Redes soma-produto e sua relação com redes bayesianas

Tiago Madeira
<madeira@ime.usp.br>

PCS5708 – Técnicas de Raciocínio Probabilístico em Inteligência Artificial

Prof. Paulo Sergio Cugnasca
Escola Politécnica – Universidade de São Paulo

Maio de 2018

Roteiro

Motivação (1/2)

Redes soma-produto (*Sum-Product Network* = *SPN*)[1] são modelos profundos tratáveis para inferência probabilística.

Resultados muito bons para diversas aplicações como modelagem de linguagem[2], classificação e reconstrução de imagens[1].

[1] POON, H. DOMINGOS, P. Sum-Product Networks: A New Deep Architecture (2011). *Uncertainty in Artificial Intelligence (UAI 2011)*

[2] CHENG, W. KOK, S. PHAM, H. CHIEU, H. MING, K. CHAI, A. Language Modeling with Sum-Product Networks (2014). Disponível em `http://spn.cs.washington.edu/papers/is14.pdf`

Motivação (2/2)

Em 2015, artigo[3] apresentou conexão teórica entre redes soma-produto e redes bayesianas (*Bayesian Network* = *BN*), junto com algoritmo polinomial para realizar conversão.

O que essa conexão pode nos dizer sobre o conhecimento probabilístico codificado nas redes soma-produto?

[3] ZHAO, H. MELIBARI, M. POUPART, P. On the Relationship between Sum-Product Networks and Bayesian Networks (2015). Disponível em `http://arxiv.org/abs/1501.01239`

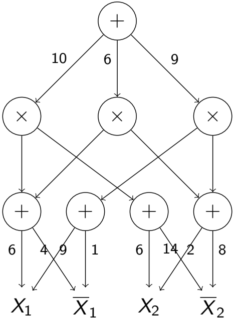
No que consiste o projeto?

- Estudo sobre redes soma-produto e sua relação com redes bayesianas
- Implementação de algoritmo para construir uma rede bayesiana a partir de uma rede soma-produto

Redes soma-produto (Poon & Domingos, 2011)

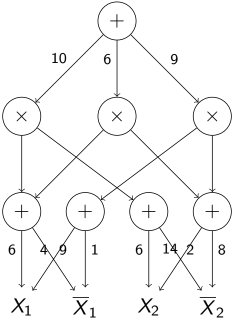
Uma rede soma-produto é um **grafo acíclico dirigido enraizado** de somas e produtos, no qual folhas são indicadores de variáveis ou distribuições univariadas.

Arestas que saem de nós do tipo soma são ponderadas com um peso não-negativo.



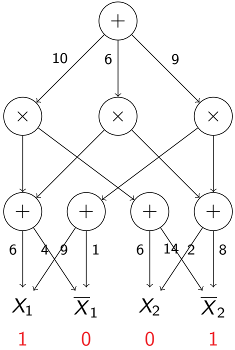
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) =$?



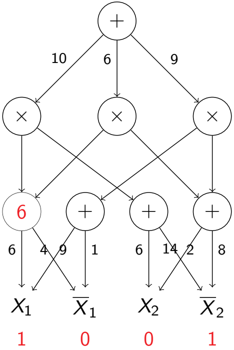
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



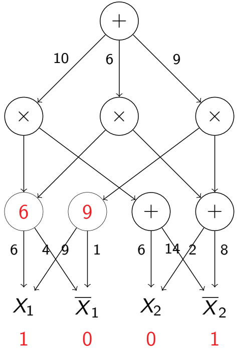
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



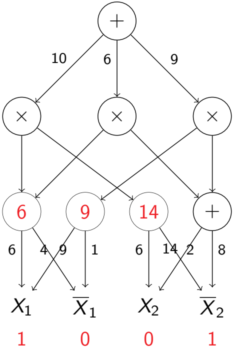
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



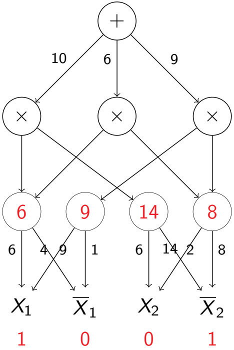
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



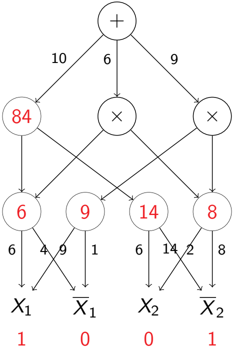
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



