



Università degli Studi di Milano–Bicocca
Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Magistrale in Informatica

Tesi di laurea magistrale

Continuous time Bayesian Network Classifiers

Sottotitolo

Candidato:
Leonardo Di Donato
Matricola 744739

Relatore:
Prof. F. Antonio Stella

Correlatore:
Dott. Daniele Codecasa

Anno Accademico 2012–2013

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

— Oscar Wilde

Dedicato a tutti gli appassionati di L^AT_EX.

INDICE

1	CONTINUOUS TIME BAYESIAN NETWORK	1
1.1	Fondamenti	1
1.1.1	Bayesian Network	1
1.1.2	Processi di Markov	6
1.2	Definizioni preliminari	10
1.3	Rappresentazione	11
1.4	Apprendimento	12
1.4.1	Statistiche sufficienti	13
1.4.2	Likelihood	13
1.4.3	Stima dei parametri	15
2	CLASSIFICAZIONE	19
2.1	Modello	19
2.2	Apprendimento	22
2.3	Inferenza	26
3	APPRENDIMENTO STRUTTURALE	31
3.1	Funzione di scoring	32
3.2	Ricerca della struttura	33
3.2.1	Hill Climbing	33
4	PACKAGE R	34
4.1	Analisi	34
4.2	Package CTBN	34
5	CREAZIONE DI DATASET RELATIVI AL TRAFFICO	36
5.1	TSIS	36
5.1.1	Descrizione	36
5.1.2	API	36
5.2	Estensione	37
5.2.1	Analisi	37
5.2.2	Sensors DLL	37
5.3	Applicativi di supporto	37
6	ESPERIMENTI NUMERICI	38
6.1	Dataset 1	38
6.1.1	Modello TSIS	38
6.1.2	Risultati	39
6.2	Dataset 2	39
6.2.1	Modello TSIS	39
6.2.2	Risultati	39

7	CONCLUSIONI	41
A	GUIDE ALL'USO	42
A.1	Utilizzo del package CTBN	42
A.1.1	Caricamento del dataset	42
A.1.2	Calcolo delle sufficient statistics	42
A.1.3	Calcolo dei parametri	42
A.1.4	Calcolo delle CIM	42
A.1.5	Apprendimento	42
A.1.6	Classificazione	42
A.1.7	Apprendimento strutturale	42
A.1.8	Cross-validation	42
A.2	Creazione di dataset	42
A.2.1	Sensors DLL	43
A.2.2	Applicativi di supporto	43
	ACRONIMI	44
	INDICE ANALITICO	45
	BIBLIOGRAFIA	46

ELENCO DELLE FIGURE

Figura 2.1	Un esempio di CTBNC	21
Figura 2.2	Un CTNBC	21
Figura 2.3	Un CTTANBC	22

ELENCO DELLE TABELLE

ELENCO DEGLI ALGORITMI

Algoritmo 2.1	Apprendimento di un classificatore CTNB . . .	23
Algoritmo 2.2	Apprendimento di un classificatore CTBN . . .	25
Algoritmo 2.3	Inferenza su un classificatore CTBN	29

SOMMARIO

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

ABSTRACT

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

*Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis.
Curabitur dictum gravida mauris.*

— Donald Ervin Knuth

RINGRAZIAMENTI

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Milano, settembre 2013

L.

INTRODUZIONE

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

IL PRIMO CAPITOLO offre una visione d'insieme della storia di \LaTeX e ne vengono presentate le idee di fondo.

IL SECONDO CAPITOLO offre una visione d'insieme della storia di \LaTeX e ne vengono presentate le idee di fondo.

IL TERZO CAPITOLO spiega le operazioni, veramente semplici, per installare \LaTeX sul proprio calcolatore.

IL QUARTO CAPITOLO descrive sinteticamente le principali norme tipografiche della lingua italiana, utili nella composizione di articoli, tesi o libri.

IL QUINTO CAPITOLO descrive sinteticamente le principali norme tipografiche della lingua italiana, utili nella composizione di articoli, tesi o libri.

IL SESTO CAPITOLO descrive sinteticamente le principali norme tipografiche della lingua italiana, utili nella composizione di articoli, tesi o libri.

In questo capitolo si introducono i concetti fondamentali relativi alle Continuous time Bayesian Network (CTBN). Le CTBN sono un framework capace di modellare processi stocastici a tempo continuo e con spazio degli stati discreto.

Prima di affrontare tale argomento si presentano alcuni concetti propedeutici a questo lavoro di tesi: le Bayesian Network (BN) e i processi di Markov (sezione 1.1).

1.1 FONDAMENTI

Le Continuous time Bayesian Network utilizzano concetti e idee provenienti da teorie afferenti l'area statistica e del machine learning. Al fine di conferire alla discussione sulle CTBN un quadro iniziale completo ed esauriente, si presentano quindi gli aspetti di maggior rilievo di tali argomenti.

BAYESIAN NETWORK

Le Continuous time Bayesian Network utilizzano una rappresentazione strutturata dello spazio degli stati propria della teoria delle Bayesian Network. Ne ereditano perciò gli aspetti chiave (e.g. indipendenza condizionale) nonché l'insieme delle tecniche algoritmiche per l'apprendimento e l'inferenza.

PROCESSI DI MARKOV

Le Continuous time Bayesian Network descrivono la dinamica evolutiva di variabili casuali tramite un processo di Markov omogeneo costituito da un insieme di processi di Markov condizionali.

1.1.1 Bayesian Network

Una Bayesian Network è un modello grafico probabilistico costituito da un grafo aciclico orientato (DAG)¹. I nodi di tale grafo rappresentano un insieme di variabili casuali mentre gli archi evidenziano le dipendenze (e le indipendenze) condizionali fra esse [Korb e Nicholson, 2011]. Una BN rappresenta la distribuzione di probabilità congiunta

¹ Un grafo aciclico orientato (anche detto grafo aciclico diretto o digrafo aciclico) è un tipo di grafo che non presenta cicli diretti: comunque si scelga un vertice non è possibile tornare ad esso percorrendo gli archi del grafo.

del suo insieme di variabili casuali tramite la distribuzione di probabilità condizionale di ognuna di essa (si veda l'equazione 1.2). Le BN sono quindi modelli grafico probabilistici con cui è possibile modellare in modo probabilistico le relazioni causali tra variabili. Esse risultano molto utili nella rappresentazione e analisi di domini caratterizzati da incertezza. Sono infatti usate in svariate applicazioni di supporto alle decisioni, bioinformatica, biologia computazionale, data mining, information retrieval e classificazione.

Rappresentazione

Di seguito si fornisce la definizione formale delle Bayesian Network e si introducono i loro aspetti basilari.

Definizione 1 (Bayesian Network). Una Bayesian Network \mathcal{B} è una coppia $\mathcal{B} = (\mathcal{G}, \theta_{\mathcal{G}})$ costituita da:

- $\mathcal{G} = (\mathbf{V}(\mathcal{G}), \mathbf{A}(\mathcal{G}))$, un grafo aciclico orientato dove:
 - $\mathbf{V}(\mathcal{G}) = \{V_1, \dots, V_n\}$ è l'insieme dei nodi, ognuno dei quali è associato ad una distribuzione di probabilità condizionale (CPD)²
 - $\mathbf{A}(\mathcal{G}) \subseteq \mathbf{V}(\mathcal{G}) \times \mathbf{V}(\mathcal{G})$ è l'insieme degli archi fra i nodi $\mathbf{V}(\mathcal{G})$
- $\theta_{\mathcal{G}}$, insieme delle CPD dei nodi che specifica $\mathbf{P}_{\mathcal{B}}$, la distribuzione di probabilità congiunta delle variabili casuali $\mathbf{X}_{\mathbf{V}(\mathcal{G})}$ a cui corrispondono i nodi $\mathbf{V}(\mathcal{G})$.

Osservazione 1.1. Ogni nodo di una BN è condizionalmente indipendente (si veda definizione 2) dai suoi non-discendenti dati i suoi nodi genitori.

La CPD di ogni variabile casuale $X_i \in \mathbf{X}_{\mathbf{V}(\mathcal{G})}$ esprime i suoi valori di probabilità in funzione dei valori assunti da $\text{Pa}(X_i)$, notazione con cui denotiamo l'insieme dei nodi genitori per ogni nodo o variabile casuale.

Un arco da un nodo genitore verso un nodo figlio di \mathcal{G} rappresenta una dipendenza diretta fra le corrispettive variabili casuali [si veda Russell e Norvig, 2003, sezione 14.1]. I nodi non direttamente connessi rappresentano variabili casuali condizionalmente indipendenti dagli altri nodi (per quanto riguarda il concetto di *indipendenza condizionale* si rimanda alla definizione 2).

Prima di procedere con la discussione si introduce la Chain Rule, proprietà fondamentale delle BN.

² Nel caso di variabili causali discrete, le CPD sono rappresentabili come delle tabelle che contengono i valori di probabilità di un nodo in funzione di tutte le possibili configurazioni dei nodi genitori (cioè l'insieme dei nodi da cui parte un arco che punta al nodo di interesse). Tali tabelle sono spesso chiamate tabelle di probabilità condizionale (CPT).

Teorema 1 (Chain Rule). *Dato un insieme di variabili casuali e una distribuzione di probabilità congiunta definita su di esse è possibile calcolare qualsiasi elemento di tale distribuzione tramite le distribuzioni di probabilità condizionale delle variabili casuali.*

Perciò, dato un insieme di variabili casuali A_1, \dots, A_n è possibile calcolare il valore di tale membro della distribuzione di probabilità congiunta applicando la definizione di probabilità condizionale:

$$\mathbf{P}(A_1, \dots, A_n) = \mathbf{P}(A_n | A_{n-1}, \dots, A_1) \cdot \mathbf{P}(A_{n-1}, \dots, A_1).$$

Ripetendo tale processo per ogni termine finale si ottiene:

$$\mathbf{P}\left(\bigcap_{k=1}^n A_k\right) = \prod_{k=1}^n \mathbf{P}\left(A_k | \bigcap_{j=1}^{k-1} A_j\right). \quad (1.1)$$

Applicando l'equazione 1.1 alle Bayesian Network diciamo che la distribuzione di probabilità congiunta $\mathbf{P}_{\mathcal{B}}$ si *fattorizza* rispetto al grafo \mathcal{G} se è possibile scrivere:

$$\mathbf{P}_{\mathcal{B}}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Pa}(X_i)). \quad (1.2)$$

L'equazione 1.2 esprime quindi la *proprietà di fattorizzazione* della distribuzione congiunta del modello grafico, detta *distribuzione di probabilità globale*, ed è ciò che permette di descriverla efficientemente in funzione delle distribuzioni condizionali dei nodi [Russell e Norvig, 2003, sezione 14.2], dette *distribuzioni di probabilità locali*. Questa proprietà contiene in sé il concetto di *proprietà di Markov* (si veda la definizione 3), il quale attesta che ogni nodo di una Bayesian Network dipende solo ed esclusivamente dai suoi nodi genitori [Korb e Nicholson, 2011, sezione 2.2.4]. Si noti inoltre, che le Bayesian Network richiedono (DAG, definizione 1) che la loro componente \mathcal{G} non contenga cicli affinché possano rispettare tale proprietà [Russell e Norvig, 2003, sezione 14.1].

Poiché, come detto, una Bayesian Network stabilisce che ogni nodo, dati i suoi genitori, è *condizionalmente indipendente* da ogni altro nodo che non sia un suo discendente, di seguito introduciamo tale concetto formalmente.

Definizione 2 (Indipendenza condizionale). Un evento A è *condizionalmente indipendente* da un evento B , data l'evidenza su un evento C , qualora la conoscenza di B non apporta alcuna variazione alla probabilità di A rispetto a quella conseguente alla conoscenza di C . Formalmente, ciò significa che:

$$\mathbf{P}(A, B | C) = \mathbf{P}(A | B, C) \cdot \mathbf{P}(B | C) = \mathbf{P}(A | C) \cdot \mathbf{P}(B | C).$$

Da cui segue che:

$$A \perp B | C \iff \mathbf{P}(A | B, C) = \mathbf{P}(A | C).$$

In termini non formali, supponendo di essere nel caso della definizione, cioè di avere una variabile casuale A *condizionalmente indipendente* da B dato C , ciò significa che è possibile ignorare B poiché essa non ha alcun riflesso sulla distribuzione condizionale di A quando sia noto l'evento C .

Si noti che il concetto appena espresso gioca un ruolo importante per i modelli probabilistici, quali sono le Bayesian Network, semplificando i calcoli richiesti per l'inferenza e l'apprendimento. Le Bayesian Network ereditano questi benefici dell'indipendenza condizionale come conseguenza della loro definizione (si veda l'osservazione 1.1). Infatti, la distribuzione condizionale di ogni variabile casuale X_i dipende solo ed esclusivamente dal valore dei suoi genitori, $Pa(X_i)$, mentre ignora completamente i valori dei nodi che non discendono da essa, $Nd(X_i)$.

Grazie alla definizione 2 è possibile esprimere in modo formale il concetto appena espresso per ogni nodo $X_i \in \mathbf{X}_{V(G)}$:

$$P(X_i | E, Pa(X_i)) = P(X_i | Pa(X_i)) \quad \forall E \in Nd(X_i),$$

dove $Nd(X_i)$ è l'insieme dei nodi non-discendenti (ed E è una variabile casuale o un insieme di variabili casuali ad essi associati). In base a ciò si dice quindi che le Bayesian Network rispettano l'*assunzione locale di Markov*.

