Parameter Learning of a Bayesian Network

Luigi Ariano

Relazione del progetto

1 Introduzione

Nella seguente relazione viene descritto il funzionamento dell' elaborato che è stato sviluppato per l' esame. In particolare, l' elaborato inerente l' apprendimento dei parametri in reti Bayesiane utilizzando l' approccio Bayesiano descritto in (Heckerman 1997). Nella seconda parte, si generano dei data sets di **n** esempi a partire da una rete nota che possiede una distribuzione **p**, campionando i dati dalla rete.
Inizialmente, infatti, viene prodotto un data set di dimensione crescente, composto da **n** righe e da un numero di colonne corrispondenti al numero di nodi presenti nella rete, considerando le probabilità condizionate che sono state prese dalla rete utilizzata.
Il passo successivo è quello di effettuare l' apprendimento dei parametri, il quale restituisce una distribuzione appresa su un data set generato dal campionamento dei dati della rete di dimensione **n**, detto $\mathbf{q_n}$.
Infine, viene misurata la distanza tra la distribuzione delle probabilità **p** fornite inizialmente dalla rete e la distribuzione dei parametri appresa $\mathbf{q_n}$ tramite la divergenza di Jensen-Shannon, la quale deve tendere a 0 al crescere delle dimensioni del data set.

2 La rete Metastatic Cancer

La rete utilizzata per il progetto è un semplice modello di rete Bayesiana per l' importante problema di un **cancro metastatico**. La sua struttura è sostanzialmente basata su un DAG, ovvero un grafo aciclico diretto: possiede 5 nodi e 5 archi e ogni nodo rappresenta una variabile casuale, in pratica un evento, con la relativa probabilità che questo si verifichi o meno. Ogni arco descrive una sorta di condizione di dipendenza: un nodo figlio possiede la probabilità che l' evento si verifichi in base all' accadere o meno degli eventi rappresentati dai nodi padri, invece, i nodi non connessi rappresentano gli eventi che sono condizionalmente indipendenti dagli altri.
La struttura della rete Bayesiana e i dati campionati da questa sono visibili nel repository del mio elaborato. Comunque, di seguito, è stata inserita la sua rappresentazione grafica:

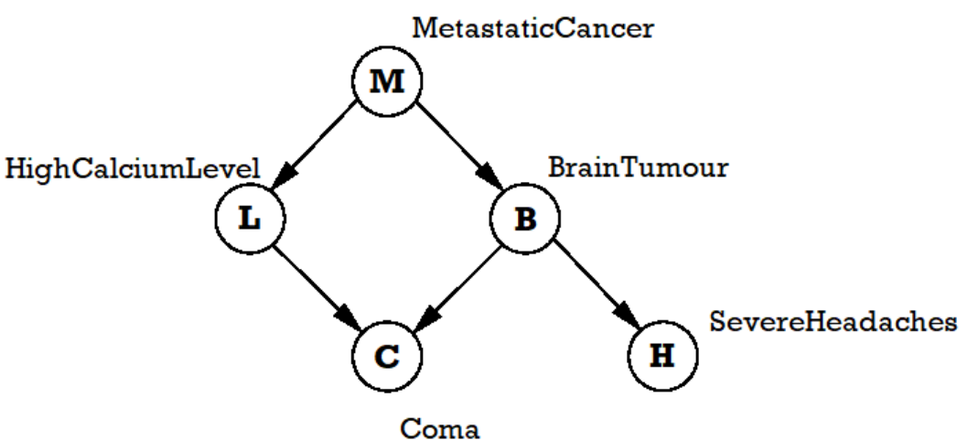

Figura 1: La rete Metastatic Cancer.

3 Ordinamento topologico

Il primo passo è dato dall' ordinamento topologico dei nodi: è infatti necessario avere un ordine degli eventi accaduti, per poter conoscere, tramite l' osservazione del comportamento di eventuali nodi padri, la possibilità con cui i figli si sono potenzialmente verificati.
L' ordinamento è eseguito tramite la visita in profondità, la quale restituisce i tempi di terminazione di ogni nodo. In uscita viene emesso il vettore dei vertici in ordine di tempo di fine decrescente.
Questo passo risulta necessario per la generazione del data set, infatti, vengono controllate le configurazioni dei nodi genitori per accertare che gli eventi rappresentati accadano o meno e, quindi, utilizzare la corretta probabilità con cui i nodi figli si verifichino.

