APRENDIZADO DE MÁQUINA

Prof. Dr. Anderson Ara

ara@ufpr.br

http://leg.ufpr.br/~ara/

Departamento de Estatística - UFPR

DSBD - UFPR

Redes Bayesianas

Bayesianismo e Redes Bayesianas

Redes bayesianas (redes causais, redes probabilísticas, redes de crenças, gráficos de dependência probabilística) emergiram na década de 1980.

Fornecer uma abordagem ao raciocínio probabilístico que engloba a Teoria dos Grafos, para o estabelecimento de relações entre variáveis e, também, a teoria da probabilidade para o tratamento da incerteza.

Bayesianismo e Redes Bayesianas

Bayesianismo e Redes Bayesianas

- Ponto de partida:

  "An Essay Towards Solving a Problem in Doctrine of Chance". Philosophical Transactions of the Royal Society of London, 1763"

  Artigo submetido por Richard Price e não apresenta a definição do clássico teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesianismo e Redes Bayesianas

- Definição dada por Laplace (1774)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

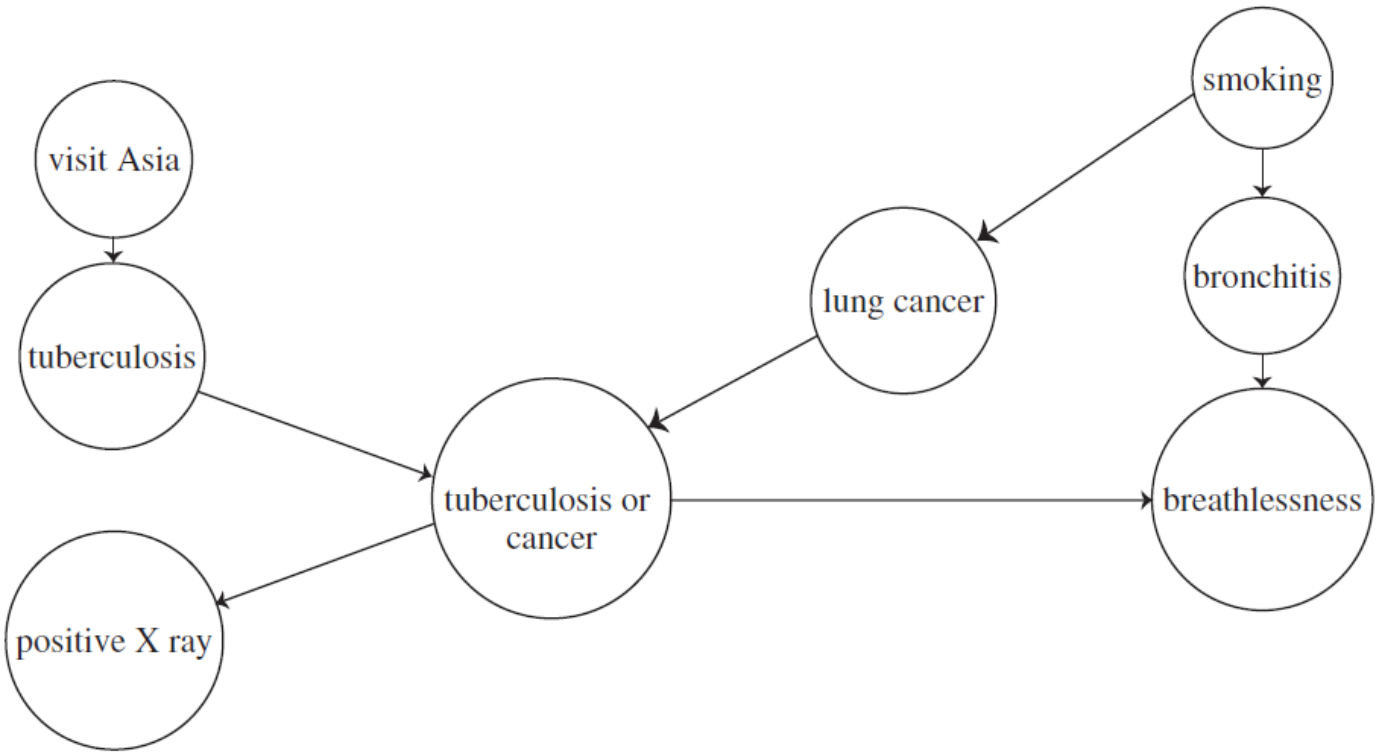
- O trabalho original de Bayes é mais focado em definições conceituais.
- Laplace foi quem formulou o teorema, tanto em sua forma discreta quanto em sua forma contínua.

Bayesianismo e Redes Bayesianas

Uma Rede Bayesiana é um Grafo Acíclico Direcionado (Directed Acyclic Graph - DAG), sendo uma representação gráfica da distribuição de probabilidade conjunta das variáveis de um problema.

Em geral, uma rede Bayesiana (discreta) consiste em uma arquitetura de rede e um conjunto de probabilidades condicionais.

Exemplo



Timo Koski and John Noble (2009). Bayesian Networks: An Introduction. page 98.