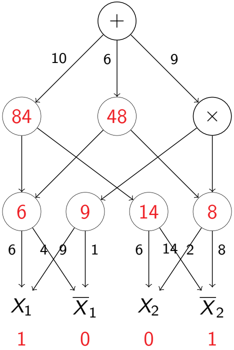
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



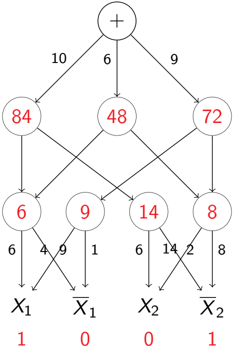
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



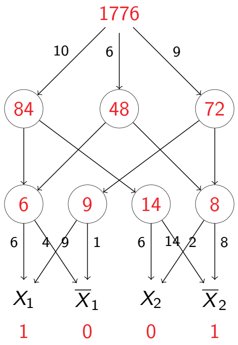
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \ ?$



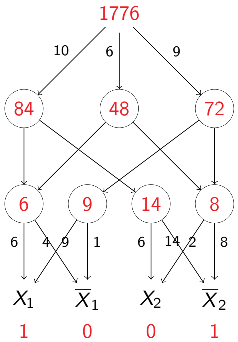
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) =$?



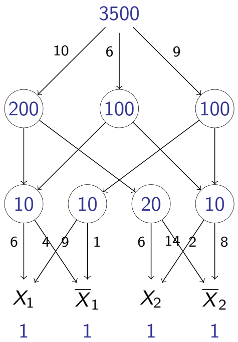
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{?}$



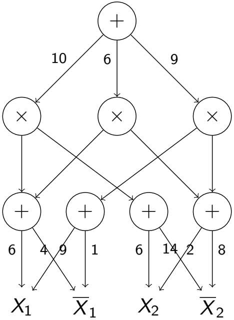
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$



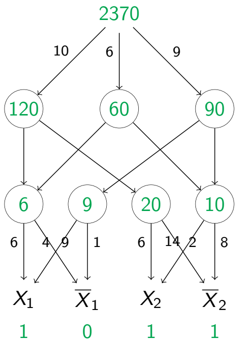
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) =$?



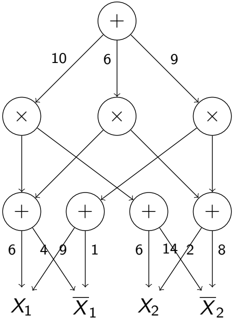
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) = \frac{2370}{3500}$



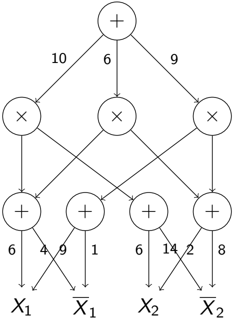
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) = \frac{2370}{3500}$
- $P(\overline{X}_2 | X_1) = \ ?$



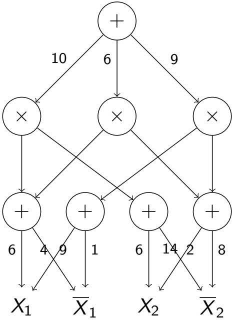
Redes soma-produto (Poon & Domingos, 2011)

- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) = \frac{2370}{3500}$
- $P(\overline{X}_2 | X_1) = \frac{P(X_1, \overline{X}_2)}{P(X_1)}$



Redes soma-produto (Poon & Domingos, 2011)

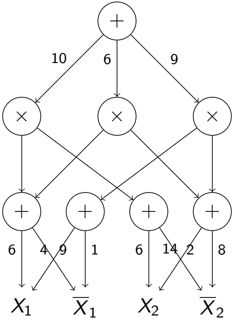
- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) = \frac{2370}{3500}$
- $P(\overline{X}_2 | X_1) = \frac{P(X_1, \overline{X}_2)}{P(X_1)} = \frac{1776}{2370}$



Redes soma-produto (Poon & Domingos, 2011)

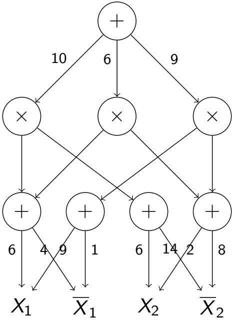
- $P(X_1, \overline{X}_2) = \frac{1776}{3500}$
- $P(X_1) = \frac{2370}{3500}$
- $P(\overline{X}_2|X_1) = \frac{P(X_1, \overline{X}_2)}{P(X_1)} = \frac{1776}{2370}$