Apprendimento e Inferenza

In questa sezione si descrivono brevemente e a scopo introduttivo i processi di apprendimento e inferenza sulle Bayesian Network.

Il problema dell'apprendimento per le Bayesian Network si divide principalmente in due casi:

- apprendere le CPD, nota la struttura
- apprendere sia le CPD, sia la struttura (incognita).

In entrambi i casi è di grande aiuto la rappresentazione efficiente delle Bayesian Network che, tramite la *fattorizzazione* della distribuzione di probabilità congiunta, permette di rappresentarla in modo compatto (tramite l'equazione 1.2) riducendo notevolmente il numero di parametri da calcolare.

Come detto, per specificare completamente una Bayesian Network è necessario rappresentare completamente la distribuzione di probabilità congiunta delle sue variabili tramite la distribuzione di probabilità condizionale di ognuna di esse. In generale, tali distribuzioni condizionali possono avere una qualsiasi forma anche se, al fine di semplificare i calcoli, è comune utilizzare distribuzioni discrete o Gaussiane per modellarle. Nel caso in cui i dati siano parzialmente osservabili solitamente si procede tramite l'algoritmo di Expectation

Maximization (EM), il quale alterna il calcolo dei valori attesi delle variabili casuali non osservate condizionalmente ai dati osservati con la massimizzazione della likelihood. Tale approccio generalmente converge ai valori di massima probabilità a posteriori per i parametri [si veda Dempster *et al.*, 1977].

Per l'apprendimento dei parametri esistono comunque una varietà di altri approcci possibili [si veda Heckerman, 1996] (e.g. trattare i parametri come variabili casuali sconosciute addizionali) che tuttavia non sono argomento di questo lavoro di tesi.

Si noti che le Bayesian Network non sono solamente un *modello discriminativo ma anche generativo* poiché possono essere utilizzate per soddisfare query arbitrarie, cioè per effettuare *inferenza probabilistica*: calcolare la distribuzione a posteriori di un insieme di variabili casuali data l'osservazione (evidenza) di altre (sfruttando il *teorema di Bayes*). In letteratura [si veda Heckerman, 1996] sono stati esplorati molti metodi di *inferenza esatta*, quali ad esempio l'eliminazione tramite integrazione o somma delle variabili non osservate che non fanno parte della query probabilistica o il metodo clique tree propagation. Questi metodi, come gli altri presenti in letteratura, dati tutti i possibili alberi di decomposizione del grafo, sono esponenziali rispetto alla larghezza minore rilevata fra essi. Per quanto riguarda invece gli algoritmi di *inferenza approssimata* si citano due tra i più comuni: l'importance sampling [Shachter e Peot, 1990] e i metodi Markov Chain Monte Carlo (MCMC) (*Gibbs sampling*, *Metropolis sampling*, e *Hybrid Monte Carlo sampling*), basati sul campionamento stocastico [si veda Geman e Geman, 1984; Gilks *et al.*, 1996; MacKay, 1998].

Nel caso in cui non si disponga della struttura di una BN è richiesto l'apprendimento strutturale. Gli algoritmi per l'apprendimento strutturale delle Bayesian Network possono essere divisi in due famiglie.

ALGORITMI BASATI SU VINCOLI

Algoritmi che apprendono la struttura del grafo analizzando le relazioni probabilistiche derivanti dalla proprietà di Markov tramite test di indipendenza condizionale e costruendo un grafo che soddisfi le proprietà di *d-separazione*³ corrispondenti. I modelli risultanti sono spesso interpretati come *modelli causali* [Pearl, 1988].

ALGORITMI BASATI SU PUNTEGGIO

Algoritmi che assegnano un punteggio (tramite una funzione di scoring) a tutte le strutture candidate e utilizzando tecniche di ottimizzazione cercano di raggiungere il punteggio massimo. Gli algoritmi di ricerca greedy (i.e. *golosi*, *aggressivi* o *avid*) a

³ Concetto di separazione direzionale tra insieme di nodi collegato al concetto di *indipendenza condizionale*. Ad esempio, quando un insieme di nodi **E** *d-separa* un insieme di nodi **X** = {A, B} allora A e B sono condizionalmente indipendenti dato E.

seconda della traduzione preferita dall'inglese) sono la scelta più comune, tuttavia qualsiasi procedura di ricerca può essere usata.

Gli *algoritmi basati su vincoli* sono basati sull'algoritmo Inductive Causation (IC) di Verma e Pearl [1991] che fornisce un contesto teorico finalizzato all'apprendimento delle strutture dei modelli causali. L'algoritmo IC può essere riassunto nei tre passi successivi.

- Apprendimento dello scheletro (i.e. grafo non diretto) della rete. Poiché la ricerca esaustiva non è, nella maggior parte dei casi, computazionalmente realizzabile, tutti gli algoritmi di apprendimento restringono la ricerca al *Markov blanket*⁴ di ogni nodo.
- Impostare la direzione degli archi che fanno parte di una *v-structure*⁵.
- Impostare la direzione degli archi fra i nodi rimanenti affinché il vincolo di aciclicità sia rispettato.

Gli *algoritmi basati su punteggio* sono invece delle applicazioni dei vari algoritmi di ricerca euristica (e.g. *hill climbing*, *tabu search*, *best first search*, *simulated annealing*) che utilizzano una funzione di *scoring*. Solitamente la funzione di *scoring* utilizza la probabilità a posteriori dell'insieme dei dati di apprendimento (i.e. training set), data la struttura in esame, ed è *score-equivalent*, affinché reti che definiscono la stessa distribuzione di probabilità abbiano lo stesso score [Chickering, 2013]. Tuttavia, per quanto questi algoritmi siano utilizzati molto frequentemente, essi sono esponenziali rispetto al numero di nodi della struttura del grafo. Inoltre, qualora si utilizzi una strategia di ricerca locale, è possibile che l'algoritmo restituisca come risultato un minimo locale (per evitare questa situazione si ricorre spesso a metodi di ricerca globale quali il MCMC). Si fa notare che è possibile ridurre il tempo necessario richiesto per l'apprendimento strutturale fissando un numero massimo di genitori candidati e cercando esaustivamente in insiemi di tale cardinalità una struttura che massimizzi l'informazione mutua fra variabili [Heckerman *et al.*, 1995].

1.1.2 Processi di Markov

Sempre al fine di preparare la discussione delle Continuous time Bayesian Network si prosegue presentando alcuni concetti propedeutici relativi ai processi di Markov, una categoria di processi stocastici con assenza di memoria [Loève, 1978].

⁴ Il *Markov blanket* di un nodo A è un insieme composto dai nodi genitori di A , dai suoi nodi figli e da tutti i nodi che condividono un figlio con A .

⁵ Una *v-structure* è una tripla di nodi $X_j \rightarrow X_i \leftarrow X_k$ incidenti su una connessione convergente.

Definizione 3 (Proprietà di Markov). Secondo la proprietà di Markov gli stati futuri di un processo stocastico sono indipendenti dagli stati passati, avendo evidenza sullo stato presente di tale processo.

Formalmente, un processo stocastico X gode di tale proprietà, se e solo se vale la seguente equazione [Loève, 1978]:

$$\mathbf{P}(X(t + \Delta t) | X(t), X(s)) = \mathbf{P}(X(t + \Delta t) | X(t)), \quad (1.3)$$

per ogni s , e t tali che $s < t < \infty$.

I modelli che rispettano tale proprietà sono detti modelli che rispettano l'*assunzione di Markov*.

Di conseguenza la distribuzione di probabilità condizionale degli stati futuri di un processo stocastico che gode di tale proprietà è indipendente dagli stati passati dato quello attuale.

In altri termini ciò indica che lo stato futuro di una variabile casuale è *condizionalmente indipendente* (si veda la definizione 2) dalla sequenza dei suoi stati passati, avendo evidenza sul suo stato presente.

Dalla proprietà di Markov deriva la definizione dei processi di Markov.

Definizione 4 (Processo di Markov). Si definisce [Loève, 1978] come processo di Markov un processo stocastico che gode della proprietà di Markov.

Definizione 5 (Catena di Markov). Un processo di Markov che può assumere solo un numero finito di stati è solitamente definito come una *catena di Markov* [si veda Norris, 1998, p. 10].

Esistono due tipi di processi di Markov: omogenei e non. Si procede quindi fornendone le definizioni.

Definizione 6 (Processo di Markov omogeneo). Un processo di Markov è detto *omogeneo* qualora $\mathbf{P}(X(t + \Delta t) | X(t))$ non dipenda dal tempo t . Affinché ciò sia vero, ponendo $t = 0$, deve risultare che:

$$\mathbf{P}(X(t + \Delta t) | X(t)) = \mathbf{P}(X(\Delta t) | X(0)). \quad (1.4)$$

Data quindi una variabile casuale X e l'insieme delle sue istanziazioni $\text{val}(X) = \{x_1, \dots, x_J\}$, $X(t)$ è un processo di Markov *omogeneo, a tempo continuo e stati finiti* se e solo se la sua dinamica è definibile in termini di:

- una distribuzione di probabilità iniziale \mathbf{P}_X^0 su $\text{val}(X)$
- una matrice di intensità \mathbf{Q}_X .

Definizione 7 (Matrice di intensità). Una matrice di intensità (IM), rappresenta un *modello di transizione Markoviano*:

$$Q_X = \begin{bmatrix} -q_{x_1} & q_{x_1 x_2} & \cdots & q_{x_1 x_k} \\ q_{x_2 x_1} & -q_{x_2} & \cdots & q_{x_2 x_k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{x_k x_1} & q_{x_k x_2} & \cdots & -q_{x_k} \end{bmatrix}.$$

Lo scopo di una matrice di intensità è descrivere il comportamento transiente di X , un processo di Markov omogeneo.

Affinché Q_X sia una matrice di intensità valida, ogni sua riga deve sommare a 0:

$$q_{x_i} = \sum_{i \neq j} q_{x_i x_j} \quad \text{con} \quad q_{x_i}, q_{x_i x_j} > 0.$$

Data quindi una matrice di intensità Q_X essa descrive il comportamento transiente di $X(t)$. Se $X(0) = x_i$ allora il processo di Markov omogeneo (e indicizzato dal tempo t) $X(t)$ rimarrà nello stato x_i una quantità di tempo *esponenzialmente distribuita* rispetto al parametro q_{x_i} . Di conseguenza la *funzione di densità* f e la corrispondente *funzione di ripartizione*⁶ F sono:

$$\begin{aligned} f(t) &= q_{x_i} \exp(-q_{x_i} t), \quad t > 0 \\ F(t) &= 1 - \exp(-q_{x_i} t), \quad t \geq 0. \end{aligned} \tag{1.5}$$

Mentre gli elementi sulla diagonale, q_{x_i} , codificano una quantità che può essere interpretata come la “*probabilità istantanea*” che X abbandoni lo stato x_i , gli elementi non sulla diagonale, $q_{x_i x_j}$, esprimono l’*intensità di transizione* dallo stato x_i allo stato x_j .

Possiamo quindi calcolare:

- il *tempo atteso* di una transizione uscente dallo stato x_i

$$1/q_{x_i}$$

- la “*probabilità istantanea*” di transizione dallo stato x_i allo stato x_j

$$\theta_{x_i x_j} = q_{x_i x_j} / q_{x_i}.$$

Quindi una matrice di intensità induce una distribuzione di probabilità locale fattorizzata in due parti:

- q_{x_i} , che esprime quando avvengono le transizioni attraverso una *distribuzione di probabilità esponenziale*

⁶ Nel calcolo delle probabilità la funzione di ripartizione di una variabile casuale X a valori reali, anche nota come funzione di distribuzione cumulativa, è la funzione che associa a ciascun valore x la probabilità che X assuma valori minori o uguali ad x .

- $\theta_{x_{ij}}$, che esprime la *distribuzione di probabilità multinomiale* tra coppie di stati $i \neq j$.

La matrice \mathbf{Q}_X fa in modo che X soddisfi la proprietà di Markov poiché il comportamento futuro di X è definito solamente in base al suo stato attuale (vale l'equazione 1.4).

Osservazione 7.1. L'ordine di una matrice di intensità corrisponde a $k = |\text{val}(X)|$, la cardinalità dell'insieme dei valori assunti da X .

Osservazione 7.2. Nel caso in cui $k = 2$, la matrice di intensità \mathbf{Q}_X è quindi una matrice 2×2 . Di conseguenza, poiché ogni riga di \mathbf{Q}_X somma a 0 per definizione, le “*probabilità istantanee*” di transizione fra gli stati di X sono sempre pari a 1.

Definizione 8 (Processo di Markov condizionale). Un processo di Markov le cui intensità di transizione variano nel tempo non in funzione del tempo ma in funzione dei valori assunti ad ogni determinato istante t da un insieme di altre variabili, che evolvono anch'esse come dei processi di Markov, è detto essere un processo di Markov condizionale (o processo di Markov non omogeneo).

Assumendo quindi che una variabile casuale X evolva come un processo di Markov $X(t)$ e che la sua dinamica sia condizionata da un insieme di altre variabili casuali $\mathbf{Pa}(X)$, anch'esse dei processi di Markov, possiamo definire per tale variabile casuale una matrice di intensità condizionale (CIM) $\mathbf{Q}_{X|\mathbf{Pa}(X)}$.

