Network Bayesiani

selezione del modello

Marco Scutari

marco.scutari@stat.unipd.it
Dipartimento di Scienze Statistiche
Università di Padova

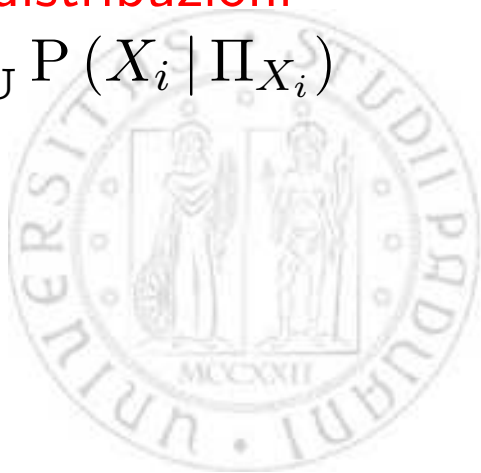
4 novembre 2008

Network bayesiani

Il modello

Un network bayesiano $BN = (\mathcal{G}, \mathrm{P})$ è costituito da:

- un grafo orientato aciclico $\mathcal{G} = (\mathbf{U}, A)$, in cui ogni nodo rappresenta una variabile aleatoria $X \in \mathbf{U}$ e gli archi orientati specificano le relazioni di dipendenza condizionale.
- una distribuzione di probabilità $\mathrm{P}(\mathbf{U})$, definita sulle stesse variabili $X \in \mathbf{U}$ e fattorizzabile in un insieme di distribuzioni di probabilità locali condizionate $\mathrm{P}(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i})$ secondo la condizione di markovianità.



Relazione tra grafo e distribuzione di probabilità

d-separazione: siano $\mathbf{X}$, $\mathbf{Y}$ e $\mathbf{Z}$ tre insiemi disgiunti di nodi in un grafo orientato aciclico; $\mathbf{X}$ e $\mathbf{Y}$ si dicono *d-separati* [15] da $\mathbf{Z}$ ($\mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$) se non esiste un cammino tra $\mathbf{X}$ e $\mathbf{Y}$ tale che:

- ogni nodo con archi convergenti appartiene a $\mathbf{Z}$ o ha un discendente che appartiene a $\mathbf{Z}$.
- qualsiasi altro nodo non appartiene a $\mathbf{Z}$.

mappe: un grafo orientato aciclico G è una *mappa di dipendenza* (*dependency map* o *d-map*) di un modello probabilistico P se

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longrightarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

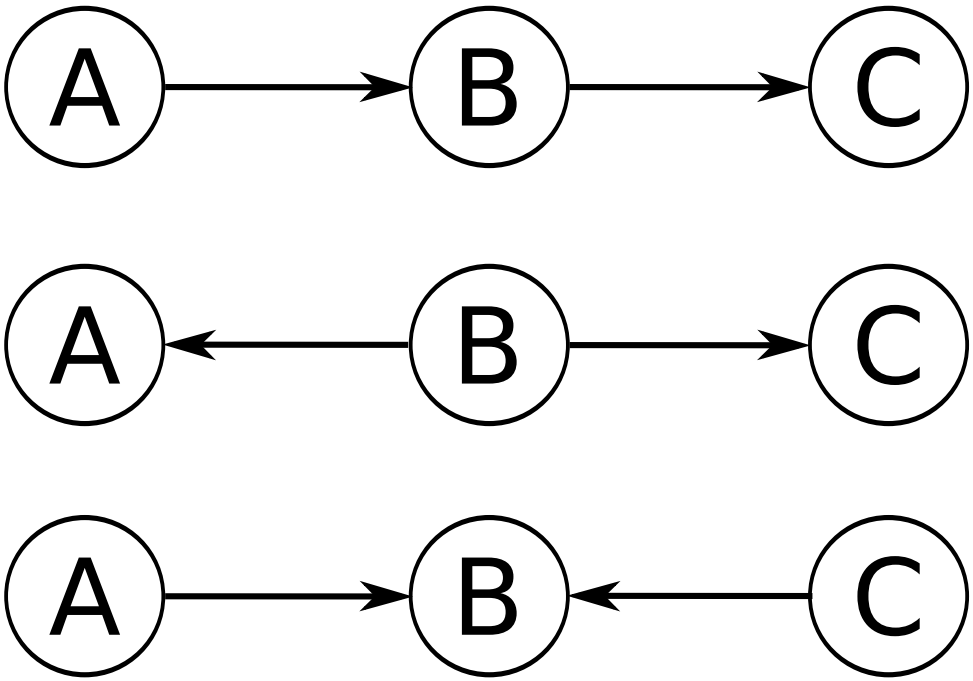
è una *mappa di indipendenza* (*independency map* o *i-map*) di P se

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longleftarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

ed è una *mappa perfetta* (*perfect map*) di P se

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longleftrightarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

