Estudo sobre Sum-Product Networks e Aprendizagem Profunda

Renato Lui Geh

Instituto de Matemática e Estatística
Universidade de São Paulo

18 de maio de 2016

INSTITUTO DE MATEMATICA E ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO

Índice

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$: conjuntos

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E
$Pa(X)$	:	pais de X

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E
$Pa(X)$	:	pais de X
$Ch(X)$	:	filhos de X

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E
$Pa(X)$	:	pais de X
$Ch(X)$	:	filhos de X
$\mathcal{O}$	:	notação assintótica

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E
$Pa(X)$	:	pais de X
$Ch(X)$	:	filhos de X
$\mathcal{O}$	:	notação assintótica
Ω	:	espaço de possibilidades

Notação

$\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}$	:	conjuntos
Pr	:	distribuição ou função de probabilidade
$\mathbf{X}$	:	conjunto de variáveis aleatórias
$\mathbf{x}$	:	instanciação de $\mathbf{X}$ (i.e. $\mathbf{X} = \mathbf{x}$)
$Val(X)$	:	domínio de X
$G = (V, E)$	:	grafo com vértices V e arestas E
$Pa(X)$	:	pais de X
$Ch(X)$	:	filhos de X
$\mathcal{O}$	:	notação assintótica
Ω	:	espaço de possibilidades
$\mathcal{F}$	:	álgebra de conjuntos

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem
  - Sócrates é mortal

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem
  - Sócrates é mortal
- Lógica difusa

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem
  - Sócrates é mortal
- Lógica difusa
  - 2.0 m é 0.8 alto

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem
  - Sócrates é mortal
- Lógica difusa
  - 2.0 m é 0.8 alto
  - 2.5 m é 0.9 alto

Representação em Inteligência Artificial

- Logica proposicional e de primeira ordem
  - Todo homem é mortal
  - Sócrates é homem
  - Sócrates é mortal
- Lógica difusa
  - 2.0 m é 0.8 alto
  - 2.5 m é 0.9 alto
- Probabilidade

Probabilidade em Inteligência Artificial

- Lógica proposicional
  - Toda ave voa.

Probabilidade em Inteligência Artificial

- Lógica proposicional
  - Toda ave voa.
- Lógica difusa
  - Sabiá é mais ave?
  - Avestruz é menos ave?

Probabilidade em Inteligência Artificial

- Lógica proposicional
  - Toda ave voa.
- Lógica difusa
  - Sabiá é mais ave?
  - Avestruz é menos ave?
- Probabilidade
  - $\Pr(Voar = 1 | Ave = 1) = 0.8$
  - $\Pr(Voar = 0 | Ave = 1) = 0.2$

Complexidade de distribuições

Exemplo

Dado de 6 faces não viesado:

X se o número da face é par
Y se o número da face é múltiplo de três

x	X	Y	$\Pr(x, y)$
$\{x = 1, y = 1\}$	1	1	1/6
$\{x = 1, y = 0\}$	1	0	1/2
$\{x = 0, y = 1\}$	0	1	2/6
$\{x = 0, y = 0\}$	0	0	1/6

O número de termos desta distribuição é 2^2.

Complexidade: $\mathcal{O}((\max_i |Val(X_i)|)^n)$

Modelos Probabilísticos Baseados em Grafo (PGMs)

1. Modelos:
   - Redes Bayesianas (Bayesian Networks)
   - Redes de Markov (Markov Random Fields)
   - Grafos de potenciais (Factor Graphs)
   - Máquinas de Boltzmann (Restricted Boltzmann Machine)
   - Redes Soma-Produto (Sum-Product Networks)

Modelos Probabilísticos Baseados em Grafo (PGMs)

1. Modelos:
   - Redes Bayesianas (Bayesian Networks)
   - Redes de Markov (Markov Random Fields)
   - Grafos de potenciais (Factor Graphs)
   - Máquinas de Boltzmann (Restricted Boltzmann Machine)
   - Redes Soma-Produto (Sum-Product Networks)
2. Aplicação:
   - Reconhecimento de voz
   - Processamento de imagens
   - Multi-classificação
   - Diagnose médica
   - Predição de interação de genes e proteínas
   - Entre outros