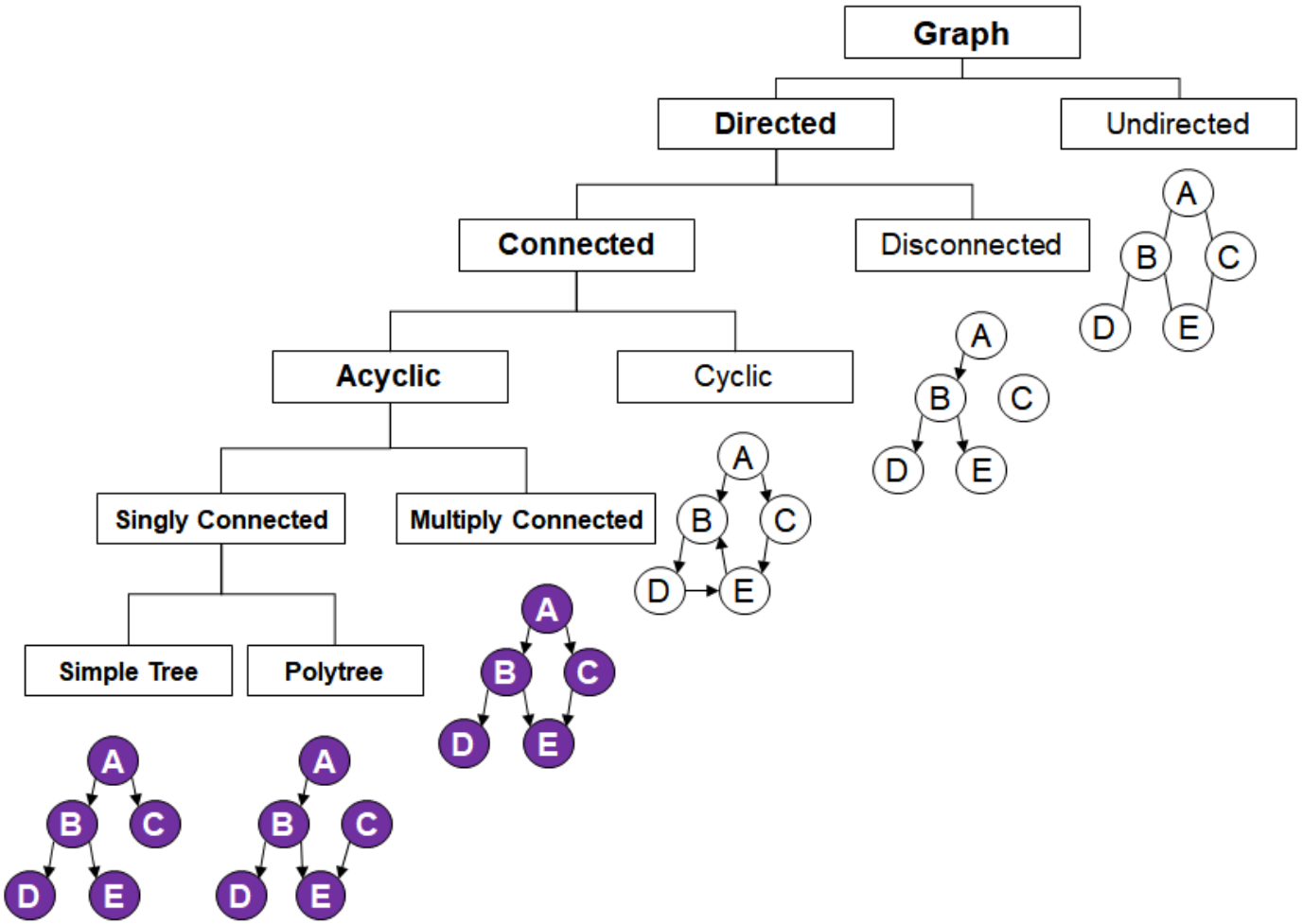
Exemplo

https://symptomate.com/

Symptomate é o verificador de sintomas mais avançado que usa inteligência artificial com redes Bayesianas para avaliar seus sintomas.

Infermedica, uma empresa polonesa com sede nos Estados Unidos.
http://www.infermedica.com

Bayesianismo e Redes Bayesianas



Bayesianismo e Redes Bayesianas

- Ferramenta Gráfica
- Integre as teorias de probabilidade e gráficos;
- Exploração de relacionamentos de (in)dependência
- Interpretação Causal
- Interpretação Visual
- Explicação e Previsão

