Parameter Learning in Bayesian Network

Luca Leuter

Gennaio 2020

1 Introduzione

Nella seguente relazione verrà illustrato il funzionamento di un software per l'apprendimento di parametri in reti Bayesiane. Essi verranno appresi da un dataset generato a partire dalle probabilità condizionate dei nodi, fornite insieme alla rete Bayesiana.
Viene infine misurata la distanza tra la distribuzione iniziale delle probabilità e la distribuzione dei parametri appresa nella prima parte tramite la divergenza di Jensen-Shannon, e i risultati mostrati su un grafico.

2 Cenni Teorici

Di seguito verranno descritte le procedure teoriche alla base del progetto del software.

2.1 Dataset

Per **generare il dataset** si tiene conto del fatto che ogni valore dei nodi (tranne per quelli che non hanno genitori) è condizionato dal valore dei suoi nodi genitori. Bisogna quindi eseguire un ordinamento topologico dei nodi, per garantire la coerenza con le dipendenze funzionali della rete. Il software per eseguire l'ordinamento topologico esegue una visita in profondità dei nodi, salvando i tempi di fine scoperta di ognuno, con il quale poi vengono ordinati. Il fatto che le reti Bayesiane siano dei DAG (grafi diretti aciclici) garantisce la presenza di almeno un ordinamento topologico. Successivamente viene creato il dataset utilizzando il file delle probabilità condizionate fornito. Si assume che tutte le righe siano indipendenti ed identicamente distribuite, in modo tale da poter ripetere la procedura un numero arbitrario di volte per formare un dataset.

2.2 Assunzioni

Vengono fatte delle assunzioni in modo che l'apprendimento sia eseguito in modo efficiente

- I nodi sono variabili discrete, ed ogni funzione di distribuzione è un insieme di distribuzioni multinomiali, una per ogni possibile configurazione dei genitori. Cioè si ha:

  $$p(x_i^k|Pa_i^j, \theta_i, G) = \theta_{ijk}$$

  dove Pa_i^j indica la j-esima configurazione dei padri del nodo i e G la struttura della rete Bayesiana.

- Il dataset D generato non ha dati mancanti, cioè è **completo**

- Dati $\theta_{ij} = (\theta_{ij1}...\theta_{ijr_i})$ con r_i possibili configurazioni della variabile i, si ha che essi sono **mutualmente indipendenti**, e lo rimangono anche dato un dataset

2.3 Parameter Learning

Generato il dataset si procede all'apprendimento dei parametri. Con le assunzioni fatte si può stimare i parametri con l'uso della **Massima Verosomiglianza (ML)**. La ML per θ_{ijk} è:

$$\widehat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

dove α_{ijk} sono gli iperparametri e gli $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$
mentre gli N_{ijk} sono il numero di volte in cui $X_i = x_i^k$ e $Pa_i = pa_i^j$. In tale relazione sono stati usati come priors pseudo-counts unitari *(Laplace Smoothing)*.

2.4 Jensen-Shannon Divergency

La parte finale del progetto consiste nel misurare la distanza tra la distribuzione iniziale p e la distribuzione dei parametri trovata q_n utilizzando la divergenza di Jensen-Shannon, definita come:

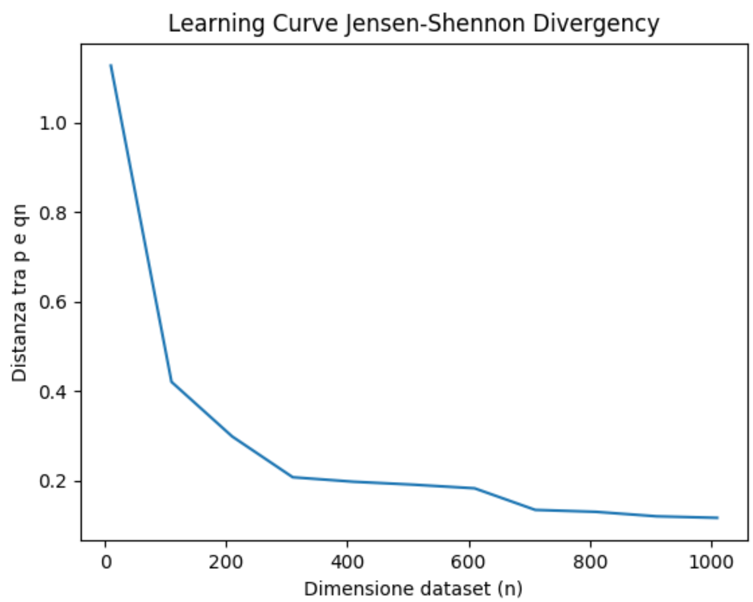
$$JS(p, q_n) = \sum_U p(U) log(\frac{p(U)}{\frac{p(U)+q_n(U)}{2}} + \sum_U q_n(U) log(\frac{q_n(U)}{\frac{p(U)+q_n(U)}{2}}$$

3 Esperimenti

Per eseguire gli esperimenti è stata utilizzata una rete Bayesiana di nome CANCER reperibile al seguente link. La rete contiene 5 nodi e 4 archi. Gli esperimenti sono stati eseguiti con un numero crescente **n** di righe del dataset, a partire da 10 fino a 1010 con passo 100. Per ogni **n** sono state eseguite 10 generazioni del dataset, per evitare casi particolari, di cui poi è stata fatta la media dei valori della divergenza di Jensen-Shennon ottenuti da ciascuno.

3.1 Risultati

I valori sopra usati si sono rivelati utili ai fini della soluzione. Viene riportato un risultato di un esperimento sottoforma di grafo e tabulare



Grandezza Dataset	Distanza tra p e q_n
10	1.127
110	0.420
210	0.298
310	0.207
410	0.197
510	0.190
610	0.182
710	0.134
810	0.129
910	0.119
1010	0.116

4 Conclusioni

Si può trarre come conclusioni che il software funziona bene poichè, come si può notare dal grafico e dalla tabella, la divergenza di Jensen-Shannon tra p e q_n si riduce all'aumentare di n tendendo a 0. Ciò vuol dire che la distribuzione dei parametri trovata dal software è molto simile alla distribuzione iniziale, e lo diventa sempre di più con l'aumentare della dimensione del dataset