$\Rightarrow$ inferência em tempo linear

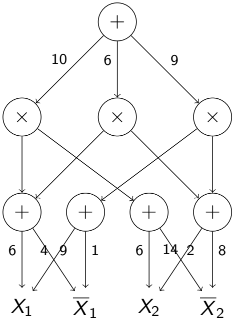


Propriedades de redes soma-produto

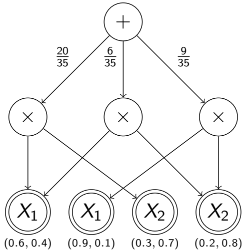
- **Escopo** de um nó: conjunto de variáveis na sub-rede enraizada em tal nó *[scope]*
- **SPN decomponível**: todo nó produto tem filhos com escopos disjuntos *[decomposable]*
- **SPN completa:** todo nó soma tem filhos com o mesmo escopo *[complete/smooth]*



Forma normal



(a) SPN completa e decomponível.

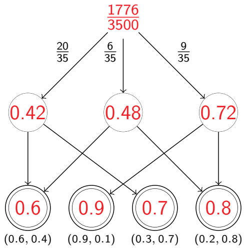


(b) Mesma SPN na forma normal.

Forma normal

Exemplo:

$$P(X_1, \overline{X}_2) = \frac{1776}{3500} \quad \checkmark$$



(b) Mesma SPN na forma normal.

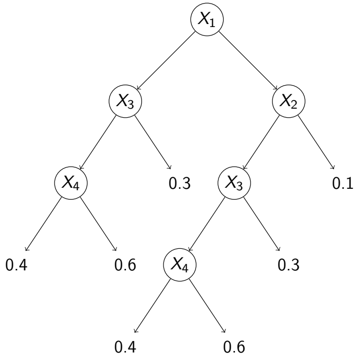
Diagramas de decisão algébrica (Bahar *et al*, 1993)

Um diagrama de decisão algébrica[4] (*Algebraic Decision Diagram* = *ADD*) é um **grafo acíclico dirigido enraizado** que representa uma função de variáveis com domínio finito.

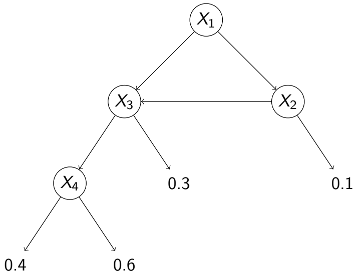
Intuitivamente é como uma árvore de decisão, mas mais compacto porque explora subgrafos isomórficos.

[4] BAHAR, R. FROHM, E. GAONA, C. HACHTEL, G. MACII, E. PARDO, A. SOMENZI, F. Algebraic decision diagrams and their applications (1993). Disponível em `https://doi.org/10.1109/ICCAD.1993.580054`

Exemplo de ADD binária e relação com árvore de decisão

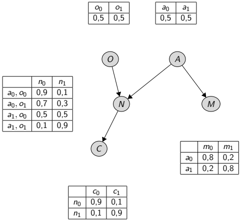


(a) Representação de uma função de variáveis binárias com árvore de decisão.

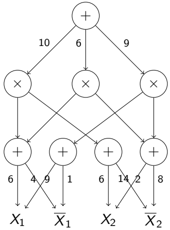


(b) Representação da mesma função com diagrama de decisão algébrica.

Representação gráfica



Redes bayesianas:
Representação gráfica de dependências diretas



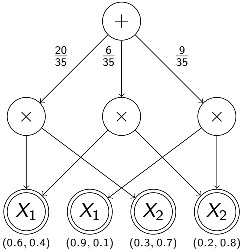
Redes soma-produto:
Representação gráfica da computação

Algoritmo de Zhao *et al.* (2015)

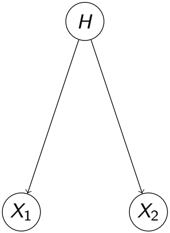
Dada uma SPN normal $\mathcal{S}$ sobre variáveis booleanas $X_{1:N}$, retorna uma rede bayesiana $\mathcal{B}$ que representa a mesma distribuição com $|\mathcal{B}| = O(N|\mathcal{S}|)$.

1. Construção da **estrutura**
2. Computação das **distribuições de probabilidade condicional** (*conditional probability distribution* = *CPD*)

Algoritmo de Zhao *et al.* (2015) - Estrutura (1/2)



(a) SPN na forma normal.



(b) Estrutura da BN correspondente.

Algoritmo de Zhao *et al.* (2015) - Estrutura (2/2)

Quando fazemos inferência, implicitamente estamos marginalizando as variáveis ocultas.

A rede bayesiana resultante tem uma **estrutura bipartida** com uma camada de variáveis ocultas apontando para uma camada de variáveis observáveis.