4 Generazione del data set

La generazione del data set ha come primo step la differenziazione di due casistiche di nodi: i nodi che possiedono uno o più genitori e i nodi, invece, senza genitori.
Questi ultimi costituiscono una maggiore facilità nella realizzazione del data set: si confronta semplicemente la probabilità del nodo con un numero random tra 0 e 1, inserendo uno dei due estremi nel data set in base all' esito del confronto (0 o 1 se l' esito ha avuto successo o meno).
Nell' altro caso, invece, bisogna considerare l' eventuale accadimento degli eventi rappresentati dai nodi genitori rispetto al vertice considerato, rendendo quindi necessario l' ordinamento topologico dei nodi della rete. Successivamente, viene nuovamente eseguito il confronto tra un numero random compreso tra 0 e 1 e la probabilità condizionata trovata del nodo in questione.
Il data set generato avrà **n** righe pari al numero di prove effettuate e come numero di colonne (**5**) pari al numero di nodi della rete.

5 Apprendimento dei parametri

L' esecuzione dell' apprendimento dei parametri nelle reti Bayesiane richiede prima una serie di assunzioni da fare:

- $\mathbf{X} = \{X_1, ..., X_n\}$ costituisce l' insieme di tutte le variabili casuali che rappresentano gli eventi considerati nella rete Bayesiana con le relative probabilità.
- $\mathbf{D} = \{X_1{=}x_1, ..., X_n{=}x_n\}$ costituisce l' insieme del data set, ovvero, l' insieme di tutte le prove i.i.d. effettuate nella verifica dell' avvenimento o meno di tutti gli eventi di **X**. L' insieme **D** generato deve essere completo, cioè senza dati mancanti.
- Ogni variabile $X_i \in \mathbf{X}$ è discreta, possiede r_i possibili valori ${x_i}^1, ..., {x_i}^{r_i}$ ed ogni funzione di distribuzione è un insieme di distribuzioni multinomiali, una per ogni possibile configurazione dei genitori:
  $p({x_i}^k \mid {pa_i}^j, \theta_i, G) = \theta_{ijk} > 0$, dove ${pa_i}^j$ indica la j-esima configurazione dei padri del nodo i e G la struttura della rete Bayesiana.
- I vettori dei parametri $\theta_{ij} = (\theta_{ij1}, ..., \theta_{ijr_i})$ sono mutualmente indipendenti per ogni i, j e per ogni possibile configurazione del nodo i-esimo, detto r_i.
- Ogni vettore θ_{ij} ha come prior la distribuzione di Dirichlet $Dir(\theta_{ij} \mid \alpha_{ij1}, ..., \alpha_{ijr_i})$, dalla quale otteniamo la distribuzione posteriore $p(\theta_{ij} \mid D, G) = Dir(\theta_{ij} \mid \alpha_{ij1}{+}N_{ij1}, ..., \alpha_{ijr_i}{+}N_{ijr_i})$. I termini α_{ijk} sono spesso indicati come iperparametri per distinguerli dal parametro θ, mentre gli N_{ijk} sono il numero di volte in cui $X_i = {x_i}^k$ e $Pa_i = {pa_i}^j$ dove k e j dipendono da i.

  Tutte le assunzioni fatte dei criteri elencati sono descritte nel link precedente.

  Successivamente, dopo aver anche generato il data set, si può procedere con l' effettivo apprendimento dei parametri, stimandoli tramite la seguente formula:

$$\theta_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

  dove gli $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ e gli $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Nel programma, quindi, con la formula precedente viene appresa la distribuzione sul data set di dimensione **n**, detta $\mathbf{q_n}$, usando come priors pseudo-counts unitari (**Laplace Smoothing**).

6 Divergenza di Jensen-Shannon

La parte finale del progetto consiste nel calcolare la distanza tra la distribuzione iniziale data di probabilità **p** e la distribuzione dei parametri appresa $\mathbf{q_n}$ dalla rete utilizzando la divergenza di Jensen-Shannon, definita come:

$$JS(p, q_{\mathrm{n}}) = \sum_U p(U) \log \frac{p(U)}{\frac{p(U) + q_{\mathrm{n}}(U)}{2}} + \sum_U q_{\mathrm{n}}(U) \log \frac{q_{\mathrm{n}}(U)}{\frac{p(U) + q_{\mathrm{n}}(U)}{2}}$$

7 Risultati

Nel programma principale viene creato il vettore di tutte le probabilità che sono state campionate dalla rete, la matrice di adiacenza che rappresenta il grafo aciclico diretto e la rete Bayesiana che è stata considerata: **Metastatic Cancer**.
Le prove sono state eseguite con un numero crescente **n** di righe del data set, partendo con un numero di 10 fino ad arrivare ad un massimo di 5010 con passo 100. Per ogni **n** sono state eseguite un numero di 50 iterazioni per le generazioni dei data sets, la distribuzione dei parametri appresa $\mathbf{q_n}$ sul precedente e la divergenza di Jensen-Shannon. Dopo l' esecuzione di questi tentativi, ne viene eseguita la media per evitare dei casi particolari e ottenere dei risultati attendibili.
Gli esiti ottenuti sono visualizzabili tramite il grafico sottostante e tramite la tabella contente i risultati numerici salvati.