Esempio: le connessioni elementari



connessione seriale

$$A \perp C \mid B$$

connessione divergente

$$A \perp C \mid B$$

connessione convergente

$$A \perp C$$
$$A \not\perp C \mid B$$

Markovianità

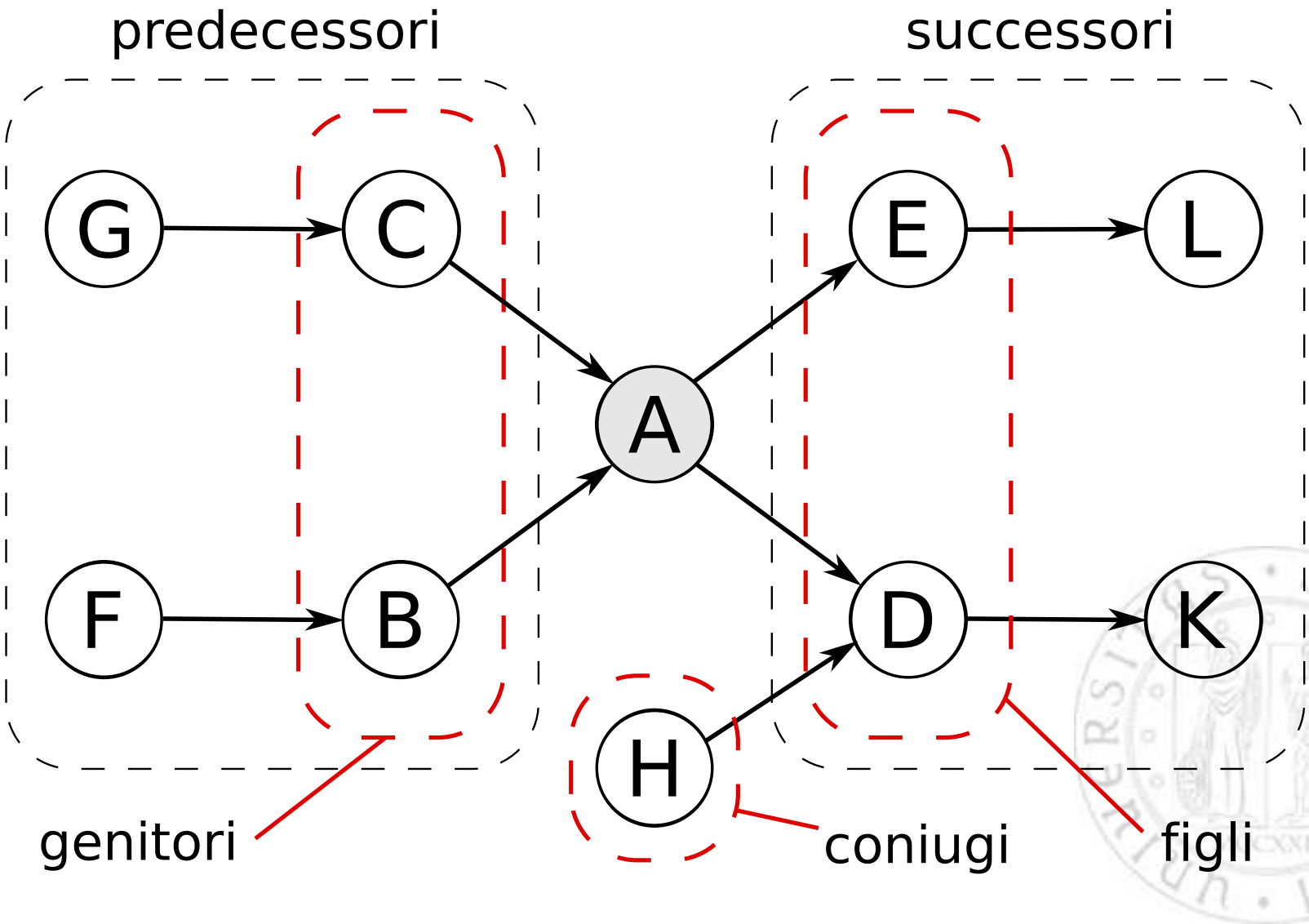
Condizione di Markov (markovianità): ogni variabile $X \in \mathbf{U}$ è indipendente dai suoi non-discendenti condizionatamente ai suoi genitori (Π_X).

Markov blanket ($Bl(X)$) di una variabile $X \in \mathbf{U}$: ogni sottoinsieme $S \subset \mathbf{U}$ di variabili per cui

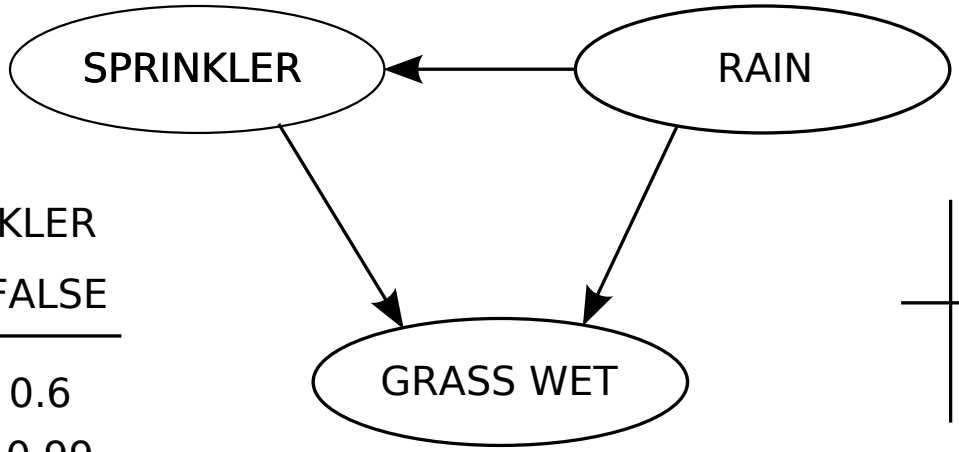
$$X \perp_G (\mathbf{U} - \mathbf{S} - X) \,|\, \mathbf{S}, \qquad X \notin \mathbf{S}$$

In ogni network bayesiano l'unione dei seguenti nodi identifica univocamente (sotto minime condizioni di regolarità) il Markov blanket minimale di X: i suoi genitori, i suoi figli ed i suoi coniugi.

Esempio: Markov blanket



Esempio: il prato di Watson



RAIN	SPRINKLER TRUE	SPRINKLER FALSE
FALSE	0.4	0.6
TRUE	0.01	0.99

RAIN TRUE	RAIN FALSE
0.2	0.8

SPRINKLER	RAIN	GRASS WET TRUE	GRASS WET FALSE
FALSE	FALSE	0.0	1.0
FALSE	TRUE	0.8	0.2
TRUE	FALSE	0.9	0.1
TRUE	TRUE	0.99	0.01

Teoria dell'informazione

Modelli probabilistici

I network bayesiani per dati discreti sono basati sulla distribuzione multinomiale

$$\mathrm{P}(x) = \frac{n!}{\prod_{i=1}^{k} n_k!} \pi_k^{n_k}, \quad \sum_{i=1}^{k} n_k = n, \prod_{i=1}^{k} \pi_k = 1$$

mentre quelli per dati continui sono basati sulla distribuzione gaussiana (normale) multivariata

$$f(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}, \quad \Sigma > 0$$

per via della trattabilità algebrica e probabilistica [12] di queste due distribuzioni multivariate.