Se specifichiamo una distribuzione di probabilità iniziale su X abbiamo così definito un processo di Markov il cui comportamento dipende dalle istanziazioni dei valori di $\mathbf{Pa}(X)$.

Definizione 9 (Matrice di intensità condizionale). Dato un insieme di processi di Markov $\mathbf{Pa}(X)$, una matrice di intensità condizionale $\mathbf{Q}_{X|\mathbf{Pa}(X)}$ è costituita da un insieme di matrici di intensità $\mathbf{Q}_{X|\mathbf{pa}_i(x)}$, una per ogni diversa istanziazione $\mathbf{pa}_i(x)$ di $\mathbf{Pa}(X)$ [Stella e Amer, 2012]:

$$\mathbf{Q}_{X|\mathbf{Pa}(X)} = \{ \mathbf{Q}_{X|\mathbf{pa}_1(x)}, \mathbf{Q}_{X|\mathbf{pa}_2(x)}, \dots, \mathbf{Q}_{X|\mathbf{pa}_n(x)} \}.$$

Ogni matrice di intensità di $\mathbf{Q}_{X|\mathbf{Pa}(X)}$ è del seguente tipo:

$$\mathbf{Q}_{X|\mathbf{pa}_i(x)} = \begin{bmatrix} -\mathbf{q}_{x_1}^{\mathbf{pa}_i(x)} & \mathbf{q}_{x_1x_2}^{\mathbf{pa}_i(x)} & \dots & \mathbf{q}_{x_1x_k}^{\mathbf{pa}_i(x)} \\ \mathbf{q}_{x_2x_1}^{\mathbf{pa}_i(x)} & -\mathbf{q}_{x_2}^{\mathbf{pa}_i(x)} & \dots & \mathbf{q}_{x_2x_k}^{\mathbf{pa}_i(x)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_{x_kx_1}^{\mathbf{pa}_i(x)} & \mathbf{q}_{x_kx_2}^{\mathbf{pa}_i(x)} & \dots & -\mathbf{q}_{x_k}^{\mathbf{pa}_i(x)} \end{bmatrix}.$$

Di seguito si presenta un breve esempio finalizzato alla comprensione pratica delle matrici di intensità condizionali (CIM) e del loro scopo.

Esempio 9.1.

Date due variabili causali, $E(t)$ e $H(t)$, delle quali la prima modella l'eventualità che un individuo stia mangiando o meno (se $e = 2$ allora l'individuo sta mangiando, viceversa se $e = 1$) mentre la seconda modella l'eventualità che lo stesso individuo abbia fame o meno (se $h = 2$ allora l'individuo è affamato, viceversa se $h = 1$) e la matrice di intensità condizionale $\mathbf{Q}_{E|H}$, che è un insieme composto dalle matrici di intensità $\mathbf{Q}_{E|h=1}$ e $\mathbf{Q}_{E|h=2}$, è possibile calcolare la probabilità degli eventi della variabile casuale E condizionatamente all'evidenza che si possiede sulla variabile casuale H .

$$\mathbf{Q}_{E|h=1} = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} -.01 & .01 \\ 10 & -10 \end{bmatrix} \end{matrix} \quad \mathbf{Q}_{E|h=2} = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} -2 & 2 \\ .01 & -.01 \end{bmatrix} \end{matrix}.$$

Ipotizzando che l'unità temporale corrisponda a un'ora:

- un individuo affamato ($h = 2$) che non sta mangiando ($e = 1$) inizierà a mangiare in 30 minuti poiché

$$\frac{1}{q_{e=1|h=2}} = \frac{1}{2},$$

- un individuo non affamato ($h = 1$) che sta mangiando ($e = 2$) smetterà di mangiare ($e = 1$) entro 6 minuti poiché

$$\frac{1}{q_{e=2|h=1}} = \frac{1}{10};$$

si osservi che la “*probabilità istantanea*” di transizione da $e = 2$ a $e = 1$, come già evidenziato dall'osservazione 7.2, è

$$\theta_{e=2,e=1|h=1} = \frac{q_{e=2,e=1|h=1}}{q_{e=2|h=1}} = \frac{10}{10} = 1.$$

1.2 DEFINIZIONI PRELIMINARI

Nelle precedenti sezioni sono stati illustrati i concetti che si pongono a fondamento delle Continuous time Bayesian Network:

- le Bayesian Network: utili a comprendere la rappresentazione strutturata dello spazio degli stati delle CTBN, l'utilizzo della nozione di indipendenza condizionale e le conseguenti tecniche di apprendimento e inferenza
- i processi di Markov, omogenei e non, al fine di introdurre le modalità di rappresentazione (qualitativa e quantitativa) delle CTBN.

Prima di presentare le Continuous time Bayesian Network come una collezione di processi di Markov a tempo continuo non omogenei e con spazio degli stati discreto [Nodelman, 2007], si forniscono alcune definizioni utili per il prosieguo della discussione.

Definizione 10 (Variabile di processo). Una variabile di processo \mathbf{X} , anche detta Process Variable (PV) [Nodelman, 2007], è un insieme di processi di Markov a tempo continuo $\mathbf{X}(t)$.

Definizione 11 (Traiettoria). Istanziatura di un insieme di valori per $\mathbf{X}(t)$ al variare di t .

Definizione 12 (J-time-segment). Partizionamento di un intervallo temporale $[0, T)$ in J intervalli chiusi a sinistra:

$$[0, t_1); [t_1, t_2); \dots; [t_{J-1}, T).$$

Nota 12.1. È possibile riferirsi a tale concetto anche tramite l'espressione "insieme dei segmenti temporali".

Definizione 13 (J-evidence-stream). Dato un insieme di segmenti temporali composto da J intervalli e una variabile di processo \mathbf{X} composta da N variabili casuali, un *J-evidence-stream* è l'insieme delle istanziazioni comuni $\mathbf{X} = \mathbf{x}$ associate ad ogni intervallo temporale per ogni sottoinsieme delle variabili casuali [Stella e Amer, 2012]. È denotato con $(\mathbf{X}^1 = \mathbf{x}^1, \mathbf{X}^2 = \mathbf{x}^2, \dots, \mathbf{X}^J = \mathbf{x}^J)$, o più concisamente $(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J)$.

Nota 13.1. È possibile riferirsi a tale concetto anche tramite l'espressione "flusso di evidenze".

Nota 13.2. Un flusso di evidenze $(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J)$ è detto essere *completamente osservato* se lo stato di tutte le variabili $X_n \in \mathbf{X}$ è conosciuto in tutto l'intervallo $[0, T)$. Viceversa, un flusso di evidenze è detto *parzialmente osservato*.

1.3 RAPPRESENTAZIONE

Una Continuous time Bayesian Network è un modello grafico in cui ogni nodo rappresenta una variabile casuale i cui stati evolvono in modo continuo nel tempo. Le dinamiche evolutive degli stati dei nodi sono governate e dipendono dal valore che gli stati dei nodi padre⁷ assumono [Stella e Amer, 2012]. Quindi ogni nodo è un processo di Markov condizionale (si veda la definizione 8) a tempo continuo e spazio degli stati discreto.

Una CTBN è composta principalmente da due componenti:

⁷ Con il termine "nodo padre", o *parent node*, si intende un nodo il cui stato condiziona quello di un altro nodo del modello grafico.

- una distribuzione di probabilità iniziale
- le dinamiche che regolano l'evoluzione nel tempo continuo della distribuzione di probabilità

Più formalmente si definisce:

Definizione 14 (Continuos time Bayesian Network). Data una variabile di processo \mathbf{X} , insieme di processi di Markov X_1, X_2, \dots, X_N a tempo continuo e con spazio degli stati finito $\text{val}(X_n) = \{x_1, \dots, x_J\}$ (dove $n = 1, \dots, N$), una CTBN \mathcal{N} su \mathbf{X} consiste di:

- una distribuzione di probabilità iniziale \mathbf{P}_X^0 specificata come una Bayesian Network \mathcal{B} su \mathbf{X}
- un modello di transizione a tempo continuo, specificato da:
 - un grafo \mathcal{G} , orientato e non necessariamente aciclico, composto dai nodi X_1, X_2, \dots, X_N , ognuno dei quali possiede un insieme di genitori denotato da $\text{Pa}(X_n)$
 - una matrice di intensità condizionale $\mathbf{Q}_{X_n | \text{Pa}(X_n)}$ per ogni nodo $X_n \in \mathbf{X}$.

Per ogni variabile causale $X_n \in \mathbf{X}$ di \mathcal{N} si ha quindi un insieme di modelli di probabilità locali: $\mathbf{Q}_{X_n | \text{Pa}(X_n)}$, la CIM di X_n , è infatti un insieme di modelli di transizione Markoviani la cui cardinalità è pari a quella dell'insieme delle diverse istanziazioni di $\text{Pa}(X_n)$.

Si riscontra, quindi, quanto già affermato in precedenza (si veda 1.1), cioè che una CTBN esprime la sua dinamica evolutiva globale tramite un unico processo di Markov omogeneo, costituito da un insieme di processi di Markov condizionali (un insieme di CIM e relative distribuzioni di probabilità iniziali).

Si noti che, diversamente dalle Bayesian Network, nelle Continuos time Bayesian Network gli archi fra i nodi rappresentano le dipendenze nel tempo. Per tale motivo è possibile che la componente \mathcal{G} del modello di transizione continuo contenga dei cicli. Tra l'altro, come vedremo nel prosieguo, la mancanza di tale vincolo di aciclicità porta a notevoli vantaggi computazionali relativamente all'apprendimento della struttura di una CTBN dai dati.

1.4 APPENDIMENTO

In questa sezione si argomenta sulla probabilità di un *insieme di dati completo* rispetto a una Continuos time Bayesian Network. A tal fine si mostra come una CTBN, essendo un modello esponenziale, possa essere decomposta in un aggregato di modelli di probabilità locali relativi alle singole variabili casuali e espressa in termini di *statistiche sufficienti* aggregate.

Si affronta infine il processo di apprendimento dei parametri delle Continuous time Bayesian Network da *dati completi*. I processi di apprendimento relativi a dati non completi sono tralasciati poiché non facenti parte degli argomenti di questo lavoro di tesi.

Definizione 15 (Insieme di dati completo). Dato un insieme di variabili casuali, un insieme di dati $\mathcal{D} = \{\delta_1, \dots, \delta_h\}$ si dice *completo* se ogni δ_i (con $i = 1, \dots, h$) è un insieme di traiettorie completamente osservate delle variabili casuali (i.e. l'istanziatura di tutte le variabili casuali è osservabile per ogni istante temporale di ogni traiettoria).

1.4.1 Statistiche sufficienti

Le *statistiche sufficienti* per un singolo processo di Markov omogeneo $X(t)$ riassumono la sua dinamica evolutiva con:

- $T[x]$: la quantità di tempo trascorsa nello stato x
- $M[x, x']$: il numero di transizioni dallo stato x allo stato x' .

Il numero totale di transizioni uscenti da uno stato x è:

$$M[x] = \sum_{x'} M[x, x'].$$

Nel caso di un processo di Markov condizionale è invece necessario considerare anche l'istanziatura dell'insieme $Pa(X)$ dei nodi genitori:

- $T[x | pa_i(x)]$: la quantità di tempo trascorsa nello stato x quando $Pa(X) = pa_i(x)$
- $M[x, x' | pa_i(x)]$: il numero di transizioni dallo stato x allo stato x' quando $Pa(X) = pa_i(x)$.

Chiaramente, il numero totale di transizioni si calcola come sopra.

1.4.2 Likelihood

Al fine di presentare il calcolo della likelihood⁸ di una CTBN rispetto a un dataset completo \mathcal{D} è bene procedere per gradi e iniziare presentando dapprima la likelihood di una singola transizione di un singolo processo di Markov omogeneo $X(t)$.

Likelihood di una singola transizione

Data una tripla $d = \langle x_d, t_d, x_{d'} \rangle \in \mathcal{D}$, la quale esprime una transizione di $X(t)$ da x_d a $x_{d'}$ dopo che esso ha trascorso t_d tempo in

⁸ La likelihood di un insieme di valori per i parametri, dato un insieme di dati, è uguale alla probabilità dei dati, dati tali valori per i parametri.

x_d , è possibile scrivere la likelihood di questa singola transizione d in funzione dei parametri (1.4.3):

$$\begin{aligned} L_X(\mathbf{q}, \boldsymbol{\theta} : d) &= L_X(\mathbf{q} : d) L_X(\boldsymbol{\theta} : d) \\ &= \mathbf{q}_{x_d} \exp(-\mathbf{q}_{x_d} \mathbf{t}_d) (\boldsymbol{\theta}_{x_d x_{d'}}). \end{aligned} \quad (1.6)$$

Si noti che l'equazione 1.6 è ricavata moltiplicando la *funzione di distribuzione di probabilità* di $X(t)$ (equazione 1.5) per la “*probabilità istantanea*” di transizione (si veda la definizione 7).

Likelihood di un dataset completo

Poiché tutte le transizioni sono osservabili, la *likelihood* del dataset \mathcal{D} può essere decomposta come un prodotto delle likelihood individuali di ogni singola transizione d [si veda Nodelman *et al.*, 2002, p. 3]. Per tale motivo \mathcal{D} è sintetizzabile aggregando le *statistiche sufficienti* relative a ogni processo di Markov condizionale di una CTBN.