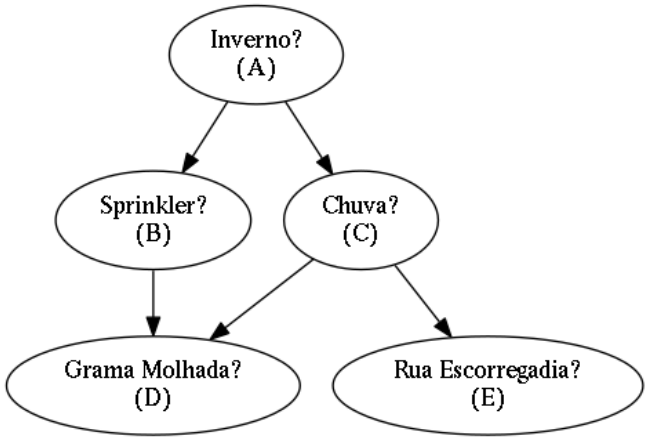
Redes Bayesianas

Definição

Uma Rede Bayesiana $\mathcal{N}$ é uma tupla $\mathcal{N} = (\Omega, \mathcal{F}, \Pr, G)$, onde Ω é o espaço de possibilidades, $\mathcal{F}$ é uma álgebra sobre Ω, Pr é uma função de probabilidade e $G = (\mathbf{X}, E)$ é um grafo onde $\mathbf{X}$ é o conjunto de variáveis de $\mathcal{N}$ e E é o conjunto de arestas. Cada variável aleatória $X_i \in \mathbf{X}$ representa uma probabilidade condicional $\Pr(X_i|Pa(X_i))$. Uma Rede Bayesiana é uma representação para a distribuição de probabilidade conjunta

$$\Pr(\mathbf{X} = \{X_1, \ldots, X_n\}) = \prod_{X \in \mathbf{X}} \Pr(X|Pa(X)). \quad (1)$$

Redes Bayesianas



A	Θ_A
true	.6
false	.4

A	B	$\Theta_{B\|A}$
true	true	.2
true	false	.8
false	true	.75
false	false	.25

A	C	$\Theta_{C\|A}$
true	true	.8
true	false	.2
false	true	.1
false	false	.9

B	C	D	$\Theta_{D\|B,C}$
true	true	true	.95
true	true	false	.05
true	false	true	.9
true	false	false	.1
false	true	true	.8
false	true	false	.2
false	false	true	0
false	false	false	1

C	E	$\Theta_{E\|C}$
true	true	.7
true	false	.3
false	true	0
false	false	1

Inferência Exata em RBs

Seja $\mathbf{Y} \subset \mathbf{X}$.

$$\Pr(\mathbf{X}) = \prod_{X \in \mathbf{X}} \Pr(X|Pa(X))$$

$$\Pr(\mathbf{Y}) = \sum_{X \in \mathbf{X} \setminus \mathbf{Y}} \Pr(X, \mathbf{Y})$$

No nosso exemplo:

$$\Pr(A, B, C, D, E) = \Pr(A) \Pr(B|A) \Pr(C|A) \Pr(D|B, C) \Pr(E|C)$$

Probabilidade condicional:

$$\Pr(\mathbf{X}|\mathbf{E}) = \frac{\Pr(\mathbf{X}, \mathbf{E})}{\Pr(\mathbf{E})}$$

Complexidade da Inferência Exata

Complexidade: $\mathcal{O}((m+n)c^{\omega+1})$

m : tamanho do maior conjunto de tabelas

n : número de variáveis

c : $\max_i |Val(X_i)|$

ω : maior escopo

Exponencial e portanto intratável.

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

$$A \vee (B \wedge (C \vee (D \wedge E \vee F)))$$

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

$$A \vee (B \wedge (C \vee (D \wedge E \vee F)))$$

Se inferência exata em RBs for subexponencial, então o problem SAT é subexponencial.

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

$$A \vee (B \wedge (C \vee (D \wedge E \vee F)))$$

Se inferência exata em RBs for subexponencial, então o problem SAT é subexponencial.

Resolvemos um problema relacionado a P vs NP! ◡̈ [KBG10]

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

$$A \vee (B \wedge (C \vee (D \wedge E \vee F)))$$

Se inferência exata em RBs for subexponencial, então o problem SAT é subexponencial.

Resolvemos um problema relacionado a P vs NP! ☺ [KBG10]

Portanto, acredita-se que não é possível. ☹

NP-completude e SAT

Inferência exata é análogo ao problema de NP-completude de SAT.

$$A \vee (B \wedge (C \vee (D \wedge E \vee F)))$$

Se inferência exata em RBs for subexponencial, então o problem SAT é subexponencial.

Resolvemos um problema relacionado a P vs NP! ⌣̈ [KBG10]

Portanto, acredita-se que não é possível. ⌢̈

Solução: inferência aproximada.

Inferência Aproximada

1. Amostragem estocástica:
   - Lógica;
   - Por importância de verossimilhança;
   - Amostragem de Gibbs;
2. Propagação de crença;
3. Algoritmo soma-produto;
4. entre outros.