Entropia e informazione reciproca

- informazione di Shannon

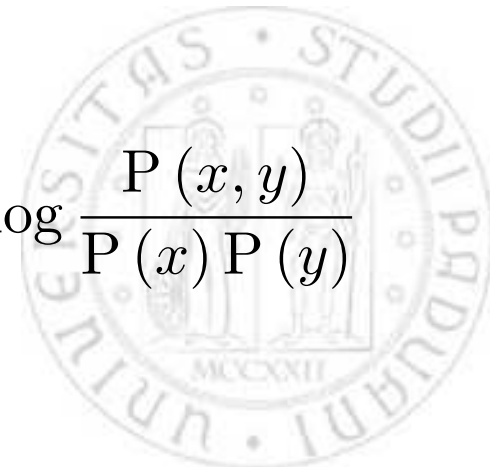
$$\mathrm{I}(X) = -\log \mathrm{P}(X = x) = -\log \mathrm{P}(x), \quad x \in \mathbb{X}$$

- entropia o *informazione attesa*

$$\mathrm{H}(X) = \mathrm{E}(-\log X) = -\sum_{x \in \mathbb{X}} \mathrm{P}(x) \log \mathrm{P}(x)$$

- informazione reciproca (*mutual information*)

$$\mathrm{MI}(X, Y) = \mathrm{E}\left(\log \frac{XY}{X \cdot Y}\right) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x, y) \log \frac{\mathrm{P}(x, y)}{\mathrm{P}(x)\,\mathrm{P}(y)}$$



Entropia differenziale e informazione reciproca

- informazione di Shannon

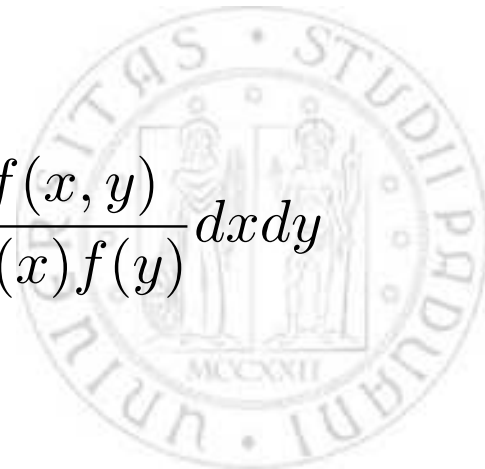
$$\mathrm{I}(X) = -\log fx, \quad x \in \mathbb{X}$$

- entropia differenziale (*differential entropy*)

$$\mathrm{H}(X) = \mathrm{E}(-\log X) = -\int f(x) \log f(x) dx$$

- informazione reciproca (*mutual information*)

$$\mathrm{MI}(X, Y) = \mathrm{E}\left(\log \frac{XY}{X \cdot Y}\right) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)} dx dy$$



Funzione di verosimiglianza

- funzione di verosimiglianza (*likelihood function*)

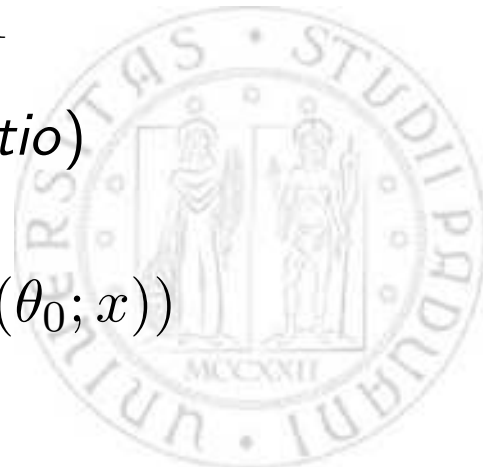
$$\mathrm{L}\left(\theta; x\right) = \prod_{i=1}^{n} \mathrm{P}\left(X = x\right) \qquad \mathrm{L}\left(\theta; x\right) = \prod_{i=1}^{n} f(x)$$

- funzione di log-verosimiglianza (*log-likelihood function*)

$$l(\theta; x) = \sum_{i=1}^{n} \log \mathrm{P}\left(X = x\right) \qquad l(\theta; x) = \sum_{i=1}^{n} \log f(x)$$

- log-rapporto di verosimiglianza (*log-likelihood ratio*)

$$\mathrm{G}^2(x) = -2 \log \frac{\mathrm{L}\left(\theta_0; x\right)}{\mathrm{L}\left(\theta_1; x\right)} = 2\left(l(\theta_1; x) - l(\theta_0; x)\right)$$



Relazione tra test statistici e teoria dell'informazione

Si può dimostrare [10] [1] che sia nel caso continuo che in quello discreto l'informazione reciproca coincide con il test log-rapporto di verosimiglianza per testare l'indipendenza tra X e Y

$$2n\,\mathrm{MI}\,(X,Y) = \mathrm{G}^2\,(X,Y) \mathrel{\dot\sim} \chi^2_r$$

ed è proporzionale o asintoticamente equivalente ai principali test statistici usati nelle verifiche di indipendenza (X^2 di Pearson, test esatto di Fisher per tabelle di contingenza, test t per la correlazione, Z di Fisher, ecc.).

Lo stesso vale per i test di indipendenza condizionale.

Relazione tra test statistici e teoria dell'informazione

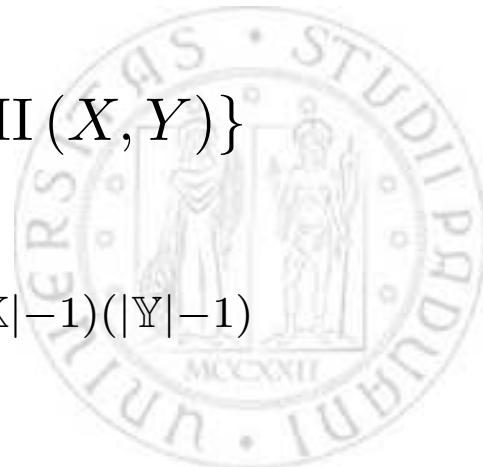
Due esempi nell'ambito dei test per dati discreti:

- X^2 di Pearson

$$X^2 = \frac{\left(\mathrm{P}\left(X = y, Y = y\right) - \mathrm{P}\left(X = x\right)\mathrm{P}\left(Y = y\right)\right)^2}{\mathrm{P}\left(X = x\right)\mathrm{P}\left(Y = y\right)}$$
$$\simeq \mathrm{MI}\left(X, Y\right) \dot{\sim} \chi^2_{(|\mathbb{X}|-1)(|\mathbb{Y}|-1)}$$