Quindi la likelihood di un dataset completo \mathcal{D} rispetto a un singolo processo di Markov omogeneo $X(t)$ è:

$$\begin{aligned} L_X(\mathbf{q}, \boldsymbol{\theta} : \mathcal{D}) &= \left[\prod_{d \in \mathcal{D}} L_X(\mathbf{q} : d) \right] \left[\prod_{d \in \mathcal{D}} L_X(\boldsymbol{\theta} : d) \right] \\ &= \left[\prod_x \mathbf{q}_x^{M[x]} \exp(-\mathbf{q}_x T[x]) \right] \left[\prod_x \prod_{x' \neq x} \boldsymbol{\theta}_{xx'}^{M[x, x']} \right]. \end{aligned} \quad (1.7)$$

Si supponga ora di traslare questo concetto a una Continuous time Bayesian Network \mathcal{N} con N nodi: per ogni nodo X_i , con $i = 1, \dots, N$ è necessario considerare tutte le transizioni contestualmente all'istanziamento dell'insieme $\text{Pa}(X_i)$ dei suoi nodi genitori. Poiché, nel caso di *dati completi*, si conosce sempre l'istanziamento di $\text{Pa}(X_i)$, allora, per ogni istante di tempo t , si conosce quale matrice di intensità $\mathbf{Q}_{X_i | \text{Pa}(X_i)}$, con $\text{pa}_i(x) \in \text{Pa}(X_i)$, governi la dinamica di X_i .

Perciò la probabilità dei dati \mathcal{D} rispetto a \mathcal{N} è il prodotto delle likelihood di ogni variabile X_i :

$$\begin{aligned} L_{\mathcal{N}}(\mathbf{q}, \boldsymbol{\theta} : \mathcal{D}) &= \prod_{X_i \in \mathcal{X}} L_{X_i}(\mathbf{q}_{X_i | \text{Pa}(X_i)}, \boldsymbol{\theta}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) \\ &= \prod_{X_i \in \mathcal{X}} L_{X_i}(\mathbf{q}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) L_{X_i}(\boldsymbol{\theta}_{X_i | \text{Pa}(X_i)} : \mathcal{D}). \end{aligned} \quad (1.8)$$

Il termine $L_X(\boldsymbol{\theta}_{X | \text{Pa}(X)} : \mathcal{D})$ esprime la likelihood delle transizioni tra stati. Si osservi, inoltre, come il tempo che intercorre fra le transizioni sia trascurato poiché esse dipendono esclusivamente dal valore di nodi genitori [si veda Nodelman *et al.*, 2002, p. 3]. Quindi, usando le *statistiche sufficienti* si può scrivere:

$$L_X(\boldsymbol{\theta}_{X | \text{Pa}(X)} : \mathcal{D}) = \prod_{\text{pa}_i(x)} \prod_x \prod_{x' \neq x} \boldsymbol{\theta}_{xx'}^{M[x, x' | \text{pa}_i(x)]}.$$

Per quanto riguarda il calcolo di $L_X(\mathbf{q}_{X|Pa(X)} : \mathcal{D})$ va considerato il caso in cui il tempo trascorso da X in uno determinato stato x termini non a causa di una sua transizione bensì a causa di una transizione di uno o più nodi appartenenti all'insieme dei suoi nodi genitori (i.e. una nuova istanziazione per l'insieme dei genitori $Pa(X)$). È quindi necessario considerare la probabilità che il nodo X rimanga in x una quantità di tempo *almeno pari* a t mentre i suoi nodi genitori $Pa(X)$ non effettuano alcuna transizione di stato [si veda Nodelman *et al.*, 2002, p. 3]. Tale quantità si ricava dalla funzione di distribuzione cumulativa di una distribuzione esponenziale (equazione 1.5):

$$1 - F(t) = \exp(-\mathbf{q}_{x|pa_i(x)} t).$$

Perciò la likelihood delle quantità di tempo trascorse in ogni stato è:

$$L_X(\mathbf{q}_{X|Pa(X)} : \mathcal{D}) = \prod_{pa_i(x)} \prod_x \mathbf{q}_{x|pa_i(x)}^{M[x|pa_i(x)]} \exp(-\mathbf{q}_{x|pa_i(x)} T[x|pa_i(x)]).$$

La likelihood di \mathcal{N} è quindi:

$$L_{\mathcal{N}}(\mathbf{q}, \boldsymbol{\theta} : \mathcal{D}) = \prod_{pa_i(x)} \prod_x \left[\mathbf{q}_{x|pa_i(x)}^{M[x|pa_i(x)]} \exp(-\mathbf{q}_{x|pa_i(x)} T[x|pa_i(x)]) \cdot \prod_{x \neq x'} \boldsymbol{\theta}_{xx'|pa_i(x)}^{M[x, x'|pa_i(x)]} \right]. \quad (1.9)$$

Mentre, scrivendola come *log-likelihood*, si ottiene:

$$\begin{aligned} \ell_{\mathcal{N}}(\mathbf{q}, \boldsymbol{\theta} : \mathcal{D}) = \sum_{pa_i(x)} \sum_x \left[M[x|pa_i(x)] \ln(\mathbf{q}_{x|pa_i(x)}) - \mathbf{q}_{x|pa_i(x)} T[x|pa_i(x)] + \right. \\ \left. + \sum_{x \neq x'} M[x, x'|pa_i(x)] \ln(\boldsymbol{\theta}_{xx'|pa_i(x)}) \right]. \quad (1.10) \end{aligned}$$

In questa sezione si è presentato come computare la likelihood di un modello di una CTBN rispetto a un dataset completo.

Tuttavia, nel caso in cui non si conoscano i parametri di una CTBN è necessario stimarli. Nella prossima sezione viene affrontato esattamente questo argomento.

1.4.3 Stima dei parametri

Si affronta ora il problema dell'apprendimento dei parametri di una Continuous time Bayesian Network (con struttura nota \mathcal{G}) da un insieme di dati completi [si veda Nodelman, 2007, sezione 5.1].

La stima dei parametri non è un processo fine a sè stesso, in quanto, da essi è possibile costruire le matrici di intensità condizionali (CIM) di ogni nodo della CTBN. Come si ricorderà, una CIM è un insieme di matrici di intensità, una per ogni istanziazione $pa_i(x)$ dei nodi genitori (si veda la definizione 9). Perciò, fissato $pa_i(x)$, si può computare la rispettiva matrice di intensità per un nodo qualsiasi ponendo sulla diagonale il suo vettore dei parametri $\mathbf{q}_{x|pa_i(x)}$ e ricavando i valori

Come costruire una CIM dai parametri.

non sulla diagonale (i.e. la “*probabilità istantanea*” di transizione fra due stati; si veda a riguardo la definizione 7) dalla relazione fra i parametri \mathbf{q} e θ :

$$\mathbf{q}_{xx'|pa_i(x)} = \theta_{xx'|pa_i(x)} \cdot \mathbf{q}_{x|pa_i(x)}. \quad (1.11)$$

Infine, come vedremo in seguito nel capitolo 3, i parametri sono un componente chiave del processo di apprendimento strutturale.

Stima maximum-likelihood

In base a quanto attestato dalla definizione stessa delle CTBN (14), la dinamica evolutiva globale di una CTBN, cioè la dinamica di tutti i nodi di \mathcal{G} (dei processi di Markov condizionali indicizzati dal tempo), è espressa tramite un processo di Markov omogeneo. Dalla definizione 7, inoltre, si deduce che tale processo di Markov induce un modello di probabilità composto da una *distribuzione esponenziale* con parametro $\mathbf{q}_{x|pa_i(x)}$, che esprime il tempo trascorso in uno stato x da un nodo X data una istanziazione $pa_i(x)$ per i nodi genitori $Pa(X)$, e una *distribuzione multinomiale* con parametro $\theta_{xx'|pa_i(x)}$, che esprime il numero di transizioni uscenti da uno stato x verso x' (sempre fermo restando il condizionamento dato dall'istanziamento dei nodi genitori).

La media della distribuzione esponenziale in questione è pari a $1/\mathbf{q}_{x|pa_i(x)}$. Questa quantità esprime il tempo medio delle transizioni uscenti da uno stato x , fermo restando che il genitore del nodo in questione abbia istanziazione costante e uguale a $pa_i(x)$. Poiché il tempo medio si calcola rapportando il tempo totale trascorso in x , $T[x|pa_i(x)]$, rispetto al numero totale di transizioni uscenti da x , $M[x|pa_i(x)]$, si ottiene:

$$\frac{1}{\mathbf{q}_{x|pa_i(x)}} = \frac{T[x|pa_i(x)]}{M[x|pa_i(x)]}.$$

Invece, la probabilità di transizione da uno stato x verso x' è data dal rapporto tra il numero totale di transizioni da x a x' diviso il numero totale di transizioni uscenti da x ; cioè:

$$\frac{M[x, x'|pa_i(x)]}{M[x|pa_i(x)]}.$$

Teorema 2. *Parametri maximum-likelihood (MLE). I parametri che massimizzano la likelihood (equazione 1.10) di una Continuous time Bayesian Network sono funzione delle statistiche sufficienti:*

$$\begin{aligned} \mathbf{q}_{x|pa_i(x)} &= \frac{M[x|pa_i(x)]}{T[x|pa_i(x)]} \\ \theta_{xx'|pa_i(x)} &= \frac{M[x, x'|pa_i(x)]}{M[x|pa_i(x)]}. \end{aligned} \quad (1.12)$$

Si noti che, in questo caso (*dataset completo*), $\mathbf{q}_{x|pa_i(x)}$ e $\theta_{xx'|pa_i(x)}$ sono delle *stime esatte*. Essi massimizzano la probabilità a posteriori di un dataset, dato un modello CTBN.

Stima bayesiana

Un approccio alternativo alla stima dei parametri è la stima bayesiana [si veda Nodelman, 2007, sezione 5.1.1].

A tal fine è necessario definire una distribuzione a priori sui parametri di una CTBN. Come si è soliti fare in situazioni di questo tipo, per tale distribuzione si sceglie di usare una *distribuzione a priori coniugata*⁹ poiché ciò risulta conveniente dal punto di vista algebrico (e quindi computazionale). Infatti, una distribuzione a priori coniugata fornisce un'espressione in forma chiusa per la distribuzione a posteriori (alternativamente potrebbe risultare necessario il calcolo di un integrale numerico).

Consideriamo innanzitutto un singolo processo di Markov. Si ricorda (si veda a tal riguardo la definizione 7) che un processo di Markov ha due insiemi di parametri: una *distribuzione multinomiale* parametrizzata dal parametro θ e una *distribuzione esponenziale* parametrizzata dal parametro \mathbf{q} .

Una distribuzione a priori coniugata per il parametro \mathbf{q} è la *distribuzione Gamma* $P(\mathbf{q}) = \text{Gamma}(\alpha, \tau)$. Perciò la distribuzione a priori coniugata su \mathbf{q} è:

$$P(\mathbf{q}) = \frac{\tau^\alpha}{\Gamma(\alpha)} \mathbf{q}^{\alpha-1} \exp(-\mathbf{q} \tau). \quad (1.13)$$

Invece, per la distribuzione multinomiale si sceglie come priori coniugata la *distribuzione di Dirichlet* $P(\theta) = \text{Dir}(\alpha_{x_{x_1}}, \dots, \alpha_{x_{x_k}})$ [si veda Heckerman, 1996; Heckerman *et al.*, 1995].

Se si assume che i parametri sono stocasticamente indipendenti, cioè che $P(\theta, \mathbf{q}) = P(\theta) P(\mathbf{q})$, allora le distribuzioni a posteriori (i.e. condizionate sui dati) dei parametri \mathbf{q} e θ sono:

$$\begin{aligned} P(\mathbf{q} | \mathcal{D}) &= \text{Gamma}(\alpha + M[x], \tau + T[x]) \\ P(\theta | \mathcal{D}) &= \text{Dir}(\alpha_{x_{x_1}} + M[x, x_1], \dots, \alpha_{x_{x_k}} + M[x, x_k]). \end{aligned} \quad (1.14)$$

Al fine di generalizzare quest'idea e ottenere una distribuzione a priori coniugata per un'intera CTBN è necessario che essa soddisfi due assunzioni (comuni per le distribuzioni a priori nelle Bayesian Network, si veda Heckerman [1996]): l'indipendenza globale e locale dei parametri. In base all'*indipendenza globale dei parametri* possiamo scrivere:

$$P(\mathbf{q}, \theta) = \prod_{X_i \in \mathbf{X}} P(\mathbf{q}_{X_i} | P_a(X_i), \theta_{X_i} | P_a(X_i)). \quad (1.15)$$

⁹ In teoria della probabilità bayesiana, se le distribuzioni a posteriori $P(\theta|x)$ sono nella stessa famiglia della distribuzione a priori $P(\theta)$, le due distribuzioni sono definite coniugate, e la distribuzione a priori è chiamata *distribuzione a priori coniugata* per la verosimiglianza (*likelihood*). Una distribuzione a priori coniugata è conveniente dal punto di vista algebrico in quanto fornisce una espressione in forma chiusa per la distribuzione a posteriori e perché può fornire delle intuizioni circa il modo con cui la funzione di verosimiglianza aggiorna la distribuzione.