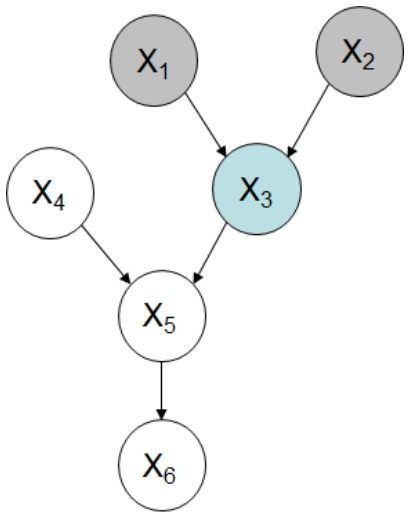
Fundamentos

Seja $\mathbb{G} = (\mathbf{V}, E)$ seja um DAG e seja $X = (X_v)$, v $\in$ V seja um conjunto de variáveis aleatórias indexadas por V. Assim, uma rede Bayesiana satisfaz a condição de Markov:

$$P(X) = \prod_{i=1}^{p} P(X_i | \mathbf{pa}(X_i))$$

Fundamentos

$$P\left(X_i|X_j, pa\left(X_i\right)\right) = P\left(X_i|pa\left(X_i\right)\right)$$



$$P\left(X_3|X_1, X_2, X_4, X_5, X_6\right) = P\left(X_3|X_1, X_2\right)$$

$$P\left(X_1, X_2, \ldots, X_p\right) = \prod_{i=1}^{p} P\left(X_i|\mathbf{pa}\left(X_i\right)\right)$$

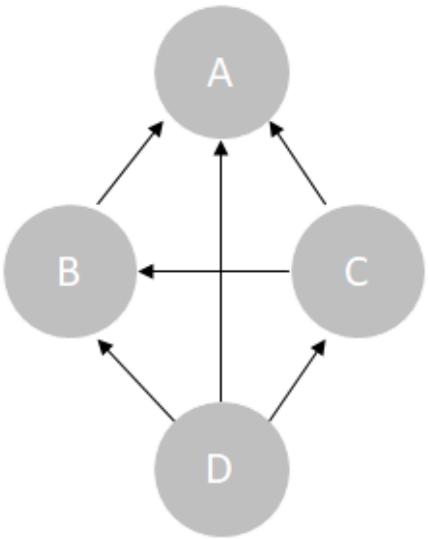
Fundamentos

- A parte qualitativa, codifica as variáveis de domínio (nós) e as influências probabilísticas (geralmente causais) entre elas (arcos)
- A parte quantitativa, codifica a distribuição de probabilidade conjunta sobre essas variáveis

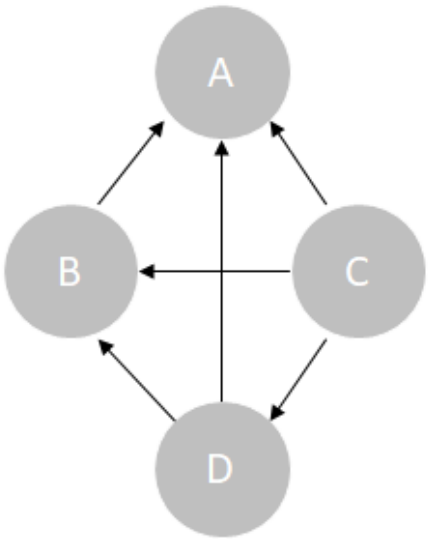
Fundamentos

A fatoração da distribuição de probabilidade conjunta fundamenta a ideia de redes bayesianas

$P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(C|D)P(D)$



$P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(D|C)P(C)$

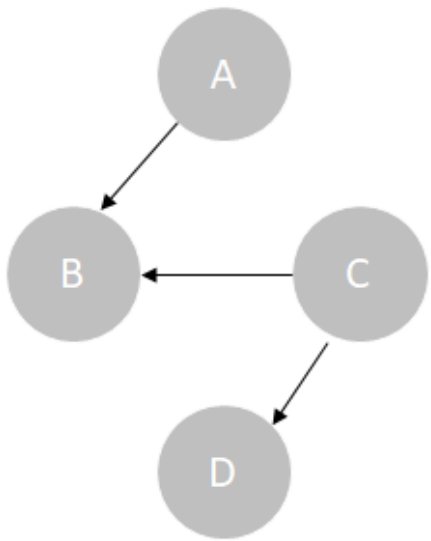


Fundamentos

A fatoração da distribuição de probabilidade conjunta fundamenta a ideia de redes bayesianas

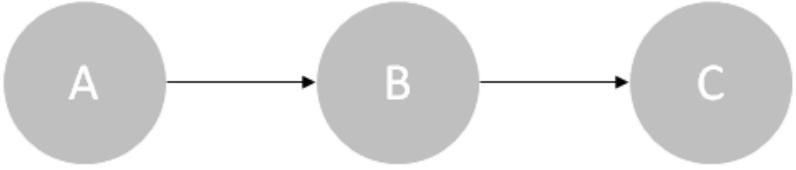
Com $B \perp D \mid C$, $D \perp A \mid C$ e $A \perp C$, temos que

$$P(A,B,C,D) = P(B \mid A,C)P(D \mid C)P(A)P(C)$$



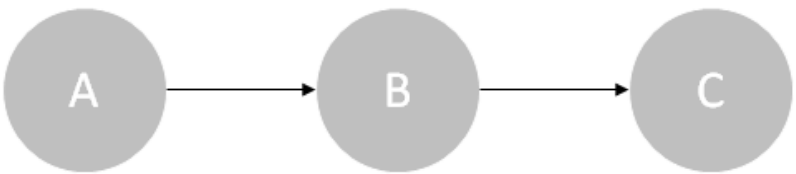
Fundamentos: Cadeia Causal

A e C são independentes dado B:



Fundamentos: Cadeia Causal

A e C são independentes dado B:



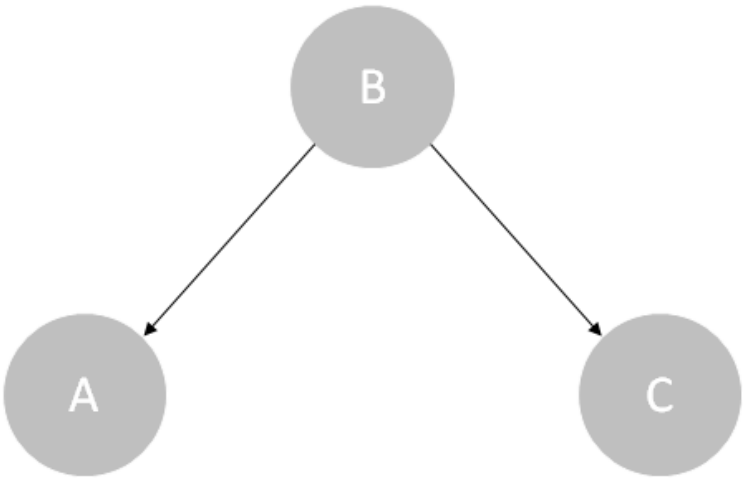
$$P(C|A,B) = \frac{P(A,B,C)}{P(A,B)}$$

$$= \frac{P(A)P(B|A)P(C|B)}{P(A)P(B|A)}$$

$$= P(C|B)$$

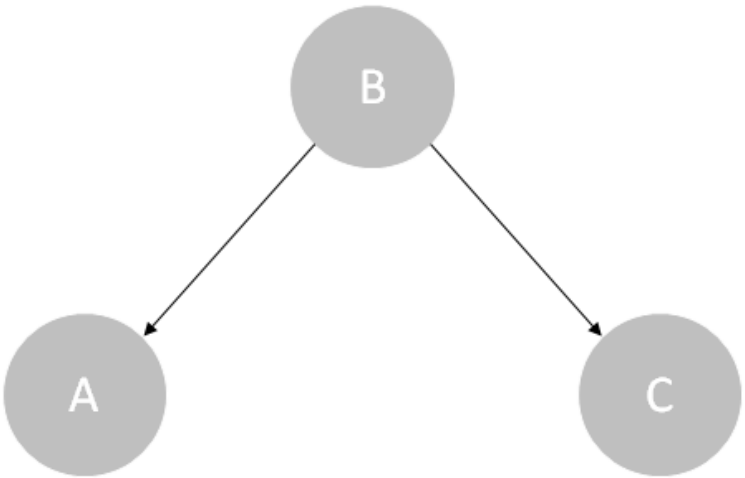
Fundamentos: Causa Comum

A e C são independentes dado B:



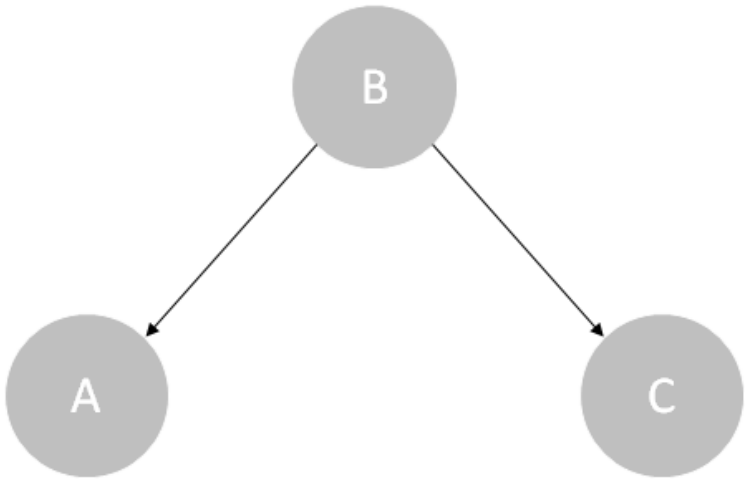
Fundamentos: Causa Comum

A e C são independentes dado B:



Fundamentos: Causa Comum

A e C são independentes dado B:



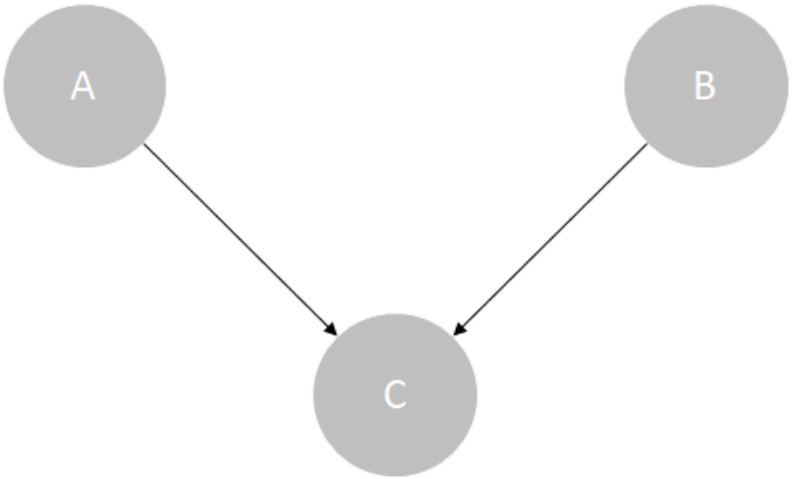
$$P(C|A,B) = \frac{P(A,B,C)}{P(A,B)}$$

$$= \frac{P(A|B)P(C|B)P(B)}{P(A|B)P(B)}$$

$$= P(C|B)$$

Fundamentos: Efeito Comum

A e B são independentes:



$$P(A,B,C) \underset{MC}{=} P(C|A,B)P(B|A)P(A).$$

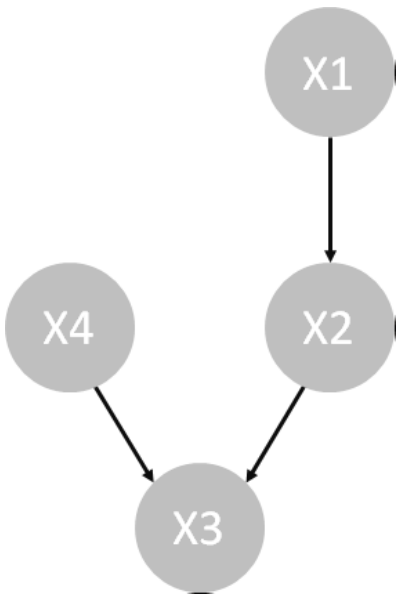
$$P(A,B,C) \underset{BN}{=} P(C|A,B)P(B)P(A)$$

$$\rightarrow P(B|A) = P(B)$$

$$\rightarrow A \perp B.$$

Fundamentos

Independencia Condicional: Cada nó é condicionalmente independente de seus não descendentes, dado seus pais imediatos.



Fundamentos

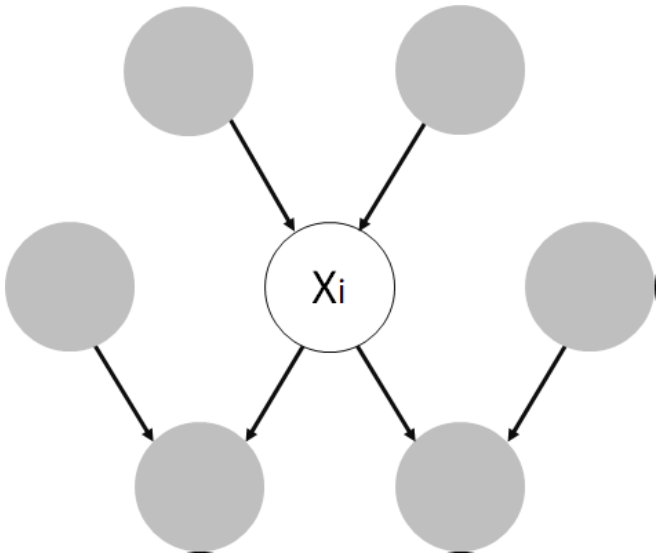
As regras anteriores fornecem toda a independência condicional relações implicadas pela rede Bayesiana?

Fundamentos

- Não!
- Por exemplo, X1 e X4 são condicionalmente indep. dado X2, X3
- Mas X1 e X4 não condicionalmente indep. dado X3
- Para isso, precisamos entender a D-separação

Fundamentos: Cobertura de Markov

Conjunto formado pelos pais, filhos e esposos de X_i.



Fundamentos

Alguns conceitos importantes:

- d - separação
- Cobertura de Markov
- Markov Equivalência

Fundamentos

Dois problemas para resolver em redes bayesianas:

- Estimação de Estrutura
- Estimação dos Parâmetros

Existem vários estudos em Redes Bayesianas Discretas, porém o caso contínuo ainda é muito incipiente e focado no pressuposto da normalidade.

Bayesianismo e Redes Bayesianas

Método geral de construção (Pearl, 1988):

- Escolha um conjunto de variáveis X_i que, em hipótese, descreve o problema;
- Escolha uma ordem para as variáveis;
- Para todas as variáveis em ordem, faça:
  - Escolha a variável X e adicione-a à rede;
  - Determina os pais da variável X com nós que já estão na rede.
  - Construa a tabela de probabilidade condicional para X.