- test esatto di Fisher per tabelle di contingenza $r \times c$

$$F(n_{ij}) = \frac{\prod_{i=1}^{r} n_{i+}! \prod_{j=1}^{c} n_{+j}!}{n! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}} \simeq \exp\left\{-n\,\mathrm{MI}\left(X, Y\right)\right\}$$
$$\Rightarrow -\frac{1}{n}\log F(n_{ij}) \simeq \mathrm{MI}\left(X, Y\right) \dot{\sim} \chi^2_{(|\mathbb{X}|-1)(|\mathbb{Y}|-1)}$$



Selezione del modello

Ipotesi di lavoro

Assunzioni alla base dei metodi automatici di apprendimento della struttura di un network bayesiano:

- *markovianità* (markov assumption): ogni nodo deve essere indipendente dai suoi non-discendenti condizionatamente ai suoi genitori.
- *causalità* (causal sufficiency): non devono esistere variabili latenti o duplicate.
- *accuratezza* (faithfulness): il grafo deve essere una mappa perfetta della distribuzione di probabilità considerata (necessario per i metodi *constraint-based*).

Ulteriori assunzioni di regolarità (spesso sottintese):

- non vi devono essere dati mancanti.
- i dati devono essere tra loro indipendenti.
- non vi devono essere parametri nulli per costruzione o il cui valore si trovi sul confine del proprio dominio.

Metodi *score-based*

1. ogni modello viene classificato in base ad un punteggio che descrive la qualità della stima.
2. si individua un modello iniziale (in genere minimale, corrispondente ad un grafo senza alcun arco).
3. si esplora lo spazio dei modelli modificando un arco alla volta ed accettando un nuovo modello quando questo è significativamente migliore del precedente [7] [8]. Non è detto che esista un unico massimo assoluto; più grafi possono essere *score equivalent* [3].
4. possibili algoritmi di ricerca: *hill climbing*, *breadth first*, algoritmi genetici, tecniche euristiche [14], ecc.

Possibili funzioni punteggio

- funzione di verosimiglianza

$$Score_{ML}(\mathcal{G}, \mathcal{D}) = \log \mathrm{L}(\mathcal{D} \mid \mathcal{G})$$

- *Akaike* e *Bayesian Information Criterion*

$$Score_{AIC}(\mathcal{G}, \mathcal{D}) = \log \mathrm{L}(\mathcal{D} \mid \mathcal{G}) - d$$

$$Score_{BIC}(\mathcal{G}, \mathcal{D}) = \log \mathrm{L}(\mathcal{D} \mid \mathcal{G}) - \frac{d}{2} \log N$$

- criterio bayesiano, *high posterior density* (come ad esempio il punteggio $K2$ [4] per dati discreti)

$$Score_{Bayes}(\mathcal{G}, \mathcal{D}) = \mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{D} \mid \mathcal{G}) \mathrm{P}(\mathcal{G})$$

Esempio: algoritmo *hill climbing* (*greedy search*)

$A = \varnothing$

$\mathcal{G} = (\mathbf{U}, A)$

$score = -\infty$

do {

  $maxscore = score$

  foreach $((X, Y) : X, Y \in \mathbf{U})$ {

    foreach $(A' \in \{A \cup (X, Y), A \cup (Y, X), A - \{(X, Y), (Y, X)\}\})$ {

      $\mathcal{G}' = (\mathbf{U}, A')$

      $newscore = Score\,(\mathcal{G}', \mathcal{D})$

      if $(newscore > score)$ {

        $\mathcal{G} = \mathcal{G}'$

        $score = newscore$

      }//*then*

    }//*foreach*

  }//*foreach*

} while $(score > maxscore)$

return $\mathcal{G}$

Metodi *constraint-based*

1. le relazioni (vincoli) di indipendenza condizionale vengono esaminate singolarmente utilizzando dei test statistici.
2. si individua un grafo non orientato che le rappresenti.
3. si orientano gli archi in base al comportamento delle connessioni fondamentali. Il risultato è una *classe di equivalenza* [20] che può contenere più di un network.
4. algoritmi, tutti derivati dall'*Inductive Causation* (IC) [20]: PC [17], *Grow-Shrink* (GS) [11], *Incremental Association Markov Blanket* (IAMB) [18] e derivati [21], ecc.

Esempio: algoritmo *Inductive Causation* (IC)

1. Per ogni coppia di nodi X e Y si cerchi un insieme di nodi $\mathbf{S}_{XY}$ che li renda condizionatamente indipendenti. Se non esiste un $\mathbf{S}_{XY}$ che soddisfi questa condizione, si colleghino X e Y con un arco non orientato.
2. Per ogni coppia di nodi X e Y non adiacenti si individuino i nodi Z adiacenti ad entrambi, e si orientino gli archi in una connessione convergente $(X \rightarrow Z \leftarrow Y)$ se $Z \notin \mathbf{S}_{XY}$.
3. Si orientino iterativamente gli archi rimanenti secondo le due regole seguenti:
   - 3.1 se X e Y non sono adiacenti ma esiste un arco orientato che porta da X a Z $(X \rightarrow Z)$ ed uno non orientato tra Y e Z, si orienti quest'ultimo in direzione di Y $(Z \rightarrow Y)$ in modo da formare una connessione seriale.
   - 3.2 se esiste un cammino costituito da soli archi orientati che porta da X a Z e questi nodi sono adiacenti, si orienti l'arco che li congiunge in modo da non creare un ciclo $(X \rightarrow Z)$.
4. Se esiste un arco orientato tra X e Y $(X \rightarrow Y)$, si orienti ogni arco tra Y e Z in direzione di quest'ultimo $(Y \rightarrow Z)$ se Z non è adiacente a X.