Invece, in base all'*indipendenza locale dei parametri* possiamo scrivere:

$$P(\mathbf{q}_X | \mathbf{p}_A(X), \boldsymbol{\theta}_X | \mathbf{p}_A(X)) = \left[\prod_x \prod_{\mathbf{p}_{A_i}(x)} P(\mathbf{q}_x | \mathbf{p}_{A_i}(x)) \right] \left[\prod_x \prod_{\mathbf{p}_{A_i}(x)} P(\boldsymbol{\theta}_x | \mathbf{p}_{A_i}(x)) \right]. \quad (1.16)$$

Se tale distribuzione a priori soddisfa le assunzioni di indipendenza allora anche la distribuzione a posteriori, essendovi coniugata e perciò appartenente alla stessa famiglia parametrica, le soddisferà. In tal caso è possibile mantenere la distribuzione parametrica in forma chiusa e aggiornarla usando le *statistiche sufficienti*:

- $M[x, x' | \mathbf{p}_{A_i}(x)]$ per il parametro $\boldsymbol{\theta}_x | \mathbf{p}_{A_i}(x)$
- $M[x | \mathbf{p}_{A_i}(x)]$ e $T[x | \mathbf{p}_{A_i}(x)]$ per il parametro $\mathbf{q}_x | \mathbf{p}_{A_i}(x)$.

Data una distribuzione sui parametri è possibile usarla per predire il prossimo evento, mediando la sua probabilità sull'insieme dei possibili valori dei parametri. Questo tipo di previsione è equivalente all'utilizzo dei valori "*attesi*" dei parametri, i quali hanno la stessa forma dei parametri maximum-likelihood ma considerano i *conteggi immaginari* degli iper-parametri:

$$\begin{aligned} \hat{\mathbf{q}}_{x | \mathbf{p}_{A_i}(x)} &= \frac{\alpha_{x | \mathbf{p}_{A_i}(x)} + M[x | \mathbf{p}_{A_i}(x)]}{\tau_{x | \mathbf{p}_{A_i}(x)} + T[x | \mathbf{p}_{A_i}(x)]} \\ \hat{\boldsymbol{\theta}}_{xx' | \mathbf{p}_{A_i}(x)} &= \frac{\alpha_{xx' | \mathbf{p}_{A_i}(x)} + M[x, x' | \mathbf{p}_{A_i}(x)]}{\alpha_{x | \mathbf{p}_{A_i}(x)} + M[x | \mathbf{p}_{A_i}(x)]}. \end{aligned} \quad (1.17)$$

Si osservi che questi parametri sono, teoricamente, validi solo per predire una singola transizione, dopo la quale la distribuzione dei parametri andrebbe aggiornata di conseguenza. Tuttavia, com'è spesso fatto in situazioni di questo tipo, si approssima la stima bayesiana congelando i parametri ai succitati parametri "*attesi*", usandoli per la predizione di una intera traiettoria (i. e. la distribuzione dei parametri non viene aggiornata ad ogni singola transizione).

2

CLASSIFICAZIONE

La classificazione è un argomento centrale nei campi di ricerca relativi all'apprendimento automatico (anche detto machine learning) e l'analisi dei dati. In generale, essa consiste nel processo di assegnare una *classe* (i.e. un'etichetta) a delle istanze descritte da un insieme di attributi. Si parla di *classificazione supervisionata* quando è necessario indurre un classificatore a partire da un insieme di dati composto da istanze già etichettate e utilizzare tale classificatore per classificare nuove istanze di dati.

In questo capitolo viene quindi introdotta una classe di modelli, che prende il nome di Continuous time Bayesian Network classifier (CTBNC), il cui scopo è la *classificazione supervisionata* di traiettorie multivariate di variabili discrete a *tempo continuo*. Si descrivono due istanze di tale classe: i classificatori Continuous time Naïve Bayes (CTNB) e i classificatori Continuous time tree augmented Naïve Bayes (CTTANB).

Mentre nella sezione 2.2 si affronta il processo di *apprendimento* in caso di *dati completi* dei CTBNC, nella sezione 2.3, si presenta un algoritmo di *inferenza esatta* per la classe dei CTBNC.

2.1 MODELLO

Al fine di risolvere il succitato problema della classificazione sono stati proposti numerosi approcci. Ad esempio naïve Bayes classifier, un classificatore semplice ma robusto proposto da Duda e Hart [1973]; rivelatosi essere uno fra i classificatori più performanti [Langley *et al.*, 1992]. Esso apprende dai dati la probabilità condizionale di ogni attributo A_i data la classe C . La classificazione di nuove istanze dei dati è effettuata applicando la *regola di Bayes* al fine di calcolare la probabilità della classe C data l'istanziatura di A_1, \dots, A_N e scegliendo quella con la maggiore probabilità a posteriori. Questo calcolo è reso possibile da un'assunzione forte: tutti gli attributi A_i sono *condizionalmente indipendenti* (si veda la definizione 2) tra di loro data evidenza sulla classe C .

Poiché tale assunzione è chiaramente irrealistica, Friedman *et al.* [1997] ha investigato come migliorare ulteriormente le prestazioni del naïve Bayes classifier evitando assunzioni di indipendenza non giustificate dai dati. A tal fine Friedman *et al.* [1997], generalizzando il naïve Bayes classifier, ha proposto una classe di modelli di *classificazione supervisionata*, chiamata Bayesian Network classifier (BNC) (di cui fa

parte il Tree Augmented Naïve Bayes (TAN) classifier, ad esempio) che ereditano dalla teoria delle Bayesian Network (si rimanda alla definizione 1 per maggiori dettagli) una rappresentazione fattorizzata delle distribuzioni di probabilità dei nodi attributo e rappresentano esplicitamente le indipendenze condizionali fra essi.

Seguendo le stesse motivazioni, in Stella e Amer [2012] viene formalizzata una classe di modelli di *classificazione supervisionata*, chiamati Continuous time Bayesian Network classifier (CTBNC), derivata dalle CTBN (si veda definizione 14).

Di seguito si definiscono quindi i Continuous time Bayesian Network classifier e due istanze di classificatori appartenenti a tale classe: il Continuous time Naïve Bayes classifier (CTNBC) e il Continuous time tree augmented Naïve Bayes classifier (CTTANBC).

Un Continuous time Bayesian Network classifier estende una CTBN tramite l'aggiunta di un nodo associato alla variabile classe Y . Si ricorda, dalla definizione 14, che una CTBN rappresenta l'evoluzione nel tempo continuo di una variabile di processo \mathbf{X} (i.e. insieme composto da N processi di Markov, si veda la definizione 10).

Di seguito si dà la definizione di questa nuova classe di modelli di *classificazione supervisionata*.

Definizione 16 (Continuous time Bayesian Network classifier). Un Continuous time Bayesian Network classifier (CTBNC) è composto da una coppia $\mathcal{C} = (\mathcal{N}, \mathbf{P}(Y))$ dove:

- \mathcal{N} è una CTBN con nodi attributo X_1, X_2, \dots, X_N
- Y è il nodo classe con valori $\text{val}(Y) = \{y_1, \dots, y_K\}$ e probabilità marginale $\mathbf{P}(Y)$.

E inoltre il grafo su \mathcal{N} (i.e. il grafo \mathcal{G} , si veda la definizione 14) rispetta le seguenti condizioni:

- \mathcal{G} è un grafo connesso¹⁰
- $\text{Pa}(Y) = \{\}$, i.e. la variabile casuale Y è associata a un nodo radice¹¹
- il nodo Y è indipendente dal tempo ed è specificato solo ed esclusivamente dalla sua probabilità marginale $\mathbf{P}(Y)$.

A supporto della definizione 16, la figura fig. 2.1 nella pagina seguente fornisce un'istanza di CTBNC composta dai nodi attributi X_1, X_2, X_3, X_4, X_5 e dal nodo classe Y (nodo radice). Si osservi come tale istanza contenga dei cicli, uno riguardante i nodi X_2, X_4, X_5, X_3 e l'altro riguardante i nodi X_1, X_3 . Si fa notare che gli archi della rete \mathcal{N} rappresentano le dipendenze causali nel tempo.

¹⁰ Il grafo $\mathcal{G} = (V, E)$ è detto *connesso* se $\forall (u, v) \in V$ esiste un cammino che collega u a v .

¹¹ In un grafo un nodo è detto *radice* qualora esso non abbia alcun genitore.

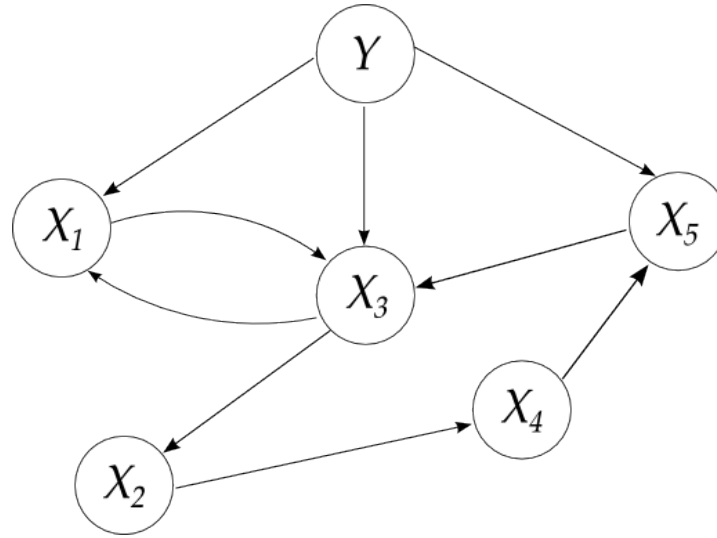


Figura 2.1: Un esempio di Continuous time Bayesian Network classifier (CTBNC) con cinque nodi attributo, X_1, \dots, X_5 , e un nodo classe, Y .

Parallelamente a quanto fatto in Langley *et al.* [1992], si presentano ora due istanze particolari di Continuous time Bayesian Network classifier.

Definizione 17 (Continuous time Naïve Bayes classifier). Un Continuous time Naïve Bayes classifier (CTNBC) è un Continuous time Bayesian Network classifier $\mathcal{C} = (\mathcal{N}, \mathbf{P}(Y))$ caratterizzato dal fatto che ogni nodo attributo ha un solo genitore, il nodo classe Y . Risulta quindi che:

$$\text{Pa}(X_i) = \{Y\} \quad \forall X_i \in \mathcal{G}.$$

Come mostrato dalla figura fig. 2.2 in questa pagina, un CTNBC possiede un nodo radice, associato alla variabile casuale Y , che è l'unico genitore di tutti i restanti nodi X_i (con $i = 1, 2, \dots, N$) che lo compongono. Si osservi come la rete di un CTNBC rappresenti l'assunzione di *indipendenza condizionale* di ogni nodo attributo dagli altri, data evidenza sulla variabile classe Y .

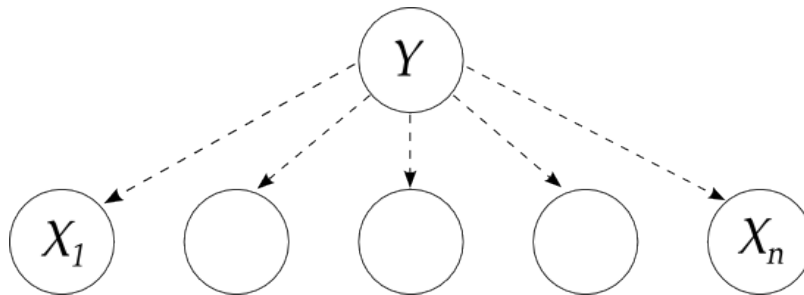


Figura 2.2: Un Continuous time Naïve Bayes classifier (CTNBC).

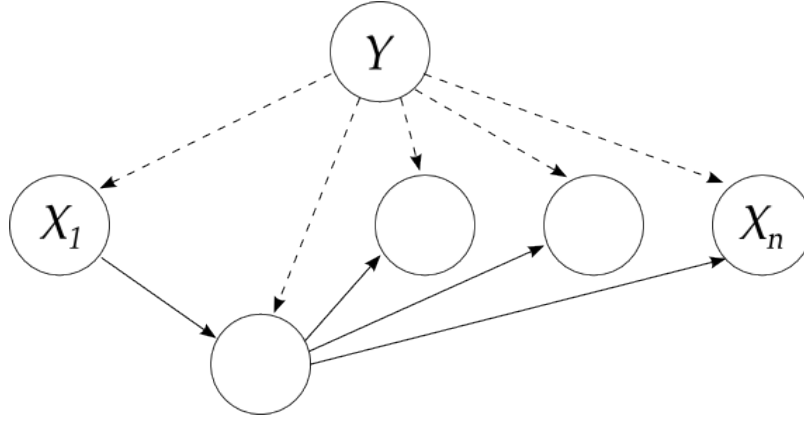


Figura 2.3: Un Continuous time tree augmented Naïve Bayes classifier (CTTANBC): qualora la variabile classe Y venga rimossa, le variabili rimanenti formano un albero.

