Modeling and Reasoning with Bayesian Networks: Compiling Bayesian Networks

Relatório semana 1 - MAC0215 (Atividade Curricular em Pesquisa)
Aluno: Renato Lui Geh (Bacharelado em Ciência da Computação)
Orientador: Denis Deratani Mauá

1 Atividades realizadas na semana

Durante a semana foram lidos os seguintes tópicos do livro *Modeling and Reasoning with Bayesian Networks*[1]:

12 - Compiling Bayesian Networks

- **12.1** - Introduction
- **12.2** - Circuit semantics
- **12.3** - Circuit propagation
  - **12.3.1** - Evaluation and differentiation passes

2 Definição das atividades

Foram estudados o processo de se compilar Redes Bayesianas em circuitos aritméticos, algumas notações usadas em Redes Bayesianas, a definição de uma *network polynomial*, algumas propriedades de Redes Bayesianas e diferenciação de uma rede a partir de uma evidência.

Esta seção será dividida em subseções para cada subtópico citado na seção anterior. Cada subseção é um resumo do que foi estudado em cada tópico, contendo os assuntos mais importantes para o projeto.

2.1 Introduction

Aqui apresentaremos algumas notações usadas em Redes Bayesianas - e que podem ser extendidas para outros Modelos Gráficos Probabilísticos (PGM).

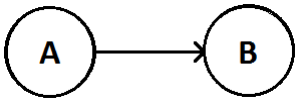

Figura 1: Uma Rede Bayesiana $A \to B$. Em Redes Bayesianas uma aresta representa uma dependência. No caso da imagem B depende de A. As CPTs associadas a esse grafo estão em Tabela 1 e Tabela 2.

Tabela 1 e Tabela 2

A	Θ_A
true	$\theta_a = 0.3$
false	$\theta_{\overline{a}} = 0.7$

A	B	$\Theta_{B\|A}$
true	true	$\theta_{b\|a} = 0.1$
true	false	$\theta_{\overline{b}\|a} = 0.9$
false	true	$\theta_{b\|\overline{a}} = 0.8$
false	false	$\theta_{\overline{b}\|\overline{a}} = 0.2$

Antes de começarmos com circuitos aritméticos, vamos primeiro apresentar algumas definições importantes.

Chamamos de CPT (Conditional Probability Table) as tabelas que representam as probabilidades de uma rede (ex.: as CPTs da Rede Bayesiana na Figura 1 são as Tabelas 1 e 2). Pode-se claramente ver que para n nós de uma Rede Bayesiana, precisamos de uma quantidade exponencial 2^n de probabilidades para representar cada instância de variáveis.

Chamamos de MPE (Most Probable Explanation) a instância mais provável de variáveis que aceitam uma evidência. Mais formalmente dizemos que:

Definição. *Sejam $X_1, ..., X_n$ as variáveis da rede e e a evidência dada. Existe uma instância $x_1, ..., x_n$ onde $Pr(x_1, ..., x_n|e)$ é maximal. Chamamos $x_1, ..., x_n$ a explicação mais provável (most probable explanation) dado e.*

Chamamos de MAP (Maximum A Posteriori hypothesis) quando a probabilidade de uma certa instância é maximal.

Definição. *Sejam X o conjunto de todas as variáveis da rede e M um subconjunto qualquer dessas variáveis. Dado uma evidência e, qualquer instância m de variáveis M onde $Pr(m|e)$ é maximal é uma hipótese máxima a posteriori (maximum a posteriori hypothesis).*

M também é chamado de variáveis MAP. MPE é uma MAP quando as variáveis MAP incluem todas as variáveis da rede.

Agora que temos uma base podemos voltar para circuitos aritméticos. Dado um circuito aritmético que representa uma Rede Bayesiana, teremos dois tipos de entradas: variáveis θ, que chamaremos de *parâmetros*, e as variáveis λ, que chamaremos de *indicadores*. Parâmetros são valorados de acordo com a CPT da rede, enquanto indicadores são valorados de acordo com a evidência. Nas próximas subseções veremos que podemos ter duas passagens pelo circuito. Uma em que vamos de baixo para cima (bottom-up) para calcular a probabilidade de uma dada evidência, e outra em que vamos de cima para baixo (top-down), chamada de *differentiation pass* para calcularmos as derivadas parciais de cada entrada do circuito.

2.2 Circuit semantics

Ao compilarmos um circuito aritmético de uma Rede Bayesiana estamos representando, de forma compacta, a distribuição de probabilidade induzida da rede. No caso da Figura 1, a distribuição de probabilidade está representada na Tabela 3.