Inferência aproximada $\Rightarrow$ aprendizado aproximado.

Network polynomial

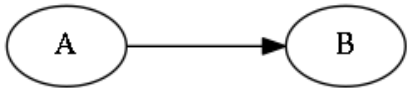
Definição

O polinômio da rede (*network polynomial*) é a função da soma de todas as instanciações da distribuição conjunta de uma Rede Bayesiana multiplicadas com as variáveis indicadoras de cada variável.

$$f(\mathbf{X}) = \sum_{\mathbf{x} \sim \mathbf{X}} \lambda_{\mathbf{x}} \theta_{\mathbf{x}|v \sim Pa(\mathbf{x})}$$

Uma variável indicadora é 1 se a variável é consistente com a instanciação e 0 caso contrário. Caso a variável não seja instanciada, a variável indicadora é 1.

Network polynomial



A	Θ_A
true	.7
false	.3

A	B	$\Theta_{B\|A}$
true	true	.2
true	false	.8
false	true	.6
false	false	.4

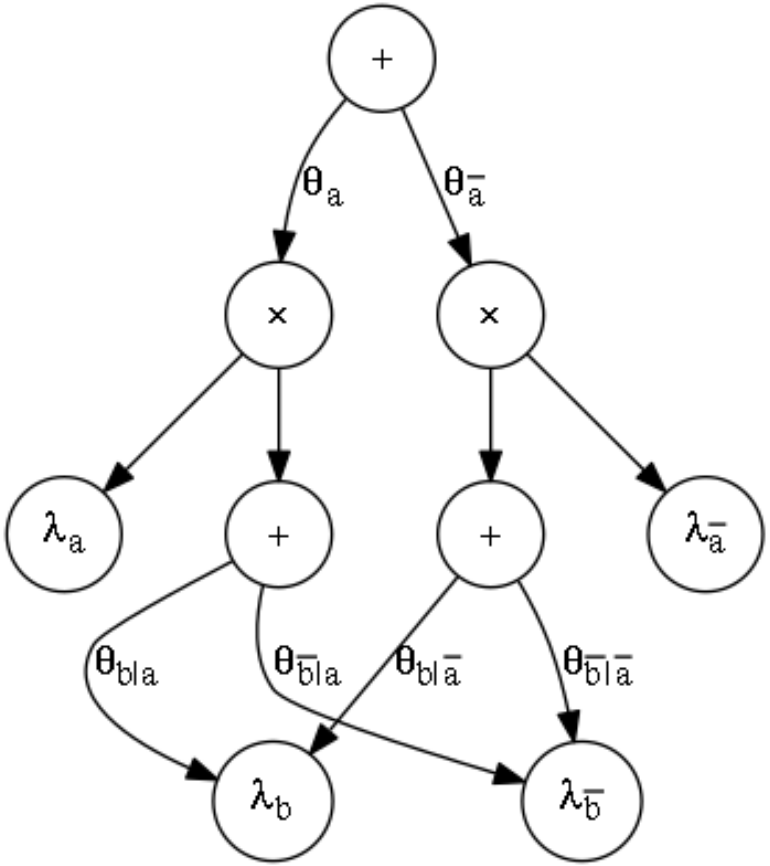
$$f(A,B) = \lambda_a\lambda_b\theta_a\theta_{b|a} + \lambda_{\bar{a}}\lambda_b\theta_{\bar{a}}\theta_{b|\bar{a}} + \lambda_a\lambda_{\bar{b}}\theta_a\theta_{\bar{b}|a} + \lambda_{\bar{a}}\lambda_{\bar{b}}\theta_{\bar{a}}\theta_{\bar{b}|\bar{a}}$$