Estimação de Estrutura

Estruturas de Estimação

Existem duas estratégias gerais de aprendizado de estruturas:

- **score based**: busca no espaço do modelo para uma pontuação ser otimizada (Métodos Hill climbing e K2);
- **constraint-based technique**: testando a independência condicional (Método PC);

Estimação de Parâmetros

Estimação de Parâmetros

- EMV (Estimador de Máxima Verossimilhança):

  Escolha de θ que maximiza a função de ligação dos dados observados no espaço paramétrico.

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

- MAP (Máximo a Posteriori)

  Escolha de *theta* que é mais provável para os dados observados poderados pela informação a priori.

$$\hat{\theta} = \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

Estimação de Parâmetros

Um dado de dois lados (moeda) é lançado n vezes, com a probabilidade de sucesso ser *theta*, sendo observados x sucessos. $X \sim Bin(n, p)$.

$$P(X|\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}$$

via EMV

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta) = \frac{x}{n}$$

Estimação de Parâmetros

Priori (Distribuição Beta):

$$P(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

com média $E(\theta) = \frac{\alpha}{\alpha+\beta}$ e moda $\frac{\alpha-1}{\alpha+\beta-2}$.

Posteriori (Distribuição Beta):

$$P(\theta|x) = \frac{1}{B(\alpha+x,\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$$

com média $E(\theta) = \frac{\alpha+x}{\alpha+\beta+n}$ e moda $\frac{\alpha+x-1}{\alpha+\beta+n-2}$.

Classificadores Bayesianos

Classificadores Bayesianos

Para tarefa de classificação, as redes Bayesianas possuem estruturas específicas e também são conhecidas como classificadores Bayesianos.

Os classificadores Bayesianos ganharam ampla aplicação devido à sua simplicidade, eficiência computacional, base teórica direta e performance de classificação competitiva.

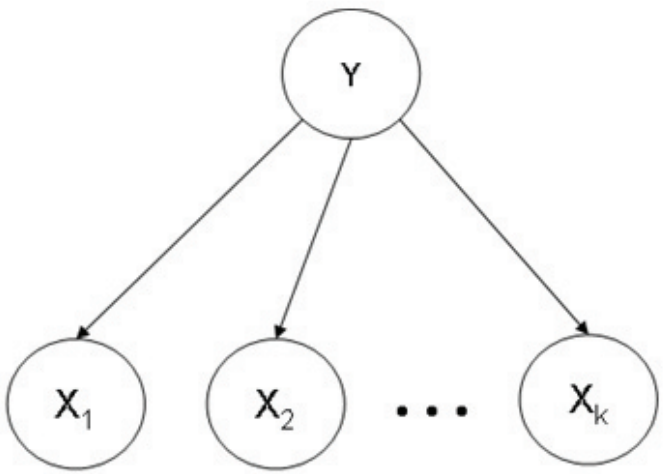
Classificadores Bayesianos

Alguns exemplos de estruturas são:

- Naïve Bayes (NB)
- Tree Augmented Network (TAN)
- k-dependence Bayesian Network (KDB)
- Averaged one dependence estimator (AODE)

Classificadores Bayesianos

O nome naïve (ingênuo, simples) deriva da premissa grosseira de que todas as variáveis explicativas são independentes, dado o Y (condicionalmente a Y).



Classificadores Bayesianos

Para o classificador Naïve Bayes em situações discretas, calculamos a probabilidade a posteriori que é dada por,

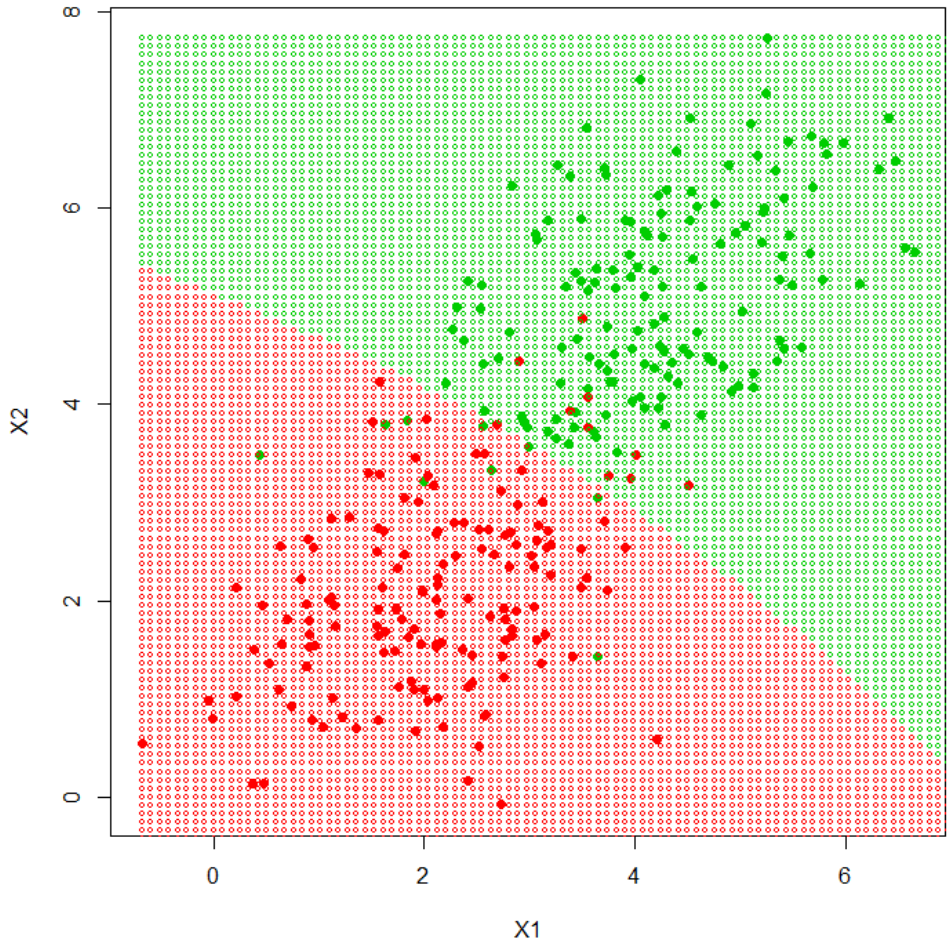
$$P(Y = y_c|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_c)\prod_{i=1}^{p} p(x_i|y_c)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} p(x_i|y_j)}$$