Equivalenza del criterio decisionale

Si considerino due network bayesiani discreti i cui grafi differiscono per la presenza di un solo arco $(X_k \rightarrow X_l)$, rispettivamente con densità $\mathrm{P}(\mathbf{U})$ e $\mathrm{P}'(\mathbf{U})$:

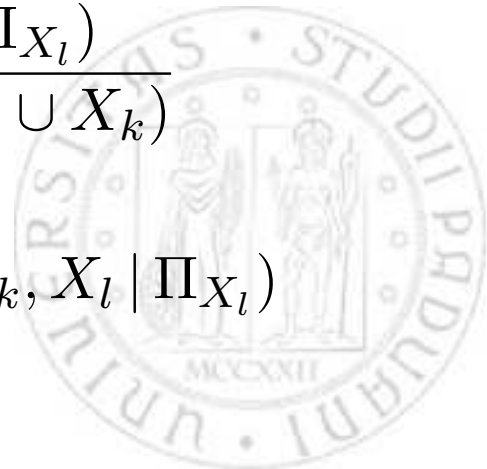
$$\mathrm{P}(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i}) \qquad \mathrm{P}'(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}\left(X_i \,|\, \Pi'_{X_i}\right)$$

sotto le ipotesi esposte in precedenza.

Equivalenza del criterio decisionale

Allora si può dimostrare che il confronto globale di due network bayesiani, effettuato utilizzando come punteggio la funzione di verosimiglianza, si riduce alla verifica locale dell'esistenza dell'arco $X_k \to X_l$ con l'informazione reciproca (condizionata):

$$
\begin{aligned}
G^2 &= -2n \log \frac{\mathrm{P}(\mathbf{U})}{\mathrm{P}'(\mathbf{U})} = -2n \log \frac{\prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i})}{\prod_{X_i \in \mathbf{U}} \mathrm{P}\left(X_i \,|\, \Pi'_{X_i}\right)} \\
&= -2n \log \frac{\mathrm{P}(X_l \,|\, \Pi_{X_l})}{\mathrm{P}\left(X_l \,|\, \Pi'_{X_l}\right)} = -2n \log \frac{\mathrm{P}(X_l \,|\, \Pi_{X_l})}{\mathrm{P}(X_l \,|\, \Pi_{X_l} \cup X_k)} \\
&= 2n \log \frac{\mathrm{P}(X_l, X_k \,|\, \Pi_{X_l})}{\mathrm{P}(X_l \,|\, \Pi_{X_l})\,\mathrm{P}(X_k \,|\, \Pi_{X_l})} = 2n\,\mathrm{MI}(X_k, X_l \,|\, \Pi_{X_l})
\end{aligned}
$$



Equivalenza del criterio decisionale

Pertanto i criteri decisionali utilizzati nelle due classi di metodi, ovvero i i punteggi per i metodi *score-based* ed i test di indipendenza condizionale per i metodi *constraint-based*, sono equivalenti:

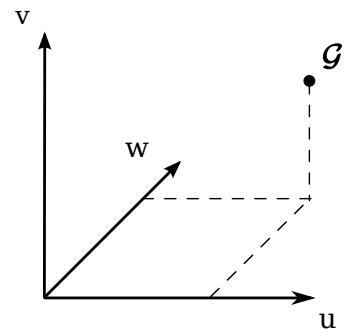
1. il tipo di confronto si riduce in entrambi i casi ad una verifica locale che coinvolge solo due variabili ed i genitori di una di esse.
2. l'indicatore utilizzato è comunque riconducibile all'informazione reciproca.

Relazione tra metodi *score* e *constraint-based*

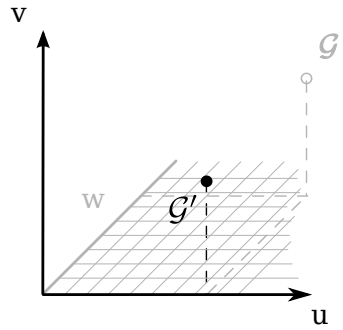
Si rappresenti lo spazio dei grafi come uno spazio n-dimensionale in cui ogni dimensione è associata ad un arco:

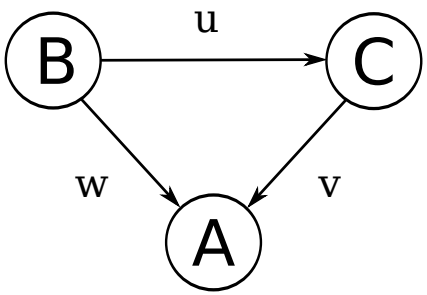
- nei metodi *score-based* ogni verifica di ipotesi si traduce in un movimento parallelo ad uno degli assi dello spazio; la procedura di selezione quindi si traduce in una sequenza di spostamenti ortogonali che, a partire da un modello iniziale, conducono ad un secondo modello il cui punteggio è più elevato.
- nei metodi *constraint-based* ogni test viene effettuato senza alcun riferimento ad un modello particolare, e le sue conclusioni sono quindi valide su tutto lo spazio dei grafi. Pertanto l'accettazione dell'ipotesi nulla corrispondente al test comporta l'eliminazione di una delle dimensioni dello spazio. Il successivo orientamento degli archi rimanenti contribuisce a restringere ulteriormente lo spazio, individuando uno o più modelli.

Relazione tra metodi *score* e *constraint-based*

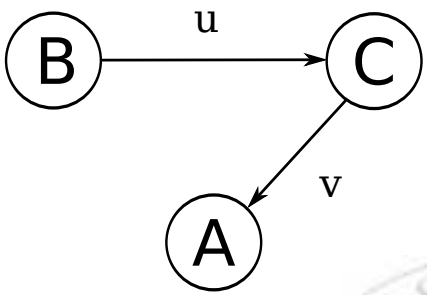


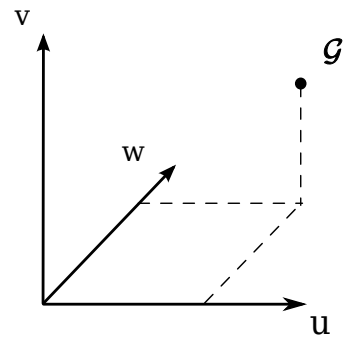
metodi
constraint–based



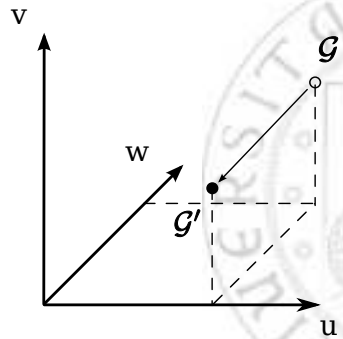


$A \perp B \,|\, C$





metodi
score–based



Statistica non parametrica

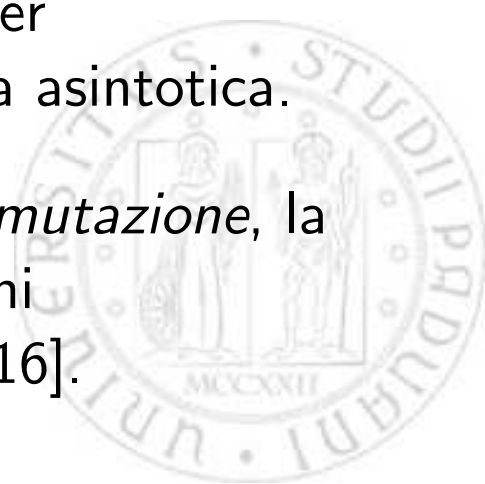
Utilizzo della distribuzione di permutazione