Tabela 3

A	B	$Pr(A, B)$
a	b	$\theta_a \theta_{b\|a}$
a	$\overline{b}$	$\theta_a \theta_{\overline{b}\|a}$
$\overline{a}$	b	$\theta_{\overline{a}} \theta_{b\|\overline{a}}$
$\overline{a}$	$\overline{b}$	$\theta_{\overline{a}} \theta_{\overline{b}\|\overline{a}}$

Multiplicando cada $Pr(A, B)$ com as suas respectivas variáveis indicadoras temos que:

Tabela 4

A	B	$Pr(A, B)$
a	b	$\lambda_a \lambda_b \theta_a \theta_{b\|a}$
a	$\overline{b}$	$\lambda_a \lambda_{\overline{b}} \theta_a \theta_{\overline{b}\|a}$
$\overline{a}$	b	$\lambda_{\overline{a}} \lambda_b \theta_{\overline{a}} \theta_{b\|\overline{a}}$
$\overline{a}$	$\overline{b}$	$\lambda_{\overline{a}} \lambda_{\overline{b}} \theta_{\overline{a}} \theta_{\overline{b}\|\overline{a}}$

Somando todas as probabilidades da distribuição temos:

$$f = \lambda_a \lambda_b \theta_a \theta_{b|a} + \lambda_a \lambda_{\overline{b}} \theta_a \theta_{\overline{b}|a} + \lambda_{\overline{a}} \lambda_b \theta_{\overline{a}} \theta_{b|\overline{a}} + \lambda_{\overline{a}} \lambda_{\overline{b}} \theta_{\overline{a}} \theta_{\overline{b}|\overline{a}} \tag{1}$$

Chamamos a função f de *network polynomial*, que representa a distribuição de probabilidade da rede na Figura 1. Para computar a probabilidade dada qualquer evidência e, definimos cada variável indicadora de forma consistente com a evidência. Vamos definir o que significa definir uma variável de forma consistente com uma evidência:

Definição. *Sejam um conjunto $X = \{x_1, ..., x_n, \overline{x}_1, ..., \overline{x}_n\}$ das variáveis da rede e uma evidência $e = \{e_p, ..., e_q\}$. Dizemos que X está consistente com e quando para cada e_i, se $e_i = 1$ então $x_i = 1$ e $\overline{x}_i = 0$; e se $e_i = 0$ então $x_i = 0$ e $\overline{x}_i = 1$. Para todo j tal que $1 \leq j \leq n$ e que não exista um $e_j \in e$ definiremos $x_j = \overline{x}_j = 1$. Por questões de visibilidade, definimos $(x_i, \overline{x}_i)$ ao invés de $(\lambda_i, \lambda_{\overline{i}})$, porém as notações são equivalentes.*

Portanto, se por exemplo tivermos evidência $e = \overline{a}$, teremos na *network polynomial* da Figura 1 os seguintes indicadores: $\lambda_a = 0, \lambda_{\overline{a}} = 1, \lambda_b = 1, \lambda_{\overline{b}} = 1$. Neste exemplo, o valor de f seria:

$$\begin{aligned} f(\mathbf{e} = \overline{a}) &= (0)(1)\theta_a \theta_{b|a} + (0)(1)\theta_a \theta_{\overline{b}|a} + (1)(1)\theta_{\overline{a}} \theta_{b|\overline{a}} + (1)(1)\theta_{\overline{a}} \theta_{\overline{b}|\overline{a}} \\ &= \theta_{\overline{a}} \theta_{b|\overline{a}} + \theta_{\overline{a}} \theta_{\overline{b}|\overline{a}} \\ &= Pr(\mathbf{e}) \end{aligned} \tag{2}$$

Como a *network polynomial* de uma rede é a distribuição de probabilidade da rede, então seu tamanho é exponencial. O circuito aritmético é uma representação compacta de f. Em alguns casos f não pode ser limitado, enquanto que o circuito pode.

Vamos definir formalmente uma *network polynomial*, mas antes vamos definir algumas notações. Dizemos que $\theta_{x|\mathbf{u}} \sim \mathbf{z}$ para dizer que $x\mathbf{u}$ é consistente com $\mathbf{z}, x\mathbf{u} \sim \mathbf{z}$. Portanto, $\prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}}$ denota o produto de todos os parâmetros $\theta_{x|\mathbf{u}}$ onde $x\mathbf{u}$ seja consistente com $\mathbf{z}$. A mesma notação aplica-se a $\lambda_x \sim \mathbf{z}$.

