello grafico

  In un modello grafico

    Due nodi X e Y sono condizionalmente indipendenti dato un insieme di nodi $\{Z_k\}$ se tutti i percorsi tra X e Y sono bloccati

      Nel determinare i possibili percorsi tra due nodi, si ignora il verso degli archi

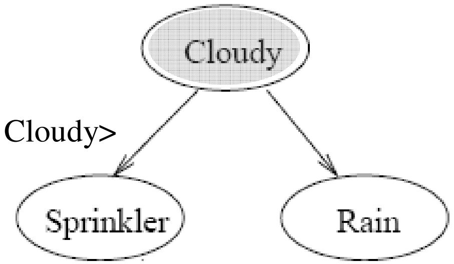


<WetGrass ⊥ Cloudy | Sprinkler>

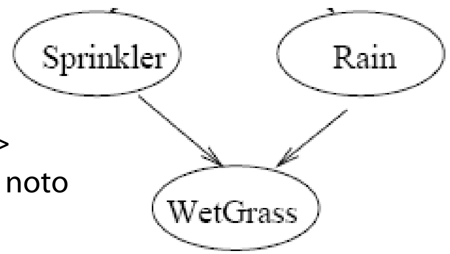
  Un percorso tra X e Y è bloccato se:

  1) Il percorso contiene una sequenza $X \rightarrow Z_i \rightarrow Y$ oppure una diramazione (*fork*) $X \leftarrow Z_i \rightarrow Y$ $(Z_i \in \{Z_k\})$

  2) Il percorso contiene una confluenza (*join*) $X \rightarrow N \leftarrow Y$ in cui N e tutti i discendenti di N <u>non</u> appartengono a $\{Z_k\}$



<Sprinkler ⊥ Rain | Cloudy>



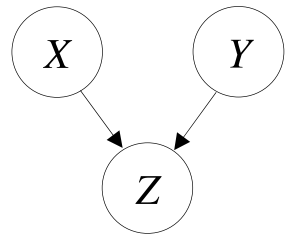
<Sprinkler ⊥ Rain>
se WetGrass non è noto

Explaining Away

Ulteriori osservazioni sulla condizione 2) della *D-separation*

Modello grafico con un *join*

Probabilità congiunta, dal grafo:



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

Probabilità marginale rispetto a X e Y (valore di Z incognito):

$$P(X, Y) = P(X)P(Y)\sum_{Z} P(Z|X,Y) = P(X)P(Y)$$

Quindi X e Y sono *marginalmente indipendenti*

Ma se il valore di Z è noto, allora X e Y sono <u>*dipendenti*</u>:

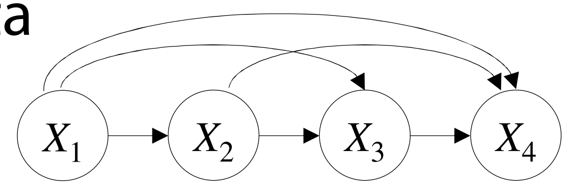
$$P(X, Y \mid Z=v) = \frac{P(X,Y,Z=v)}{P(Z=v)} = \frac{P(X)P(Y)P(Z=v|X,Y)}{\sum_{X,Y} P(X)P(Y)P(Z=v|X,Y)}$$

Non è un paradosso.

Esempio:
X e Y sono due lanci della stessa moneta, Z=1 se il risultato è lo stesso, Z=0 altrimenti.

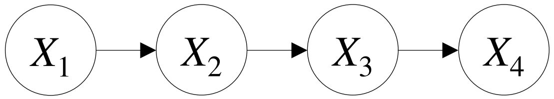
Esempi di modelli grafici

- Dipendenza completa



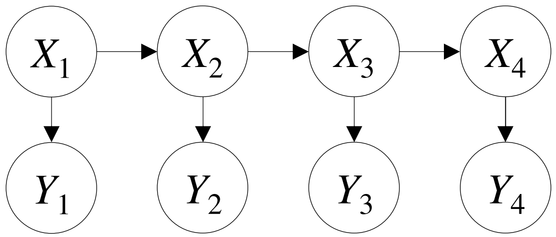
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)P(X_4 \mid X_1, X_2, X_3)$$

- Modello di Markov



$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1})$$

- Modello 'Hidden Markov'



In genere, i nodi X_i sono *hidden*, nel senso di *non-osservabili*

$$P(X_1, X_2, X_3, X_4, Y_1, Y_2, Y_3, Y_4) = P(X_1)P(Y_1 \mid X_1)P(X_2 \mid X_1)P(Y_2 \mid X_2)P(X_3 \mid X_2)P(Y_3 \mid X_3)P(X_4 \mid X_3)P(Y_4 \mid X_4)$$
$$= P(X_1)P(Y_1 \mid X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1})P(Y_i \mid X_i)$$