Definizione 18 (Continuous time tree augmented Naïve Bayes classifier). Un Continuous time tree augmented Naïve Bayes classifier (CTTANBC) è un Continuous time Bayesian Network classifier $\mathcal{C} = (\mathcal{N}, \mathbf{P}(Y))$ che rispetta i seguenti vincoli:

- $Y \in \text{Pa}(X_i)$ con $i = 1, 2, \dots, N$
- i nodi attributo X_i , $i = 1, 2, \dots, N$, formano un albero:

$$\exists j : |\text{Pa}(X_j)| = 1 \quad \text{mentre per } i \neq j : |\text{Pa}(X_i)| = 2.$$

Come mostrato dalla figura 2.3 un classificatore CTTANB è un estensione del classificatore CTNB: tutti i nodi attributo della rete \mathcal{N} sono vincolati ad avere come genitore, oltre al nodo radice, al massimo un altro nodo attributo. Ciò comporta che tutti i nodi attributo facciano parte del *Markov blanket*⁴ del nodo radice associato con la variabile classe Y .

2.2 APPENDIMENTO

In questa sezione si affronta il problema dell'apprendimento (da *dati completi*) dei CTBNC.

Per definizione (si veda 16) i CTBNC sono basati sul modello delle CTBN, rappresentano perciò un insieme di modelli di probabilità locali (relativi alle variabili casuali) esprimibili in termini di statistiche sufficienti (per maggiori dettagli relativi a questo aspetto si rimanda alla sezione 1.4). Ne deriva che il problema dell'apprendimento di un classificatore CTBN si riduce alla computazione delle statistiche sufficienti dei suoi nodi attributo, da cui è successivamente possibile (nonché semplice) stimare i parametri (argomento trattato in dettaglio nella sottosezione 1.4.3) delle distribuzioni di probabilità codificate dalle matrici di intensità condizionali (CIM).

Di conseguenza, per l'apprendimento di un CTNBC è richiesto uno sforzo computazionale minimo. Per l'apprendimento di un CTTANBC, poiché questo modello prevede archi anche fra i nodi attributo, è invece richiesto uno sforzo computazionale leggermente maggiore [Stella e Amer, 2012].

Si presenta di seguito l'algoritmo 2.1 relativo all'apprendimento di un classificatore CTNB (definizione 17).

Esso richiede in input un *dataset completo* $\mathcal{D} = \{\delta_1, \dots, \delta_h\}$, il corrispondente insieme delle classi $\{y_1, y_2, \dots, y_h\}$, con $y_i \in \text{val}(Y)$, e il grafo \mathcal{G} di una CTBN \mathcal{N} (rispettivamente chiamati *data*, *classes* e *graph* nella firma della funzione `ctnbclearn`).

Per completezza si osservi (in base alla definizione 15) che ogni δ_i (con $i = 1, \dots, h$) è un flusso di evidenze $(x^1, x^2, \dots, x^{J_i})$ (a tal riguardo si veda la definizione 13).

Il risultato dell'applicazione dell'algoritmo 2.1 è un Continuous time Naïve Bayes classifier $\mathcal{C} = (\mathcal{N}, \mathbf{P}(Y))$.

Algoritmo 2.1: Apprendimento di un classificatore CTNB

```

1 function ctnbclearn(data, classes, graph) {
2   var h = len(data)
3   var klass = unique(classes)
4   var nk = len(klass)
5   var priors[nk]
6   for (i in index(data)) {
7     var k = index(classes[i])
8     priors[k] = priors[k] + (1 / h)
9   }
10  var m(), t()
11  for (i in index(data)) {
12    var y = classes[i]
13    var j = 1
14    while (tj ≤ Ti) {
15      for (n in index(graph.nodes)) {
16        m(xnj, xnj+1, y) = m(xnj, xnj+1, y) + 1
17        t(xnj, y) = t(xnj, y) + (tj - tj-1)
18      }
19      j = j + 1
20    }
21  }
22  var q(), thet()
23  foreach (y in klass) {
24    for (n in index(graph.nodes)) {
25      for (xn in val(Xn)) {
26        var mm(xn, y) =  $\sum_{x'_n \neq x_n} m(x_n, x'_n, y)$ 
27        q(xn, y) = mm(xn, y) / t(xn, y)
28        thet(xn, y) = m(xn, x'_n, y) / mm(xn, y)

```

```

29         }
30     }
31 }
32 var ctbn = new ctbn(graph, q, thet)
33 return (priors, ctbn)
34 }

```

L'algoritmo (2.1) di apprendimento appena presentato consiste nella stima delle matrici di intensità condizionali (CIM) di ogni variabile casuale di \mathcal{N} per ogni classe $y_i \in Y$. Più in dettaglio, esso è composto da tre fasi consecutive:

1. da linea 2 a linea 9 viene calcolata la *probabilità a priori* della variabile classe Y in base alla frequenza di ogni sua istanziazione $y_i \in Y$ in \mathcal{D}
2. da linea 10 a linea 21 vengono calcolate le *statistiche sufficienti* di ogni nodo attributo X_i , con $i = 1, 2, \dots, N$, sull'insieme di dati di apprendimento (anche detto *training set*) \mathcal{D}
3. da linea 22 a linea 31 vengono infine stimati, a partire dalle statistiche sufficienti, i *parametri maximum-likelihood* (MLE).

Si osservi che, poiché il processo di apprendimento è eseguito su un classificatore CTNB, l'algoritmo condiziona sia il calcolo delle statistiche sufficienti (i.e. variabili t e m) che la stima dei parametri (i.e. variabili q e thet) di ogni nodo attributo X_i solo ed esclusivamente al valore della variabile classe Y (i.e. variabile y). Questa semplificazione è dovuta al vincolo che caratterizza i classificatori CTNB: ogni nodo attributo X_i ha un solo genitore, il nodo associato alla variabile classe Y (si veda la definizione 17).

Come anticipato, al costo di un leggero incremento di complessità computazionale, è possibile estendere l'algoritmo 2.1 al fine di creare un algoritmo di apprendimento generale che apprenda un qualsiasi classificatore CTBN. Affinché tale obiettivo sia raggiunto è necessario rimuovere il succitato vincolo sull'insieme dei genitori di ogni nodo attributo. Mentre il calcolo della probabilità a priori della variabile classe Y non varia, il calcolo delle statistiche sufficienti e la stima dei parametri, invece, necessitano di tale generalizzazione.

Nello specifico:

1. il calcolo delle statistiche sufficienti di ogni nodo attributo X_i va condizionato all'istanziatura attuale (i.e. al tempo j) del suo insieme di nodi genitori $\text{Pa}(X_i)$; perciò a linea 15 dell'algoritmo 2.2 si prende in considerazione tale valore (i.e. variabile p)
2. la stima dei parametri maximum-likelihood (MLE) di ogni nodo attributo X_i va eseguita in base a ogni istanziazione del suo insieme di nodi genitori $\text{Pa}(X_i)$; perciò a linea 25 si itera in

base a ogni valore (i.e. variabile p) assunto da $\text{Pa}(X_i)$ (i.e. $\text{val}(\text{Pa}(X_i))$) mentre l'iterazione per classe non è più coerente e di conseguenza rimossa.

Si riporta di seguito l'algoritmo 2.2, il quale include le succitate modifiche finalizzate alla creazione di un algoritmo di apprendimento generale per i CTBNC.

Algoritmo 2.2: Apprendimento di un classificatore CTBN

```

1 function learn(data, classes, graph) {
2   var h = len(data)
3   var klass = unique(classes)
4   var nk = len(klass)
5   var priors[nk]
6   for (i in index(data)) {
7     var k = index(klass[i])
8     priors[k] = priors[k] + (1 / h)
9   }
10  var m(), t()
11  foreach (i in index(data)) {
12    var j = 1
13    while ( $t_j \leq T_i$ ) {
14      foreach (n in index(graph.nodes)) {
15        var p =  $\text{val}^j(\text{Pa}(X_n))$ 
16         $m(x_n^j, x_n^{j+1}, p) = m(x_n^j, x_n^{j+1}, p) + 1$ 
17         $t(x_n^j, p) = t(x_n^j, p) + (t_j - t_{j-1})$ 
18      }
19      j = j + 1
20    }
21  }
22  var q(), thet()
23  foreach (n in index(graph.nodes)) {
24    for ( $x_n$  in  $\text{val}(X_n)$ ) {
25      foreach (p in  $\text{val}(\text{Pa}(X_n))$ ) {
26         $\text{mm}(x_n, p) = \sum_{x'_n \neq x_n} m(x_n, x'_n, p)$ 
27         $q(x_n, p) = \text{mm}(x_n, p) / t(x_n, p)$ 
28         $\text{thet}(x_n, p) = m(x_n, x'_n, p) / \text{mm}(x_n, p)$ 
29      }
30    }
31  }
32  var ctbn = new ctbn(graph, q, thet)
33  return (priors, ctbn)
34 }

```

2.3 INFERENZA

In questa sezione si affronta il problema della *classificazione* di un *flusso di evidenze completamente osservato*, indicato con $(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J)$ (si veda la definizione 13), rispetto a un classificatore CTBN (CTBNC). Si osservi che, in tale situazione (i.e. *dati completi*), l'unica variabile casuale non osservata è la variabile classe, perciò è possibile sfruttare le relazioni di indipendenza fra variabili casuali così come si fa per le Bayesian Network. L'argomento di questa sezione è quindi il processo di *classificazione supervisionata* (i.e. è necessario un CTBNC appreso da un *training set*), di cui si presentano in primis le basi teoriche e successivamente l'implementazione algoritmica che ne consegue.

Il processo di classificazione di un flusso di evidenze è effettuato in base alla regola *maximum a posteriori*¹² (MAP) [si veda Stella e Amer, 2012]: un flusso di evidenze completamente osservato viene classificato assegnandogli la classe la cui probabilità a posteriori (rispetto al flusso di evidenze stesso) è massima. A tale scopo è necessario calcolare la probabilità a posteriori della variabile classe Y del CTBNC, rispetto al flusso di evidenze in input, per tutti i suoi possibili stati (i.e. classi, o etichette).

Il classificatore CTBN classifica quindi il flusso di evidenze massimizzando la seguente probabilità a posteriori, ricavata applicando la *regola di Bayes*:

$$P(Y | (\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J)) = \frac{P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J) | Y) P(Y)}{P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J))}. \quad (2.1)$$

Si specifica di seguito la semantica dei componenti dell'equazione 2.1:

- la probabilità marginale associata alla variabile classe Y

$$P(Y)$$

- la probabilità del flusso di evidenze

$$P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J))$$

- la *likelihood*⁸ del flusso di evidenze dato il valore della variabile classe, a cui ci si riferisce nel prosieguo usando l'espressione "*likelihood temporale*"

$$P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J) | Y).$$

¹² La stima della probabilità maximum a posteriori (MAP) è una moda della distribuzione a posteriori che può essere usata per ottenere una stima puntuale di una quantità inosservata sulla base di dati empirici. Può essere vista come una regolarizzazione della stima maximum-likelihood (MLE) poiché è strettamente correlata ad essa; vi differisce perché impiega un obiettivo di massimizzazione incrementato che incorpora una distribuzione a priori sopra la quantità che si vuole stimare.

La probabilità del flusso di evidenze, similmente a quanto accade per i BNC [Friedman *et al.*, 1997], non è richiesta per l'implementazione della regola MAP, perciò è possibile ometterla e riscrivere la precedente equazione sotto forma di relazione proporzionale:

$$P(Y | (\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J)) \propto P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J) | Y) P(Y). \quad (2.2)$$

Il primo termine dell'equazione 2.2, cioè la *likelihood temporale*, è invece fondamentale per la classificazione tramite regola MAP ed è possibile riformularlo nel seguente modo:

$$P((\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^J) | Y) = \prod_{j=1}^J P(\mathbf{x}^j | Y) P(\mathbf{x}^{j+1} | \mathbf{x}^j, Y), \quad (2.3)$$

dove:

- $P(\mathbf{x}^j | Y)$ rappresenta la probabilità che il *vettore aleatorio*¹³ \mathbf{X} resti nello stato \mathbf{x}^j durante l'intervallo temporale $[t_{j-1}, t_j)$ data evidenza sulla variabile classe Y
- $P(\mathbf{x}^{j+1} | \mathbf{x}^j, Y)$ rappresenta la probabilità che in \mathbf{X} si verifichi una transizione da \mathbf{x}^j a \mathbf{x}^{j+1} all'istante di tempo t_j data evidenza sulla variabile classe Y .

Inoltre, al fine di assicurare la consistenza dell'equazione 2.3 si assume che $P(\mathbf{x}^{J+1} | \mathbf{x}^J, Y) = 1$.

Il passo successivo consiste nel calcolo dei due termini da cui è composta l'equazione 2.3. A tal fine si utilizzano le distribuzioni di probabilità locali associate ad ogni nodo del CTBN su cui è costruito il classificatore. Come già descritto nella sottosezione 1.1.2, tali modelli di probabilità sono espressi tramite le matrici di intensità condizionali e quindi tramite i parametri \mathbf{q} e $\boldsymbol{\theta}$.