Definição. *Seja $\mathcal{N}$ uma Rede Bayesiana sob variáveis $\mathbf{Z}$. Para cada variável X com pais $\mathbf{U}$, chamamos de λ_x o indicador de x e $\theta_{x|\mathbf{u}}$ o parâmetro. A network polynomial de $\mathcal{N}$ é definida como:*

$$f \stackrel{\text{def}}{=} \sum_x \prod_{\theta_{x|\mathbf{u}} \sim \mathbf{z}} \theta_{x|\mathbf{u}} \prod_{\lambda_x \sim \mathbf{z}} \lambda_x. \tag{3}$$

O valor da network polynomial f com evidência e é o resultado da função $f(e)$ substituindo cada indicador λ_x em f com um valor consistente com e.

Teorema. *Sejam $\mathcal{N}$ uma Rede Bayesiana, Pr a distribuição de probabilidade induzida e f a polynomial. Para toda evidência e, temos que $f(e) = Pr(e)$.*

2.3 Circuit propagation

Nesta subseção definiremos o que são derivadas parciais e sua utilidade em Redes Bayesianas.

A derivada parcial de uma variável representa o quanto a mudança de uma variável impacta na saída final da rede. Portanto, seja $\lambda_{\overline{a}} = 0$ e sua derivada parcial $\partial f / \partial \lambda_{\overline{a}} = 0.4$, e dada evidência $e = a\overline{c}$ e seja o valor final de $f(e) = 0.1$, ao mudarmos o valor de $\lambda_{\overline{a}} = 1$, estaremos na realidade mudando a evidência de $a\overline{c}$ para $\overline{c}$, e portanto teremos um valor final de 0.5 ao invés de 0.1, que representa justamente $f(\overline{c})$. Como pode-se notar, ter a derivada parcial evita termos de computar toda vez a rede, já que temos as diferenças no próprio circuito.

Vamos definir uma notação de mudança de evidência. Sejam uma evidência e e X um conjunto de variáveis. Dizemos que $e - X$ é todos os elementos de e menos aqueles que são equivalentes a qualquer elemento do conjunto X. Por exemplo, se $e = ab\overline{c}$, então $e - A = b\overline{c}$ e $e - AC = b$.

Teorema. *Sejam $\mathcal{N}$ uma Rede Bayesiana, Pr a distribuição de probabilidade, f a network polynomial e $\boldsymbol{e}$ uma evidência qualquer. Para cada indicador λ_x temos que:*

$$\frac{\partial f}{\partial \lambda_x}(\boldsymbol{e}) = Pr(x, \boldsymbol{e} - X) \tag{4}$$

Além disso, para cada parâmetro $\theta_{x|\boldsymbol{u}}$ temos que:

$$\theta_{x|\boldsymbol{u}} \frac{\partial f}{\partial \theta_{x|\boldsymbol{u}}}(\boldsymbol{e}) = Pr(x, \boldsymbol{u}, \boldsymbol{e}) \tag{5}$$

Disso chega-se diretamente que:

$$\frac{\partial f}{\partial \theta_{x|\boldsymbol{u}}}(\boldsymbol{e}) = \frac{Pr(x, \boldsymbol{u}, \boldsymbol{e})}{\theta_{x|\boldsymbol{u}}}, \qquad \text{quando } \theta_{x|\boldsymbol{u}} \neq 0. \tag{6}$$

Essa derivada é mais comumente escrita como:

$$\frac{\partial f}{\partial \theta_{x|\boldsymbol{u}}}(\boldsymbol{e}) = \frac{\partial Pr(\boldsymbol{e})}{\partial \theta_{x|\boldsymbol{u}}}, \qquad \text{onde } Pr \text{ é a distribuição de } f. \tag{7}$$

2.4 Evaluation and differentiation passes

Na subseção 12.3.1 do livro vimos como implementar as passagens *top-down* e *bottom-up*. Em seguida analisamos a complexidade das duas passagens. Não vamos citar os algoritmos neste relatório pois não é de muita importância para o projeto. No entanto, é interessante citar as complexidades dos algoritmos.

No algoritmo de *bottom-up*, a complexidade em tempo foi linear ao tamanho do circuito, onde o tamanho foi definido como o número de arestas no circuito.

No algoritmo de *top-down*, a princípio o algoritmo era linear se o número de nós crianças era limitado. Porém, no segundo algoritmo garantimos que fosse linear sempre, já que se um nó era zero, não guardávamos os nós crianças dele.

3 Conclusão

Estudou-se algumas notações em Redes Bayesianas que são indispensáveis para aprender Sum-Product Networks. Além disso, o estudo de *network polynomial*, que é bastante presente em PGMs é um requerimento importante para SPNs. Também adquiriu-se uma noção básica de circuitos aritméticos e de como implementa-los. Um outro ponto importante foi diferenciação parcial de variáveis, que evita termos de passar várias vezes pelo circuito quando mudamos a evidência.

Referências

[1] Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. 1st Edition. Cambridge University Press, 2009.