La correttezza delle verifiche di ipotesi utilizzate nell'apprendimento della struttura è estremamente importante, dato che ogni errore causa distorsioni cumulative in entrambe le classi di metodi.

Tuttavia:

- la natura condizionata dei test riduce di molto il numero di osservazioni disponibili per ogni modalità delle variabili in esame, dato che aumenta la dimensionalità del problema.
- la distribuzione χ^2 non è sempre soddisfacente per l'informazione reciproca, a causa della sua natura asintotica.

Pertanto è preferibile utilizzare la *distribuzione di permutazione*, la cui precisione dipende solo dal numero di permutazioni considerate, e la sua *combinazione non parametrica* [16].



Distribuzione di permutazione

Si supponga:

- di avere una verifica di ipotesi di uguaglianza in distribuzione $H_0 : X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} \ldots \stackrel{d}{=} X_C$ (che nel caso di test di indipendenza si traduce in $X_1 X_2 \stackrel{d}{=} X_1 \cdot X_2$).
- che sotto l'ipotesi nulla le osservazioni siano *scambiabili*, e possano essere permutate tra i vari gruppi. L'indipendenza implica la scambiabilità.

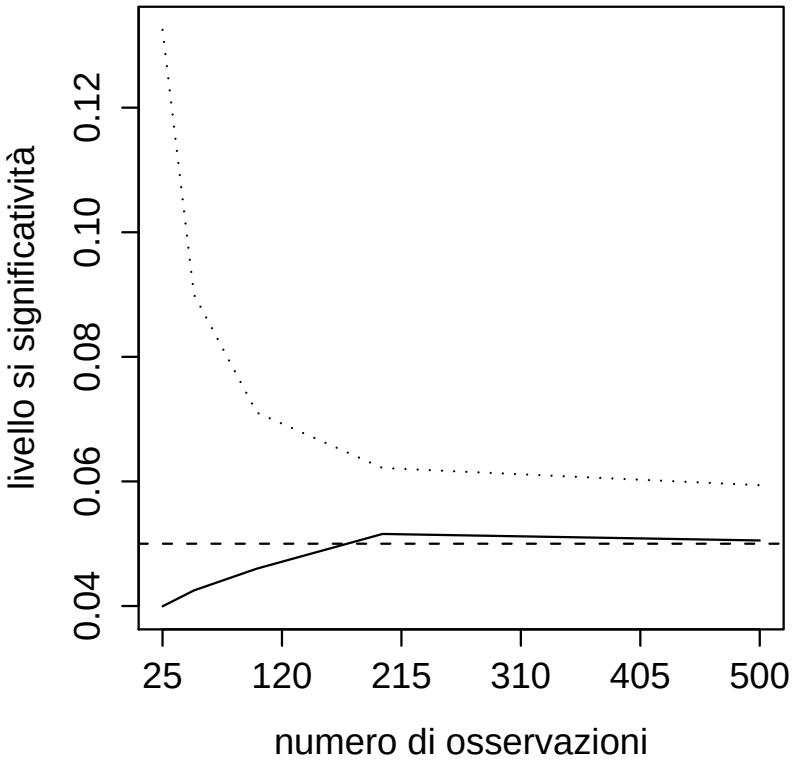
In questo modo si ottiene uno *spazio delle permutazioni* $\mathcal{X}^*$, condizionato ai dati osservati su cui valutare

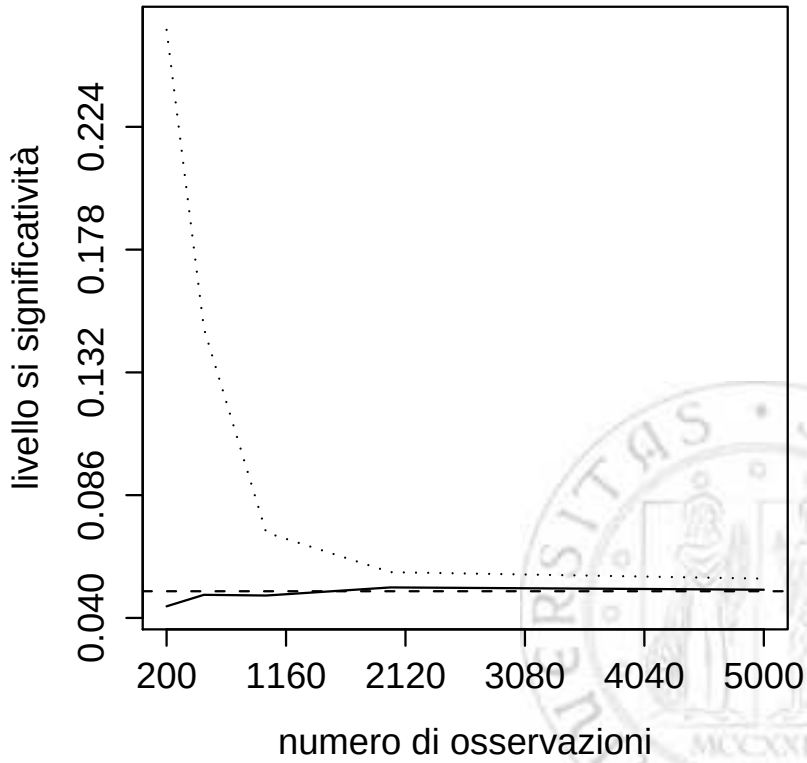
$$\lambda = \mathrm{P}\left(T(X^*) \geqslant T(X) \,|\, X\right), \qquad X^* \in \mathcal{X}^*$$

senza dover fare alcuna assunzione sulla distribuzione della statistica test T.