Sum-Product Networks

Definição

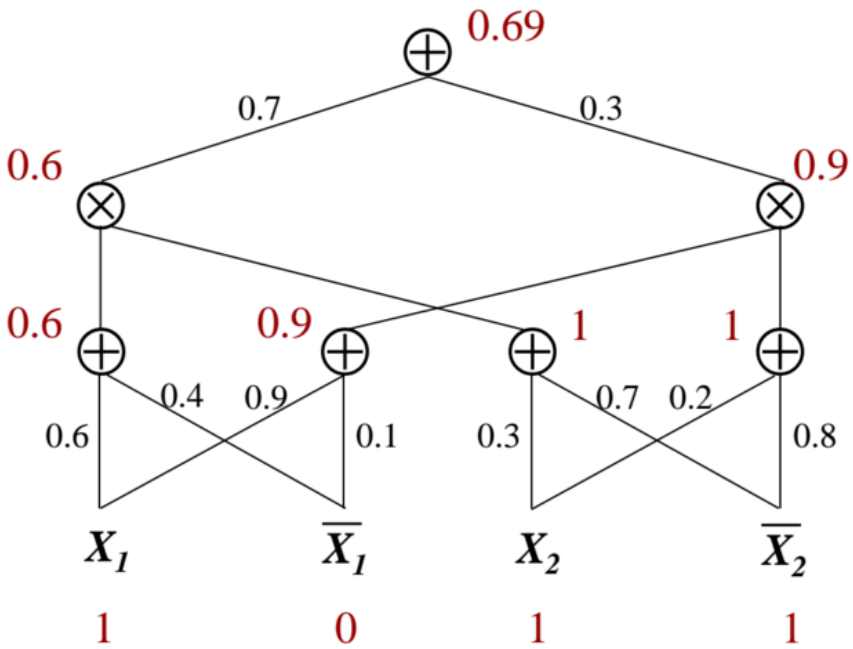
Uma SPN S é um DAG com três tipos de nós: soma, produto e indicadores. Todo nó indicador é uma folha. Todo nó soma tem pais produto, e todo nó produto tem pais soma. Toda aresta com destino a um nó soma tem uma aresta com um peso associado. O valor de um nó soma i é $\sum_{j \in Ch(i)} w_{ij} v_j$ e o valor de um nó produto i é $\prod_{j \in Ch(i)} v_j$, onde $Ch(i)$ é o conjunto de filhos de i, v_i é o valor do nó i e w_{ij} é o peso associado a aresta $i \rightarrow j$. Uma SPN representa uma função que mapeia uma distribuição de probabilidade. O valor de uma SPN é o valor do nó raíz.

Estrutura de uma SPN



$$f(A,B) = \lambda_a\lambda_b\theta_a\theta_{b|a} + \lambda_{\bar{a}}\lambda_b\theta_{\bar{a}}\theta_{b|\bar{a}} + \lambda_a\lambda_{\bar{b}}\theta_a\theta_{\bar{b}|a} + \lambda_{\bar{a}}\lambda_{\bar{b}}\theta_{\bar{a}}\theta_{\bar{b}|\bar{a}}$$

Inferência por Retropropagação



$$\lambda_{X_1} = 1, \lambda_{\overline{X}_1} = 0, \lambda_{X_2} = 1, \lambda_{\overline{X}_2} = 1$$

$$S = \Pr(X_1 = true) = f(x_1) = 0.69$$

Aprendizado

Duas classes de aprendizado:

1. Paramétrico [PD11]
   - Gradiente
   - EM (expectation-maximization)
2. Estrutural [GD13]

Semelhanças com Redes Neurais

1. Estrutura
2. Neurônios
3. Retropropagação (backpropagation)
4. RNs de Convolução
5. Arquitetura profunda
6. Representa uma função
7. Mais camadas ocultas, melhor

Relação com o Cortex

- Neurônios piramidais $\equiv$ nós soma
- Neurônios estrelados $\equiv$ nós max (produto)
- Cortex $\equiv$ SPNs com múltiplas raízes
- Raciocínio humano é mais probabilístico do que lógico

Mais informações no artigo [PD11].

Estudo planejado

1. Base teórica
   1. Teoria de Probabilidade
   2. Modelos Probabilísticos Baseados em Grafos
2. Inferência em SPNs
   1. Função de partição
   2. Marginais
   3. MAP
   4. MPE
3. Aprendizado de SPNs
   1. Paramétrico
   2. Estrutural

Referências e Bibliografia I

Gregory F. Cooper. "The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks". Em: (1988).

Adnan Darwiche. "A Differential Approach to Inference in Bayesian Networks". Em: (2003).

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

Robert Gens e Pedro Domingos. "Learning the Structure of Sum-Product Networks". Em: *International Conference on Machine Learning* 30 (2013).

Referências e Bibliografia II


J. H.P. Kwisthout, Hans L. Bodlaender e L. C. van der Gaag. "The Necessity of Bounded Treewidth for Efficient Inference in Bayesian Networks". Em: *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference in Artificial Intelligence* (2010).


Daphne Koller e Nir Friedman. *Probabilistic Graphical Models: Principals and Techniques*. The MIT Press, 2009.


Hoifung Poon e Pedro Domingos. "Sum-Product Networks: A New Deep Architecture". Em: *Uncertainty in Artificial Intelligence* 27 (2011).


Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Referências e Bibliografia III

Robert Peharz. “Foundations of Sum-Product Networks for Probabilistic Modeling”. Tese de doutorado. Graz University of Technology, 2015.