In tale contesto è quindi possibile calcolare il termine $P(\mathbf{x}^j | Y)$ come segue:

$$P(\mathbf{x}^j | Y) = \prod_{n=1}^N \exp\left(-\mathbf{q}_{\mathbf{x}_n^j}^{\mathbf{p}a_j(\mathbf{x}_n)} (t_j - t_{j-1})\right), \quad (2.4)$$

dove $\mathbf{q}_{\mathbf{x}_n^j}^{\mathbf{p}a_j(\mathbf{x}_n)}$ è il valore del parametro della *distribuzione esponenziale* quando la variabile casuale X_n è nello stato x_n durante il j -esimo intervallo temporale $[t_{j-1}, t_j)$ e contemporaneamente l'istanziamento dei genitori di X_n è $\mathbf{p}a_j(\mathbf{x}_n)$.

Uguualmente, il termine $P(\mathbf{x}^{j+1} | \mathbf{x}^j, Y)$ è così calcolabile:

$$P(\mathbf{x}^{j+1} | \mathbf{x}^j, Y) = \prod_{n=1}^N P(x_n^{j+1} | x_n^j, Y), \quad (2.5)$$

¹³ Un *vettore aleatorio* $\mathbf{X} = (X_1, X_2, \dots, X_n)$ è una *n-upla* (composta da n variabili casuali) i cui elementi sono dati da numeri aleatori.

dove

$$P(x_n^{j+1} | x_n^j, Y) = \begin{cases} q_{x_n^j x_n^{j+1}}^{pa_j(x)} & \text{se } x_n^j \neq x_n^{j+1} \\ 1, & \text{altrimenti} \end{cases}. \quad (2.6)$$

Il termine $P(x_n^{j+1} | x_n^j, Y)$ rappresenta la probabilità che nel vettore aleatorio \mathbf{X} si verifichi una transizione dallo stato x^j allo x^{j+1} , dato il valore della variabile classe Y .

Poiché il modello CTBN implica che, ad ogni istante t_j , solo un componente X_n del vettore aleatorio \mathbf{X} può essere soggetto a transizione, allora $P(x_n^{j+1} | x_n^j, Y)$ rappresenta la probabilità che la variabile casuale X_n effettui una transizione da x_n^j a x_n^{j+1} mentre tutti gli altri componenti del vettore aleatorio \mathbf{X} (i.e. X_i con $i \neq n$) non cambiano il proprio stato.

Perciò, come specificato dall'equazione 2.6, nel caso in cui avvenga un cambio di stato di X_n , il termine $P(x_n^{j+1} | x_n^j, Y)$ equivale alla quantità $q_{x_n^j x_n^{j+1}}^{pa_j(x)}$, ricavata dall'equazione 1.11, moltiplicando la “*probabilità istantanea*” di transizione da x_n^j a x_n^{j+1} e il tempo atteso di transizione uscente dallo stato x_n^j . Tale quantità rappresenta il parametro associato alla transizione da x_n^j , stato in cui la variabile casuale X_n si trovava durante il j -esimo intervallo temporale $[t_{j-1}, t_j)$, a x_n^{j+1} , stato in cui X_n si troverà durante il successivo intervallo temporale $[t_j, t_{j+1})$; data l'istanziatura $pa_j(x_n)$ dei genitori di X_n durante il j -esimo intervallo temporale.

Combinando l'equazione 2.4 e l'equazione 2.5 si ottiene:

$$P(x^j | Y) P(x^{j+1} | x^j, Y) = \prod_{n=1}^N \exp\left(-q_{x_n^j}^{pa_j(x_n)} (t_j - t_{j-1})\right) P(x_n^{j+1} | x_n^j, Y). \quad (2.7)$$

Utilizzando l'equazione 2.7 appena ricavata è possibile riformulare l'equazione della *likelihood temporale* (2.3) nel seguente modo:

$$P((x^1, x^2, \dots, x^J) | Y) = \prod_{j=1}^J \prod_{n=1}^N \exp\left(-q_{x_n^j}^{pa_j(x_n)} (t_j - t_{j-1})\right) P(x_n^{j+1} | x_n^j, Y). \quad (2.8)$$

Sostituendo infine l'equazione appena scritta (2.8) nell'equazione 2.2 si formula definitivamente la probabilità a posteriori della variabile classe Y dato un flusso di evidenze:

$$P(Y | (x^1, x^2, \dots, x^J)) \propto P(Y) \cdot \prod_{j=1}^J \prod_{n=1}^N \left[\exp\left(-q_{x_n^j}^{pa_j(x_n)} (t_j - t_{j-1})\right) \cdot P(x_n^{j+1} | x_n^j, Y) \right]. \quad (2.9)$$

Di conseguenza, dato un classificatore CTBN $\mathcal{C} = \{N, P(Y)\}$ e un flusso di evidenze completamente osservato (x^1, x^2, \dots, x^J) , la regola MAP seleziona la classe $y^* \in \text{val}(Y)$ massimizzando l'equazione 2.9:

$$y^* = \arg \max_{y \in \text{val}(Y)} P(Y) \prod_{j=1}^J \prod_{n=1}^N \exp\left(-q_{x_n^j}^{pa_j(x_n)} (t_j - t_{j-1})\right) P(x_n^{j+1} | x_n^j, Y) \quad (2.10)$$

Si presenta di seguito l'algoritmo per l'*inferenza esatta*) di un flusso di evidenze completamente osservato rispetto a un classificatore CTBN.

Tuttavia si osservi che tale algoritmo rappresenta le probabilità come *log-probabilità*¹⁴ per motivi computazionali. Ciò significa che l'algoritmo 2.3 implementa la probabilità a posteriori della classe dato un flusso di evidenze (equazione 2.9) come segue:

$$\ell_{P(Y|...)} = \log(P(Y)) + \sum_{j=1}^J \sum_{n=1}^N -q_{x_n^j}^{pa_j(x_n)} (t_j - t_{j-1}) + \log(P(x_n^{j+1} | x_n^j, Y)). \quad (2.11)$$

A tal proposito si noti che, nel caso in cui non avvenga alcun cambio di stato durante un determinato istante di tempo t_j , la quantità $P(x_n^{j+1} | x_n^j, Y)$ in 2.11 sarà pari a 1 (si veda l'equazione 2.6), il cui logaritmo è pari a 0. Ciò spiega la struttura di controllo condizionale alla linea 11.

Algoritmo 2.3: Inferenza su un classificatore CTBN

```

1 function infer(ctbnc, timeseg, stream) {
2     var priors = ctbnc.priors
3     var logp[len(priors)]
4     for (k in index(priors)) {
5         logp[k] = log(priors[k])
6     }
7     for (k in index(priors)) {
8         for (j in index(timeseg)) {
9             for (n in index(ctbnc.graph.nodes)) {
10                logp[k] = logp[k] -  $q_{x_n^j}^{pa_j(x_n)}$  * timeseg[j]
11                if ( $x_{n_j} \neq x_{n_{j+1}}$ ) {
12                    logp[k] = logp[k] + log( $q_{x_n^j x_n^{j+1}}^{pa_j(x)}$ )
13                }
14            }
15        }
16    }
17    return which(max(logp))
18 }
```

L'algoritmo 2.3 di inferenza esatta appena presentato è composto principalmente da due fasi corrispondenti ai due termini principali dell'equazione 2.11: da linea 2 a linea 6 converte la distribuzione della variabile classe Y del classificatore CTBN in forma logaritmica; da linea 7 a linea 16 aggiorna, incrementandola o decrementandola, il

¹⁴ La *log-probabilità* è un modalità di rappresentazione della probabilità che porta con sé alcuni vantaggi computazionali. Ad esempio, l'utilizzo della *log-probabilità* generalmente comporta una maggior velocità dovuta alla trasformazione delle moltiplicazioni, più computazionalmente costose, in addizioni. In informatica è molto comune l'utilizzo della sua variante negativa, la quale codifica un valore di probabilità $x \in [0, 1]$ come $x' = -\log(x) \in \mathbb{R}$.

valore di *log-probabilità* relativo a ogni classe (si veda il ciclo `for` alla linea 7) del classificatore CTBN iterando il flusso di evidenze per ogni segmento temporale cui sono associate le sue istanze (i.e. `variable timeseg`, che si assume venga fornita in input; si veda il ciclo `for` alla linea 8), infine iterando sui nodi del grafo \mathcal{N} del CTBNC (ciclo `for` a linea 9). Il passo finale dell'algoritmo, corrispondente alla linea finale (i.e. linea 17), consiste nella restituzione della classe la cui *log-probabilità* è maggiore di quella delle restanti classi, implementando così l'equazione 2.10.

Si osservi infine che l'algoritmo in esame implementa una conversione delle probabilità in forma logaritmica non negativa (i.e. $x \in [0, 1]$ rappresentato come $x' = \log(x)$). Ne consegue che l'applicazione della regola MAP non varia.

Anche l'algoritmo di inferenza esatta presentato in questa sezione è, così come l'algoritmo di apprendimento, computazionalmente efficiente.

3 | APPRENDIMENTO STRUTTURALE

Uno dei casi principali che costituisce il problema dell'*apprendimento* di modelli grafico probabilistici è l'apprendimento della struttura incognita sottostante ogni modello.

Il problema dell'*apprendimento strutturale* da *dati completi* di una CTBN è quindi l'argomento trattato in questo capitolo.

Generalmente, questo problema può essere informalmente descritto nel seguente modo: dato un training set composto da istanze di un insieme di variabili casuali si trovi un grafo che rappresenti tali dati e le relazioni fra le variabili casuali. Si noti come tale problema possa essere catalogato come una forma di apprendimento non supervisionato; nel senso che il processo di apprendimento non distingue la variabile classe dalle variabili attributo nei dati. L'obiettivo è quindi indurre una struttura (i.e. grafo) che descriva nel miglior modo possibile la distribuzione di probabilità sui dati (i.e. *training set*). Si osservi, inoltre, che questo problema di ottimizzazione è solitamente intrattabile per le Bayesian Network [Chickering, 1994]. Tuttavia, imponendo alcuni vincoli sulle strutture delle BN da includere nello spazio di ricerca, esistono algoritmi efficienti che risolvono tale problema in un tempo polinomiale rispetto al numero di variabili casuali rappresentate dalla struttura.

Per quanto riguarda invece il caso delle Continuous time Bayesian Network (CTBN), Nodelman *et al.* [2002] hanno dimostrato che, grazie alla mancanza del vincolo di aciclicità, come già accennato nella sezione 1.3), il problema dell'apprendimento strutturale di una CTBN è significativamente più facile, sia teoricamente che in pratica, rispetto all'apprendimento strutturale di una Bayesian Network, o di modelli da esse derivanti, e.g. le Dynamic Bayesian Networks (DBN). Inoltre, nel caso si vincoli la procedura di ricerca a strutture con un numero massimo di genitori per nodo, questo problema può essere risolto in tempo polinomiale.

L'approccio che si presenta in questo capitolo è quindi un approccio basato sul punteggio: si definisce una funzione che computa uno *score bayesiano* finalizzato alla valutazione di ogni struttura rispetto ai dati di addestramento e si usa una tecnica di ricerca euristica (ad esempio, la ricerca *hill climbing*) per cercare nello spazio delle strutture candidate quella che esibisce il maggior punteggio.

Si osservi che l'apprendimento dei parametri (si veda la "Stima bayesiana", sottosezione 1.4.3) è propedeutico per tale obiettivo poiché essi costituiscono la base dello *score bayesiano*.

3.1 FUNZIONE DI SCORING

Qualsiasi processo di apprendimento strutturale basato su punteggio è costituito da due componenti: una *funzione di scoring* e una procedura di ottimizzazione.

L'obiettivo di questa sezione è quindi presentare una funzione di scoring per l'apprendimento strutturale delle Continuous time Bayesian Network (CTBN). Lo scopo di tale funzione è calcolare il punteggio (i.e. lo *score*) di una struttura relativamente al *training set* \mathcal{D} fornito.

Si definisce lo *score bayesiano* sul grafo \mathcal{G} di una CTBN nel seguente modo:

$$\text{score}_B(\mathcal{G} : \mathcal{D}) = \ln P(\mathcal{D} | \mathcal{G}) + \ln P(\mathcal{G}) \quad (3.1)$$

Come mostra l'equazione 3.1 la funzione di scoring utilizza la probabilità a posteriori dell'insieme dei dati di apprendimento (i.e. il training set \mathcal{D}) data la struttura candidata (i.e. \mathcal{G}), oltre alla probabilità a priori della struttura stessa.

È possibile aumentare in modo significativo l'efficienza dell'algoritmo di ricerca che si affronta nella prossima sezione qualora si facciano determinate assunzioni. Nello specifico, se si assume che la probabilità a priori della struttura, $P(\mathcal{G})$, soddisfi la *modularità strutturale*, ne consegue:

$$P(\mathcal{G}) = \prod_{X_i} P(\text{Pa}(X_i) = \text{Pa}_{\mathcal{G}}(X_i)). \quad (3.2)$$

Se si assume, inoltre, che la probabilità a priori dei parametri soddisfi la *modularità dei parametri*, allora per ogni due strutture \mathcal{G} e \mathcal{G}' tali che $\text{Pa}_{\mathcal{G}}(X) = \text{Pa}_{\mathcal{G}'}(X)$ risulta:

$$P(\mathbf{q}_X, \boldsymbol{\theta}_X | \mathcal{G}) = P(\mathbf{q}_X, \boldsymbol{\theta}_X | \mathcal{G}'). \quad (3.3)$$

Combinando l'assunzione di *indipendenza dei parametri* con l'equazione 3.3 derivante dalla *modularità dei parametri*, si ottiene:

$$P(\mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}} | \mathcal{G}) = \prod_{X_i} \left[P(\mathbf{q}_{X_i} | \text{Pa}(X_i) | \text{Pa}(X_i) = \text{Pa}_{\mathcal{G}}(X_i)) \cdot P(\boldsymbol{\theta}_{X_i} | \text{Pa}(X_i) | \text{Pa}(X_i) = \text{Pa}_{\mathcal{G}}(X_i)) \right]. \quad (3.4)$$

Si osservi che, poiché la *penalità del grafo*, corrispondente al termine $P(\text{Pa}(X_i) = \text{Pa}_{\mathcal{G}}(X_i))$ dell'equazione 3.2, è legata alla dimensione del grafo ma indipendente dalla quantità dei dati, è possibile ignorare il termine $P(\mathcal{G})$ della funzione di scoring (equazione 3.1).