Livello di significatività osservata

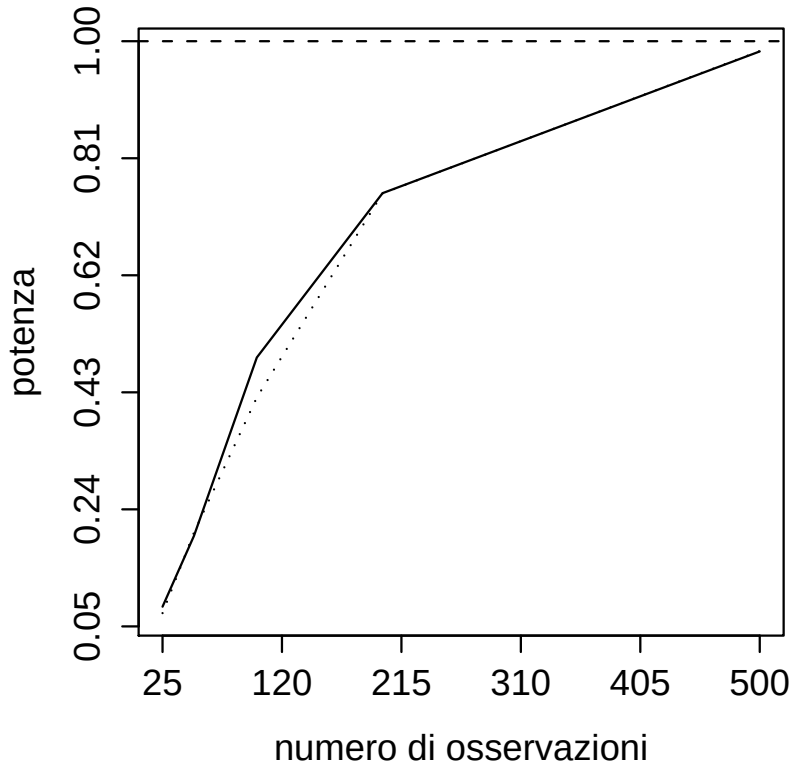
significatività = probabilità di rifiutare erroneamente l'ipotesi di indipendenza

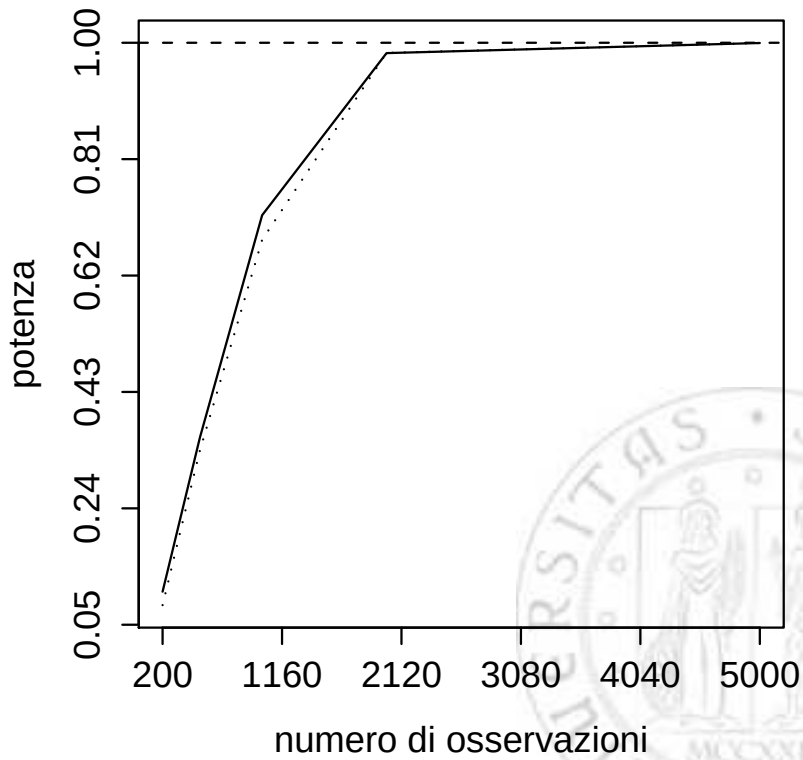




Livello di potenza osservata

potenza = probabilità di rifiutare correttamente l'ipotesi di indipendenza





Combinazione non parametrica

Si supponga di avere una verifica di ipotesi di uguaglianza in distribuzione rappresentabile come un insieme di ipotesi parziali:

$$H_0 : \bigcap_{i=1}^{k} \{H_{0_i}\} \qquad H_1 : \bigcup_{i=1}^{k} \{H_{1_i}\} \qquad i = 1, \ldots, k$$

In questo caso è possibile utilizzare gli indicatori parziali $T_i(\mathbf{X})$, calcolati simultaneamente su ogni permutazione dei dati, per costruire un unico indicatore $\mathbf{T}''(\mathbf{X})$ per l'ipotesi globale tramite una opportuna *funzione di combinazione* ψ dei loro livelli di significatività λ_i.

Applicazione per dati discreti

Ad esempio un'ipotesi di indipendenza condizionale

$$H_0 : X_l \perp X_k \,|\, \Pi_{X_l} \qquad H_1 : X_l \not\perp X_k \,|\, \Pi_{X_l}$$

per dati discreti può essere scomposta in un insieme di ipotesi parziali corrispondenti alle varie configurazioni di Π_{X_l}

$$H_0 : \bigcap_{i=1}^{|\Pi_{X_l}|} \{X_{l_i} \perp X_{k_i}\} \qquad H_1 : \bigcup_{i=1}^{|\Pi_{X_l}|} \{X_{l_i} \not\perp X_{k_i}\}$$

che sono equivalenti ad altrettante ipotesi di uguaglianza in distribuzione:

$$H_0 : \bigcap_{i=1}^{|\Pi_{X_l}|} \left\{X_{l_i} X_{k_i} \stackrel{d}{=} X_{l_i} \cdot X_{k_i}\right\} \quad H_1 : \bigcup_{i=1}^{|\Pi_{X_l}|} \left\{X_{l_i} X_{k_i} \stackrel{d}{\neq} X_{l_i} \cdot X_{k_i}\right\}$$

e possono quindi essere studiate tramite la distribuzione di permutazione.

