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 in un file di testo, e dalla rete Bayesiana fornita sottoforma di matrice di adiacenza in un altro file di testo.
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 e in forma tabulare.

2 Cenni Teorici

Di seguito verranno descritte le procedure teoriche alla base del progetto del software, in particolare verranno presentati in ordine di esecuzione:

- Ordinamento Topologico
- Generazione del Dataset
- Apprendimento dei Parametri

Ogni esecuzione è condizionata dalla precedente

2.1 Ordinamento Topologico

E' necessario eseguire un ordinamento topologico dei nodi, per garantire la coerenza con le dipendenze funzionali della rete durante la generazione del dataset. 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. L'ordinamento può non essere unico.

2.2 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, rendendo necessario l'uso dell'ordinamento topologico. Viene quindi creato il dataset utilizzando il file delle probabilità condizionate fornito, generando un valore casuale tra 0 e 1 e confrontandolo con la probabilità condizionata del nodo in esame data l'eventuale configurazione dei genitori.

2.3 Assunzioni per il learning

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 i vettori $\theta_{ij} = (\theta_{ij1}...\theta_{ijr_i})$ con r_i possibili configurazioni della variabile i, si ha che essi sono **mutualmente indipendenti**, e lo rimangono anche dato un dataset.
- Inoltre assumendo che ognuno di essi abbia come prior la distribuzione di Dirichlet $Dir(\theta_{ij}|\alpha_{ij1}, ..., \alpha_{ijr_i})$ si ottiene come posterior la distribuzione: $p(\theta_{ij}|D, G) = Dir(\theta_{ij}|\alpha_{ij1} + N_{ij1}, ..., \alpha_{ijr_i} + N_{ijr_i})$ dove α_{ijk} sono gli iperparametri e gli N_{ijk} sono il numero di volte in cui $X_i = x_i^k$ e $Pa_i = pa_i^j$.

2.4 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 gli $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ e gli $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. In tale relazione sono stati usati come priors pseudo-counts unitari*(Laplace Smoothing)*.

2.5 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:

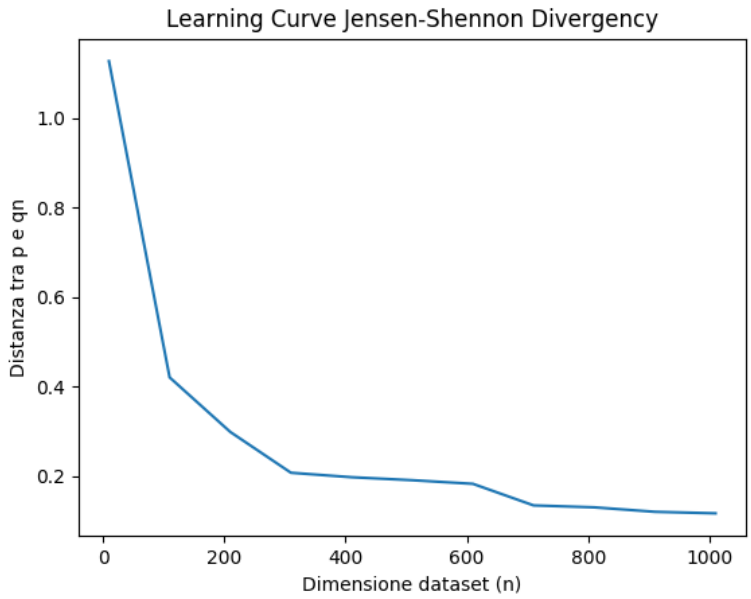
$$JS(p, q_n) = \sum_U p(U) log \frac{p(U)}{\frac{p(U)+q_n(U)}{2}} + \sum_U q_n(U) log \frac{q_n(U)}{\frac{p(U)+q_n(U)}{2}}$$

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 poiché, 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 con un andamento esponenziale.
Si nota anche che il guadagno in termini di distanza è maggiore per n compresi fra 0 e 400, mentre per $n > 400$ diventa meno significativo. Ciò dipende dalla grandezza della rete. Infatti più la rete è grande e più è necessario un dataset grande per apprendere più correttamente i parametri e dunque ridurre la divergenza di Jensen-Shannon, mentre per reti più piccole bastano meno valori.