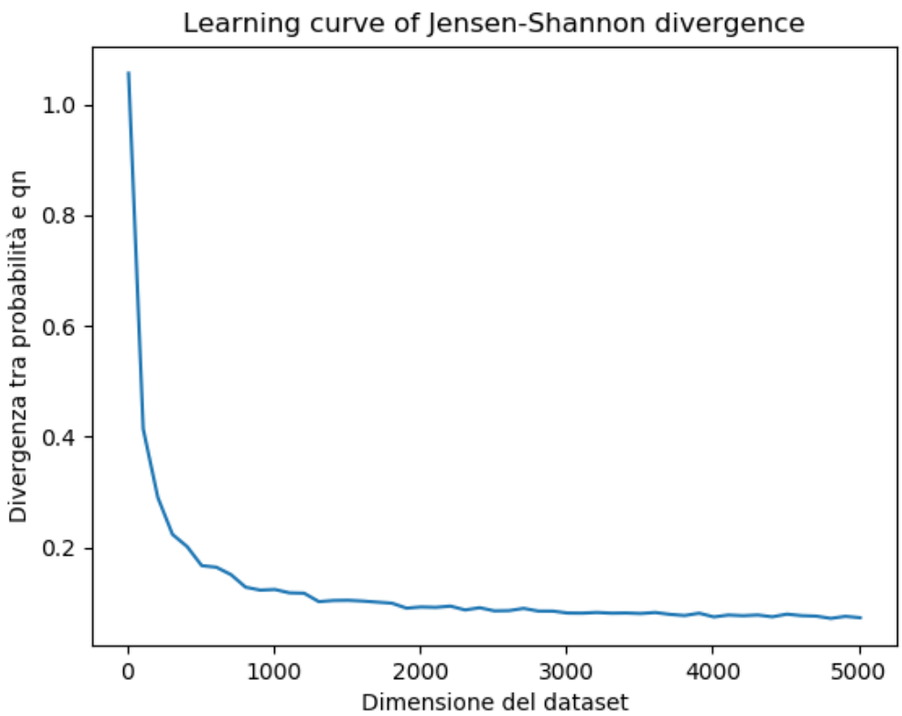

Figura 2: Divergenza di Jensen-Shannon.

Dimensione del data set	**Divergenza tra p e $\mathbf{q_n}$**
10	1.069
110	0.404
210	0.293
310	0.239
410	0.174
510	0.160
610	0.157
710	0.133
810	0.130
910	0.127
1010	0.125
1510	0.099
2010	0.090
2510	0.087
3010	0.083
3510	0.079
4010	0.076
4510	0.075
5010	0.071

Tabella 1: Risultati ottenuti dalla divergenza di Jensen-Shannon.

8 Conclusione

In conclusione, il programma ha un corretto funzionamento: come si può notare dal grafico e dalla tabella, al crescere della dimensione del data set, la divergenza di Jensen-Shannon tra la distribuzione delle probabilità iniziali $\mathbf{p}$ e la distribuzione appresa sul data set di dimensione $\mathbf{n}$, $\mathbf{q_n}$, diminuisce velocemente tendendo a 0. Questo risultato implica che la distribuzione appresa $\mathbf{q_n}$ dalla rete, si avvicina molto alla distribuzione di probabilità iniziale $\mathbf{p}$ ottenendo, quindi, una riduzione della distanza notevole tra le distribuzioni già dopo pochissime iterazioni. Infatti, dopo che il data set ha raggiunto la dimensione di 1010 elementi, non sono stati memorizzati nella tabella tutti i passi delle iterazioni per la poca rilevanza captata tra i risultati ottenuti. Questa rapida riduzione della distanza tra le distribuzioni $\mathbf{p}$ e $\mathbf{q_n}$ è dipesa molto anche dalla rete scelta, poiché una maggiore grandezza e complessità del network comporta più operazioni nella generazione del data set, nell' apprendimento dei parametri e nel calcolo della divergenza generando una learning curve simile, ma che tende più lentamente verso 0.