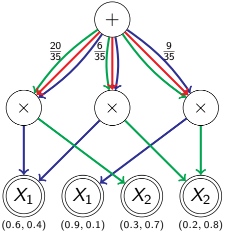
Algoritmo de Zhao *et al.* (2015) - CPDs (1/3)

Variáveis ocultas: Seja H_v a variável oculta correspondente ao nó soma v de $\mathcal{S}$. Seja l o grau de saída de v. Como a SPN é normal temos $\sum_{i=1}^{l} w_i = 1$ e $w_i \geq 0 \quad \forall i$.

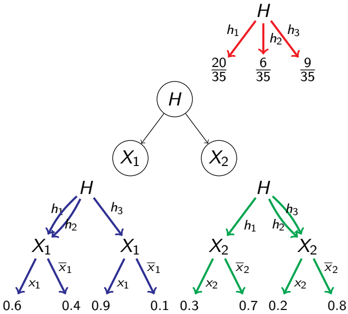
Isso sugere tomar $P(H_v = i) = w_i$.

E as variáveis observáveis?

Algoritmo de Zhao *et al.* (2015) - CPDs (2/3)

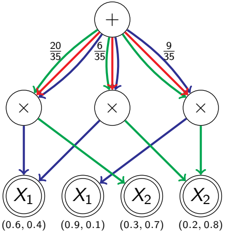


(a) Sub-SPNs usadas para calcular CPDs.



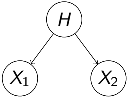
(b) BN correspondente com ADDs.

Algoritmo de Zhao *et al.* (2015) - CPDs (2/3)



(a) Sub-SPNs usadas para calcular CPDs.

h_1	h_2	h_3
$\frac{20}{35}$	$\frac{6}{35}$	$\frac{9}{35}$



	x_1	$\overline{x}_1$
h_1	0.6	0.4
h_2	0.6	0.4
h_3	0.9	0.1

	x_2	$\overline{x}_2$
h_1	0.3	0.7
h_2	0.2	0.8
h_3	0.2	0.8

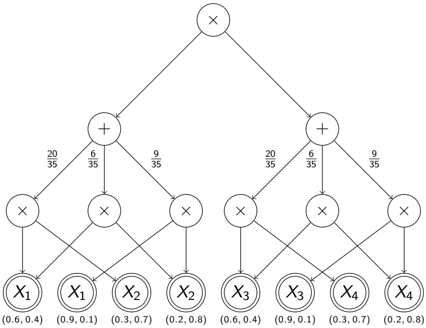
(b) BN com tabelas de CPDs.

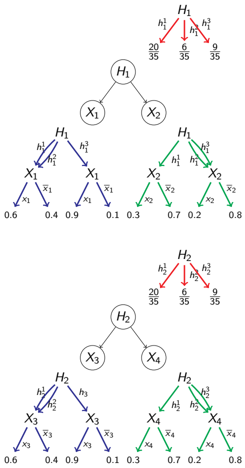
Algoritmo de Zhao *et al.* (2015) - CPDs (3/3)

Como a SPN é decomponível, cada nó produto tem filhos com escopos disjuntos. Para cada variável observável X, construímos uma ADD extraindo de $\mathcal{S}$ a sub-SPN induzida por X e contraindo todos seus nós produto.

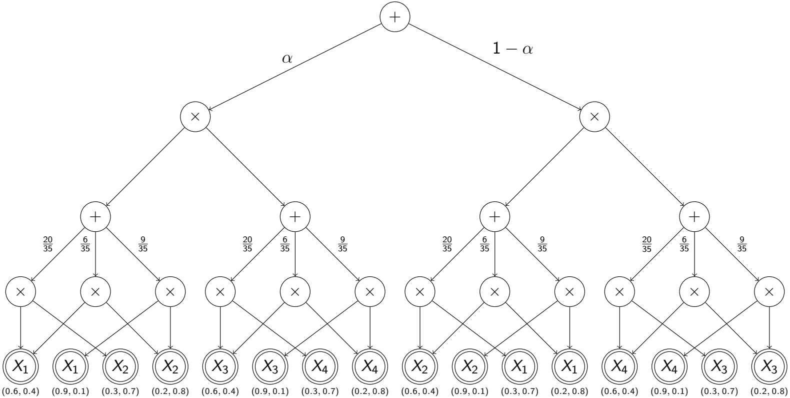
Percorremos a ADD para encontrar $P(X|H_1 = h_1^*, \cdots, H_m = h_m^*)$.