Di conseguenza l'unico termine significativo dell'equazione 3.1 è la *likelihood marginale*, $P(\mathcal{D} | \mathcal{G})$. Tale termine, infatti, codifica l'incertezza sui parametri integrando su tutti i possibili valori che essi possono assumere:

$$P(\mathcal{D} | \mathcal{G}) = \int_{\mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}}} P(\mathcal{D} | \mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}}) P(\mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}} | \mathcal{G}) d\mathbf{q}_{\mathcal{G}} d\boldsymbol{\theta}_{\mathcal{G}}. \quad (3.5)$$

Come per l'equazione 1.9, la likelihood marginale può essere decomposta come un prodotto di likelihood:

$$\begin{aligned}
 P(\mathcal{D} | \mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}}) &= \prod_{X_i} L_{X_i}(\mathbf{q}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) L_{X_i}(\boldsymbol{\theta}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) \\
 &= \underbrace{\left[\prod_{X_i} L_{X_i}(\mathbf{q}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) \right]}_{L(\mathbf{q} : \mathcal{D})} \underbrace{\left[\prod_{X_i} L_{X_i}(\boldsymbol{\theta}_{X_i | \text{Pa}(X_i)} : \mathcal{D}) \right]}_{L(\boldsymbol{\theta} : \mathcal{D})}.
 \end{aligned} \tag{3.6}$$

Combinando l'*indipendenza globale dei parametri* e la decomposizione ottenuta dall'equazione 3.6 si può riformulare la likelihood marginale (equazione 3.5) nel seguente modo:

$$\begin{aligned}
 P(\mathcal{D} | \mathcal{G}) &= \int_{\mathbf{q}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{G}}} L(\mathbf{q}_{\mathcal{G}} : \mathcal{D}) L(\boldsymbol{\theta}_{\mathcal{G}} : \mathcal{D}) P(\mathbf{q}_{\mathcal{G}}) P(\boldsymbol{\theta}_{\mathcal{G}}) d\mathbf{q}_{\mathcal{G}} d\boldsymbol{\theta}_{\mathcal{G}} \\
 &= \underbrace{\left[\int_{\mathbf{q}_{\mathcal{G}}} L(\mathbf{q}_{\mathcal{G}} : \mathcal{D}) P(\mathbf{q}_{\mathcal{G}}) d\mathbf{q}_{\mathcal{G}} \right]}_a \cdot \underbrace{\left[\int_{\boldsymbol{\theta}_{\mathcal{G}}} L(\boldsymbol{\theta}_{\mathcal{G}} : \mathcal{D}) P(\boldsymbol{\theta}_{\mathcal{G}}) d\boldsymbol{\theta}_{\mathcal{G}} \right]}_b.
 \end{aligned} \tag{3.7}$$

Ottenuta tale equazione, si affronta di seguito l'analisi e la decomposizione dei due termini che la compongono.

Utilizzando l'assunzione di *indipendenza locale dei parametri*, il termine termine a dell'equazione 3.7 è decomponibile nel seguente modo: ... Invece, poiché le distribuzioni sui parametri $\boldsymbol{\theta}$ sono di *Dirichlet* l'analisi del termine b dell'equazione 3.7 è analoga a quella che tradizionale per le Bayesian Network. Ne consegue che il termine b si semplifica: ...

3.2 RICERCA DELLA STRUTTURA

In questa sezione si affronta il secondo componente del processo di apprendimento strutturale: l'utilizzo di una *procedura di ottimizzazione* finalizzata a ricercare fra le strutture candidate quella che è caratterizzata dal massimo *score bayesiano*.

3.2.1 Hill Climbing

...

4

PACKAGE R

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.1 ANALISI

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

4.2 PACKAGE CTBN

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus.

Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5

CREAZIONE DI DATASET RELATIVI AL TRAFFICO

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

5.1 TSIS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

5.1.1 Descrizione

...

5.1.2 API

...

5.2 ESTENSIONE

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5.2.1 Analisi

...

5.2.2 Sensors DLL

...

5.3 APPLICATIVI DI SUPPORTO

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

6

ESPERIMENTI NUMERICI

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6.1 DATASET 1

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

6.1.1 Modello TSIS

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus

eu enim. Vestibulum pellentesque felis eu massa.

6.1.2 Risultati

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

6.2 DATASET 2

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

6.2.1 Modello TSIS

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

6.2.2 Risultati

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellente-

sque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

7 | CONCLUSIONI

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

A.1 UTILIZZO DEL PACKAGE CTBN

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

A.1.1 Caricamento del dataset

A.1.2 Calcolo delle sufficient statistics

A.1.3 Calcolo dei parametri

A.1.4 Calcolo delle CIM

A.1.5 Apprendimento

A.1.6 Classificazione

A.1.7 Apprendimento strutturale

A.1.8 Cross-validation

A.2 CREAZIONE DI DATASET

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt ur-

na. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

A.2.1 Sensors DLL

...

Installazione

...

Guida all'uso

...

A.2.2 Applicativi di supporto

...

ACRONIMI

BN	Bayesian Network	1
BNC	Bayesian Network classifier	19
CIM	Conditional Intensity Matrix	9
CPD	Conditional Probability Distribution	2
CPT	Conditional Probability Table	2
CTBN	Continuos time Bayesian Network	1
CTBNC	Continuos time Bayesian Network classifier	19
CTNB	Continuos time Naïve Bayes	19
CTNBC	Continuos time Naïve Bayes classifier	20
CTTANB	Continuos time tree augmented Naïve Bayes	19
CTTANBC	Continuos time tree augmented Naïve Bayes classifier	20
DAG	Directed acyclic graph	1
DBN	Dynamic Bayesian Networks	31
EM	Expectation Maximization	4
IC	Inductive Causation	6
IM	Intensity Matrix	8
MAP	Maximum a posteriori	26
MCMC	Markov Chain Monte Carlo	5
MLE	Maximum Likelihood Estimation	16
PV	Process Variable	11
TAN	Tree Augumented Naïve Bayes	20

INDICE ANALITICO

apprendimento

- aciclicità, 31
- dati completi, 31
- dati di addestramento, 31
- funzione di scoring, 32
- grafo, 32
- hill climbing, 31
- ottimizzazione, 33
- parametri, 31
- punteggio, 31, 32
- ricerca euristica, 31
- score bayesiano, 31–33
- struttura, 32
- strutturale, 31, 32
- training set, 31, 32

apprendimento automatico, 19

classificazione

- maximum a posteriori, 26
- apprendimento, 19
- classificatore, 19
- dati completi, 26
- inferenza, 19, 29
- likelihood temporale, 26, 27
- regola di Bayes, 26
- supervisionata, 19
- training set, 24
- vettore aleatorio, 27, 28

classificazione, 2, 26

conteggi immaginari, 18

definizione

- segmenti temporali, 11

distribuzione

- Dirichlet, 17, 33
- esponenziale, 8, 15–17
- Gamma, 17
- multinomiale, 9, 16, 17
- priori coniugata, 17

machine learning, 1, 19

statistiche sufficienti, 22

stime esatte, 16

BIBLIOGRAFIA

Chickering, David Maxwell

- 1994 *Learning Bayesian networks is NP-hard*, rapp. tecn., Microsoft Research. (Citato a p. 31.)
- 2013 «A Transformational Characterization of Equivalent Bayesian Network Structures», *CoRR*, p. 87-98, <http://arxiv.org/abs/1302.4938>. (Citato a p. 6.)

Dempster, A P, N M Laird e D B Rubin

- 1977 «Maximum likelihood from incomplete data via the EM algorithm», *Journal of the Royal Statistical Society Series B Methodological*, Series B, 39, 1, p. 1-38, ISSN: 00359246, DOI: 10.2307/2984875, <http://www.jstor.org/stable/2984875>. (Citato a p. 5.)

Duda, R. O. e P. E. Hart

- 1973 *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York. (Citato a p. 19.)

Friedman, N, D Geiger e M Goldszmidt

- 1997 «Bayesian network classifiers», *Machine learning*, 163, p. 131-163, <http://link.springer.com/article/10.1023/A%3A1007465528199>. (Citato alle p. 19, 27.)

Geman, Stuart e Donald Geman

- 1984 «Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images», *IEEE Trans. Pattern Anal. Mach. Intell.*, 6, 6, p. 721-741, ISSN: 0162-8828, DOI: 10.1109/TPAMI.1984.4767596, <http://dx.doi.org/10.1109/TPAMI.1984.4767596>. (Citato a p. 5.)

Gilks, WR, S Richardson e DJ Spiegelhalter

- 1996 «Markov chain Monte Carlo in practice». (Citato a p. 5.)

Heckerman, David

- 1996 «A Tutorial on Learning With Bayesian Networks», *Innovations in Bayesian Networks*, Studies in Computational Intelligence, 1995, November, a cura di Dawn E Holmes e Lakhmi C Jain, p. 33-82, ISSN: 1860949X, DOI: 10.1007/978-3-540-85066-3, <http://www.springerlink.com/index/62mv333389016034.pdf>. (Citato alle p. 5, 17.)

Heckerman, David, Dan Geiger e David M. Chickering

- 1995 «Learning Bayesian networks: The combination of knowledge and statistical data», *Machine Learning*, 20, 3 [set. 1995], p. 197-243, ISSN: 0885-6125, DOI: 10.1007/BF00994016, <http://link.springer.com/10.1007/BF00994016>. (Citato alle p. 6, 17.)

Korb, K.B. e A.E. Nicholson

- 2011 *Bayesian Artificial Intelligence*, Chapman & Hall / CRC Computer Science and Data Analysis, CRC PressINC, ISBN: 9781439815915. (Citato alle p. 1, 3.)

Langley, Pat, Wayne Iba e Kevin Thompson

- 1992 «An Analysis of Bayesian Classifiers», in *AAAI*, p. 223-228, ISBN: 0-262-51063-4. (Citato alle p. 19, 21.)

Loève, Michel

- 1978 *Probability theory. II Edition*. Fourth, Graduate Texts in Mathematics, Vol. 46, Springer-Verlag, New York, p. xvi+413, ISBN: 0-387-90262-7. (Citato alle p. 6, 7.)

MacKay, D. J. C.

- 1998 «Introduction to Monte Carlo methods», in *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, Kluwer Academic Publishers, Norwell, MA, USA, p. 175-204, <http://dl.acm.org/citation.cfm?id=299068.299077>. (Citato a p. 5.)

Nodelman, Uri, CR Shelton e Daphne Koller

- 2002 «Learning continuous time Bayesian networks», *Proceedings of the Nineteenth ...*, X, arXiv:/arxiv.org/abs/1212.2498 [http:], <http://dl.acm.org/citation.cfm?id=2100639>. (Citato alle p. 14, 15, 31.)

Nodelman, Uri D.

- 2007 *Continuous Time Bayesian Networks*, tesi di dott., Stanford University. (Citato alle p. 11, 15, 17.)

Norris, James R.

- 1998 *Markov chains*, Cambridge series in statistical and probabilistic mathematics, Cambridge University Press, p. I-XVI, 1-237, ISBN: 978-0-521-48181-6. (Citato a p. 7.)

Pearl, Judea

- 1988 *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN: 0-934613-73-7. (Citato a p. 5.)

Russell, Stuart J. e Peter Norvig

- 2003 *Artificial Intelligence: A Modern Approach*, Pearson Education, ISBN: 0137903952, <http://portal.acm.org/citation.cfm?id=773294>. (Citato alle p. 2, 3.)

Shachter, Ross D. e Mark A. Peot

- 1990 «Simulation Approaches to General Probabilistic Inference on Belief Networks», in *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '89, North-Holland Publishing Co., Amsterdam, The Netherlands, p. 221-234, ISBN: 0-444-88738-5, <http://dl.acm.org/citation.cfm?id=647232.719570>. (Citato a p. 5.)

Stella, F e Y Amer

- 2012 «Continuous time Bayesian network classifiers.», *Journal of biomedical informatics*, 45, 6 [dic. 2012], p. 1108-19, ISSN: 1532-0480, DOI: 10.1016/j.jbi.2012.07.002, <http://www.ncbi.nlm.nih.gov/pubmed/22846170>. (Citato alle p. 9, 11, 20, 23, 26.)

Verma, Thomas S. e Judea Pearl

- 1991 «Equivalence and synthesis of causal models», in *Uncertainty in Artificial Intelligence*, North Holland, p. 255-268. (Citato a p. 6.)

DICHIARAZIONE

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices.

Milano, settembre 2013

Leonardo Di Donato