Sviluppi recenti

Metodi *Sparse Candidate*

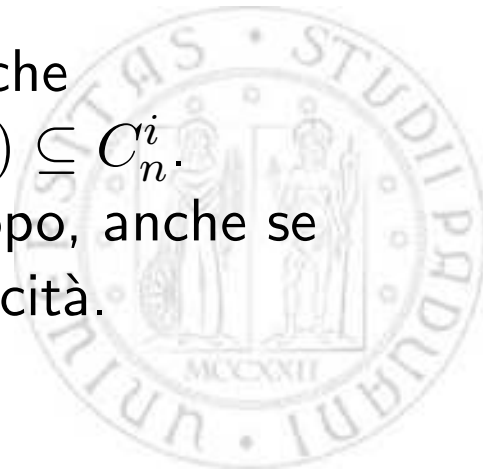
La classe di metodi *sparse candidate* [6], la cui implementazione più recente è data dal *Min-Max Hill Climbing* (MMHC) [19], combina i metodi *score* e *constraint-based* per ottenere una convergenza più veloce ed una maggiore robustezza:

- i test di indipendenza condizionale tipici dei metodi *constraint-based* vengono usati per ridurre rapidamente lo spazio dei network bayesiani da valutare.
- la massimizzazione di un punteggio (come nei metodi *score-based*) viene effettuata in un sottoinsieme ridotto dello spazio dei network bayesiani, ed è quindi meno probabile che rimanga bloccata in un massimo locale.

Algoritmo *Sparse Candidate*

Si ripetano i due passi seguenti per $n = 1, 2, \ldots$ volte, fino alla convergenza.

1. **Restrict:** sulla base del network bayesiano $\mathcal{B}_{n-1}$ si selezioni per ogni variabile X_i un insieme C_n^i di possibili genitori attraverso dei test di indipendenza condizionale. Questo corrisponde al primo punto dell'algoritmo *Inductive Causation*, ed ai punti corrispondenti di tutti gli algoritmi *constraint-based* ad esso ispirati.
2. **Maximize:** si individui il network bayesiano $\mathcal{B}_n$ che massimizza $Score\,(\mathcal{B}, \mathcal{D})$ tra quelli in cui $Pa(X_i) \subseteq C_n^i$. Qualsiasi algoritmo *score-based* è adatto allo scopo, anche se l'*hill climbing* è preferito per via della sua semplicità.



Bibliografia I


A. Agresti.
Categorical Data Analysis.
John Wiley & Sons, Inc., 2002.


A. Azzalini.
Inferenza Statistica.
Springer, 2001.


D. M. Chickering.
A transformational characterization of equivalent bayesian network structures.
In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann Publishers Inc., 1995.


G. F. Cooper and E. Herskovits.
A bayesian method for the induction of probabilistic networks from data.
Machine Learning, 9(4):309–347, 1992.


T. A. Cover and J. A. Thomas.
Elements of Information Theory.
John Wiley & Sons, Inc., 2006.

Bibliografia II


N. Friedman, I. Nachman, and D. Peter.
Learning bayesian network structure from massive datasets: The Sparse Candidate algorithm.
In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 206–215. Morgan Kaufmann, 1999.


D. Geiger and D. Heckerman.
Learning gaussian networks.
Technical Report MSR-TR-94-10, Redmond, WA, 1994.


D. Heckerman, D. Geiger, and D. M. Chickering.
Learning bayesian networks: The combination of knowledge and statistical data.
Machine Learning, 20(3):197–243, 1995.


K. Korb and A. Nicholson.
Bayesian artificial intelligence.
Chapman and Hall, 2004.


S. Kullback.
Information theory and statistics.
John Wiley & Sons, Inc., 1959.

Bibliografia III

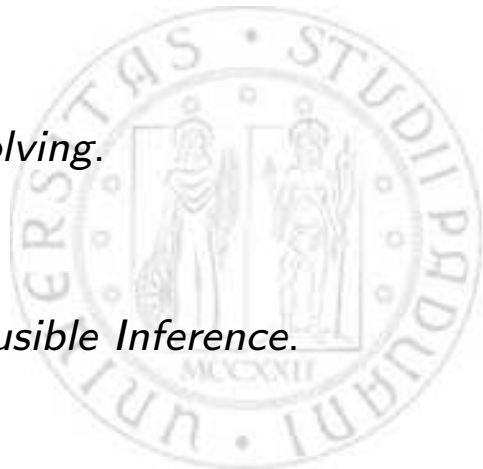

D. Margaritis.
Learning Bayesian Network Model Structure from Data.
PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, May 2003.
Available as Technical Report CMU-CS-03-153.


C. Meek.
Strong completeness and faithfulness in bayesian networks.
In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–41. Morgan Kaufmann, 1995.


L. Pace and A. Salvan.
Teoria della Statistica: Metodi, Modelli, Approssimazioni asintotiche.
Cedam, 1996.


J. Pearl.
Heuristics: intelligent search strategies for computer problem solving.
Addison-Wesley, 1984.


J. Pearl.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann Publishers Inc., 1988.



Bibliografia IV


F. Pesarin.
Multivariate Permutation Tests with Applications in Biostatistics.
John Wiley & Sons, Inc., 2001.


P. Spirtes, C. Glymour, and R. Scheines.
Causation, Prediction and Search.
MIT Press, 2001.


I. Tsamardinos, C. F. Aliferis, and A. Statnikov.
Algorithms for large scale markov blanket discovery.
In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pages 376–381. AAAI Press, 2003.


I. Tsamardinos, L. E. Brown, and C. F. Aliferis.
The Max-Min Hill-Climbing bayesian network structure learning algorithm.
Machine Learning, 65(1):31–78, October 2006.


T. S. Verma and J. Pearl.
Equivalence and synthesis of causal models.
Uncertainty in Artificial Intelligence, 6:255–268, 1991.

Bibliografia V


S. Yaramakala.
Fast Markov Blanket Discovery.
PhD thesis, Iowa State University, 2004.