se X_i contínuo,

$$p(x_i|\, y_j) \sim N(\mu_{i|\, y_j}, \sigma^2_{i|\, y_j}),$$

com média $\mu_{i|y_j}$ e variância $\sigma^2_{i|y_j}$ de X_i condicional a categoria yc. O caso contínuo é conhecido como Gaussian Naïve Bayes.

Classificadores Bayesianos



- Para o classificador naïve Bayes as probabilidades são calculadas com base na frequência das observações do conjunto de treinamento.
- Utilização do m-estimador (conjugada Dirichlet-multinomial via média a posteriori de uma priori uniforme)

$$\frac{n_c + mp}{n + m}$$

- Quando a suposição de independência condicional de todos os atributos dada a classe é satisfeita, o classificador naïve Bayes é um classificador Bayesiano ótimo.
- Apresenta bons resultados mesmo que a suposição de independência condicional não seja satisfeita

Redes Bayesianas

Para redes Bayesianas de k dependência (KDB), calculamos as probabilidades a posteriori dada por,

$$P(Y = y_c|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_c)\prod_{i=1}^{p} f(x_i|pais(X_i), y_c)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} f(x_i|pais(X_i), y_j)}$$

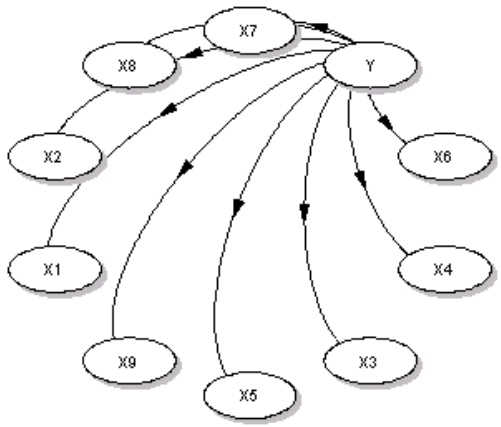
onde, se X_i contínuo,

$$p(x_i|\, pais_i, y_j) \sim N(\mu_{i|\, pais_i, y_j}, \sigma^2_{i|\, pais_i, y_j}),$$

sendo $\mu_{i|pais_i\, y_j}$ e $\sigma^2_{i|pais_i\, y_j}$ a média e a variância da variável x_i condicionada aos pais de X_i e a categoria y_j.

CLASSIFICAÇÃO: Redes Bayesianas

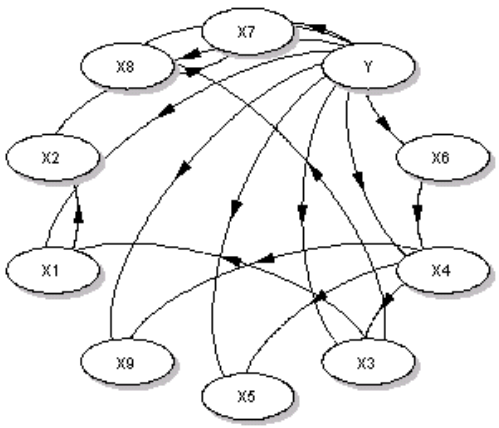
A rede com 1-dependência (KDB1) possui a mesma estrutura que uma rede probabilística para classificação e bastante difundida na literatura, conhecida como Tree Augmented Network (TAN).



KDB0

CLASSIFICAÇÃO: Redes Bayesianas

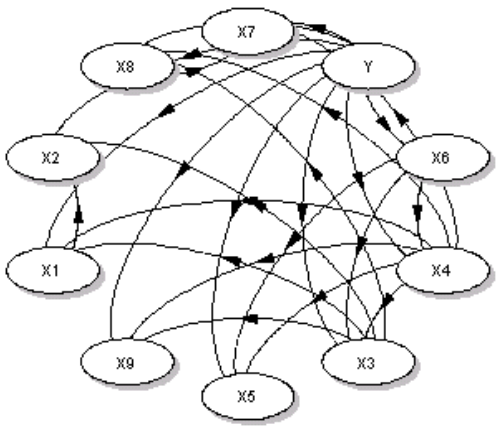
A rede com 1-dependência (KDB1) possui a mesma estrutura que uma rede probabilística para classificação e bastante difundida na literatura, conhecida como Tree Augmented Network (TAN).