Exemplos (1/2)

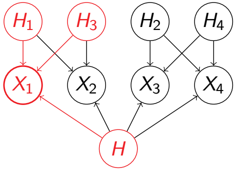

×
+
+
20/35
6/35
9/35
20/35
6/35
9/35
X_1
X_1
X_2
X_2
X_3
X_3
X_4
X_4
(0.6, 0.4)
(0.9, 0.1)
(0.3, 0.7)
(0.2, 0.8)
(0.6, 0.4)
(0.9, 0.1)
(0.3, 0.7)
(0.2, 0.8)

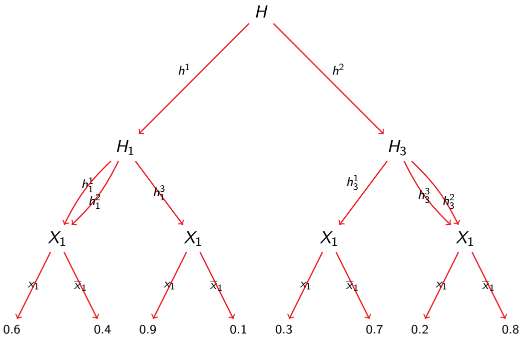

H_1
h_1^1
h_1^2
h_1^3
20/35
6/35
9/35
X_1
X_2
0.6
0.4
0.9
0.1
0.3
0.7
0.2
0.8
H_2
h_2^1
h_2^2
h_2^3
X_3
X_4

Exemplos (2/2)


α
$1-\alpha$
$\frac{20}{35}$ $\frac{6}{35}$ $\frac{9}{35}$
X_1 X_1 X_2 X_2 X_3 X_3 X_4 X_4 X_2 X_2 X_1 X_1 X_4 X_4 X_3 X_3
(0.6, 0.4) (0.9, 0.1) (0.3, 0.7) (0.2, 0.8)

Exemplos (2/2)

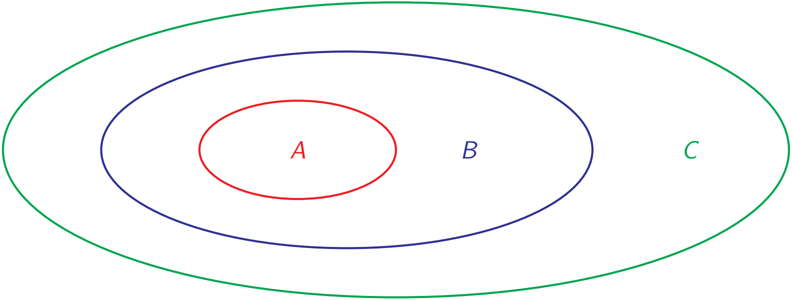

H_1
H_3
H_2
H_4
X_1
X_2
X_3
X_4
H


H
h^1
h^2
H_1
H_3
h_1^1
h_1^2
h_1^3
h_3^1
h_3^3
h_3^2
X_1
x_1
$\overline{x}_1$
0.6
0.4
0.9
0.1
0.3
0.7
0.2
0.8

Conclusões de Zhao *et al.* (2015)

1. A BN resultante da conversão tem estrutura simples (bipartida), mas é possível relacionar a profundidade de uma SPN com a *treewidth* da BN — **mais camadas** $\rightarrow$ **maior treewidth, distribuições mais complexas**.
2. Pode haver **outras técnicas** para converter uma SPN numa BN com uma representação mais compacta e um *treewidth* menor.
3. Algoritmos para aprender estrutura e parâmetros de SPNs **podem ser usados** para aprender BNs com ADDs.

Distribuições, tratabilidade e compactibilidade[5]



A: SPNs tratáveis = SPNs compactas = BNs tratáveis

B: BNs compactas

C: SPNs gerais = BNs gerais

[5] POUPART, P. Guest Lecture in STAT946: Deep Learning, University of Waterloo. Disponível em `https://www.youtube.com/watch?v=Nm0jNqOnQ2o`

Considerações finais

- Redes soma-produto são um **modelo relativamente novo e promissor**. Estudar sua teoria é interessante porque há muitas questões em aberto e alguns equívocos na literatura, como mostra Peharz (2015)[6].
- Pretende-se disponibilizar o **código implementado** (em Go) publicamente na Internet.

[6] PEHARZ, R. Foundations of Sum-Product Networks for Probabilistic Modeling (2015). PhD Thesis.

Fim

Se alguém tiver interesse na apresentação ou na monografia:
`<madeira@ime.usp.br>`

Perguntas e comentários?