KDB1

CLASSIFICAÇÃO: Redes Bayesianas

A rede com 1-dependência (KDB1) possui a mesma estrutura que uma rede probabilística para classificação e bastante difundida na literatura, conhecida como Tree Augmented Network (TAN).



KDB2

Redes Bayesianas

Algoritmo KDB (Sahami, 1996)

1. Para cada variável X_i, calcule a medida de informação mútua, denotada por $\hat{I}(X_i, Y)$;
2. Para cada par de variáveis explicativas, calcule a medida de informação mútua condicional, denotada por $\hat{I}(X_i, X_j | Y)$;
3. Defina S como a lista de variáveis explicativas utilizadas, inicialmente considere S como vazio;
4. Inicie a rede com a variável de classificação Y;
5. Repita até a lista S conter todas as variáveis explicativas:
    1. Selecione a variável explicativa X_{max} que ainda não está contida em S e que possua a maior medida $\hat{I}(X_{max}, Y)$;
    2. Adicione à rede a variável X_{max};
    3. Adicione um arco de Y para X_{max};
    4. Adicione $m = min(|S|, K)$ arcos partindo das m variáveis explicativas X_j com o maior valor $\hat{I}(X_{max}, X_j | Y)$;
    5. Adicione X_{max} à lista S.

CLASSIFICAÇÃO: Redes Bayesianas

Informação Mútua: Desenvolvida em um ramo da teoria da probabilidade e da matemática estatística que lida com problemas relacionados a comunicação denominada Teoria da Informação e introduzida por Shannon (1948):

$$I(X_i, X_j) = \sum_x \sum_y p(x_i, x_j) \log \frac{p\left(x_i, x_j\right)}{p\left(x_i\right), p\left(x_j\right)} = H(X_i, X_j) - H(X_i|X_j) - H(X_j|X_i)$$

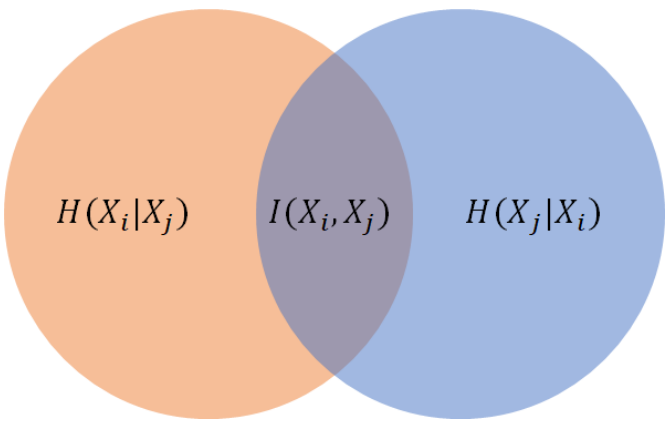
$$H(X_i|X_j) = -\sum_x p(x_i|x_j) \log p\left(x_i|x_j\right)$$

$$H(X_i, X_j) = -\sum_x p(x_i, x_j) \log p\left(x_i, x_j\right)$$

Redes Bayesianas

Informação Mútua:

$I(X_i, X_j)$ expressa a quantidade de informação que X_i compartilhada com X_j, $H(X_i|X_j)$ a entropia condicional de X_i dado X_j, valor médio do conteúdo da informação em X_i condicional a X_j.



$H(X_i, X_j)$

Redes Bayesianas

Informação Mútua Condicional: Expressa a informação mútua de duas variáveis aleatórias condicionadas a um terceiro vetor aleatório.

$$I(X_i, X_j|Z) = E_z\left(I(X_i, X_j|Z)\right)$$

$$= \sum_z \sum_x \sum_y p(z)p(x_i, x_j|z) \log \frac{p\left(x_i, x_j|z\right)}{p\left(x_i|z\right), p\left(x_j|z\right)}$$

Comentários

- Redes Bayesianas são uma abordagem gráfica e probabilística baseada em grafos acíclicos dirigidos utilizados para modelagem de dados;
- Elas consideram propriedades de (in)dependência condicional;
- Elas são flexíveis e podem suportar desde modelos simples até modelos mais complexos para objetivos diferentes;
- Há também a abordagem de redes Bayesianas dinâmicas e outras generalizações.

Comentários

- Classificadores bayesianos são um caso particular de redes bayesianas;
- Classificadores bayesianos tem principal foco em predição (machine learning);
- Geralmente os classificadores bayesianos possuem baixa complexidade computacional.
- Existem diversos métodos para a construção de classificadores bayesianos, sendo o mais popular o método de naïve Bayes.