Reposição da aula de 08/11

- Aula de 08/11 cancelada
- Reposição:

Tema 09

Redes Bayesianas

Professora:
Ariane Machado Lima

Teorema de Bayes

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

Teorema de Bayes

$$P(C|\mathbf{X}) = P(\mathbf{X}|C)\, P(C)/P(\mathbf{X})$$

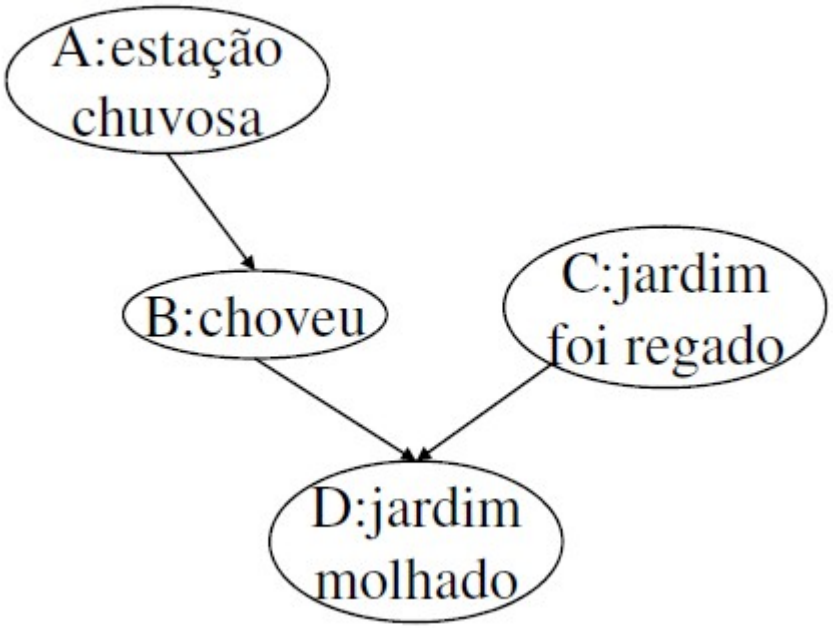
posteriori = verossimilhança * priori / evidência

Teorema de Bayes

- Probabilidades para descrever crenças ou incertezas
- Forma de atualizar o conhecimento, com base em conhecimento prévio e dados disponíveis
- É encadeável, permitindo combinar diferentes eventos (desde que se possa atribuir probabilidades a eles)
- Alguns problemas podem envolver várias variáveis, dependentes entre si
- Pode-se descrever uma REDE dessas variáveis e suas relações de dependências

Redes Bayesianas

- Modelo gráfico para representação dessa rede
- Grafo dirigido e acíclico:
  - Vértices: variáveis aleatórias
  - Arestas: relações de dependência
  - Ausência de aresta: independência condicional

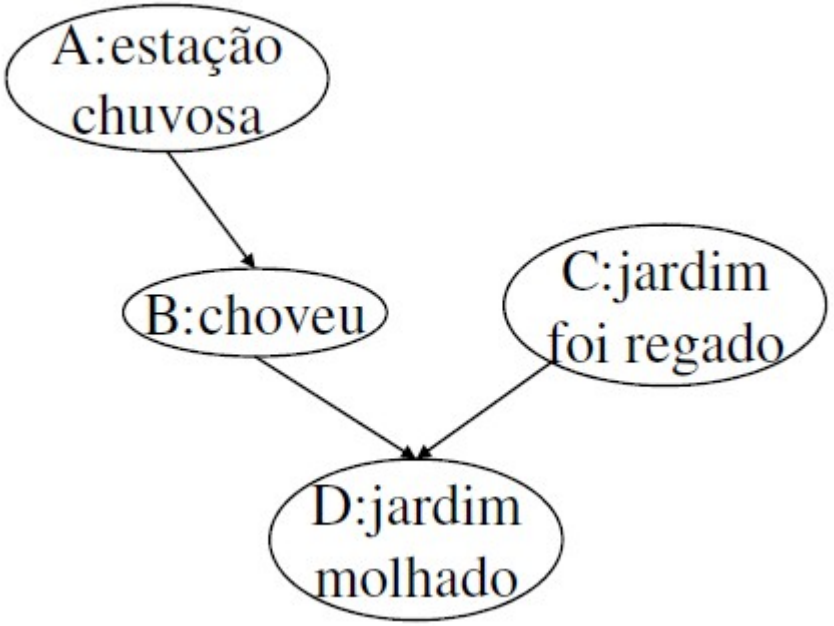


Rede Bayesiana - Definição

- Par (S, P) onde:
  - S é a estrutura da rede (nós $x = \{x_1, \ldots, x_n\}$ e arestas)
  - P é um conjunto de distribuições de probabilidades $p(x_i \mid pa(x_i))$, no qual $pa(x_i)$ são os nós pais de x_i
- Probabilidade conjunta da rede:

$P(S)$ ou $P(x) = \prod_i p(x_i \mid pa(x_i))$

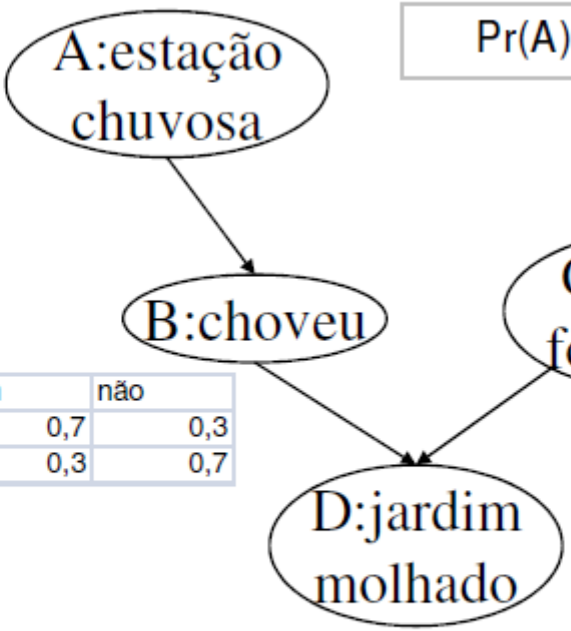
Probabilidade conjunta da rede



$$P(S) \text{ ou } P(x) = \prod_i p(x_i \mid pa(x_i))$$

$$Pr(A,B,C,D)=Pr(D|B,C)*Pr(B|A)*Pr(C)*Pr(A)$$

Exemplo



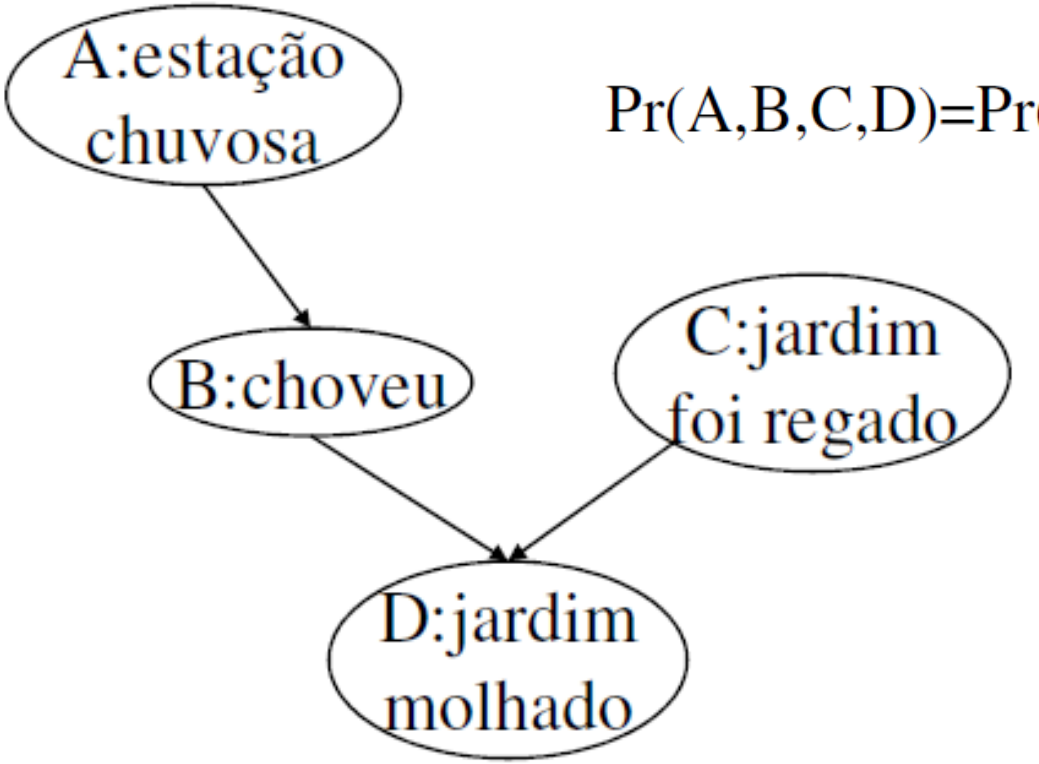
Pr(A)	sim	0,5
	não	0,5

Pr(C)	sim	0,2
	não	0,8

	A	sim	não
Pr(B|A)	sim	0,7	0,3
	não	0,3	0,7

	B	sim		não	
	C	sim	não	sim	não
Pr(D|B,C)	sim	0,9	0,9	0,9	0
	não	0,1	0,1	0,1	1

Probabilidade conjunta da rede Exemplo



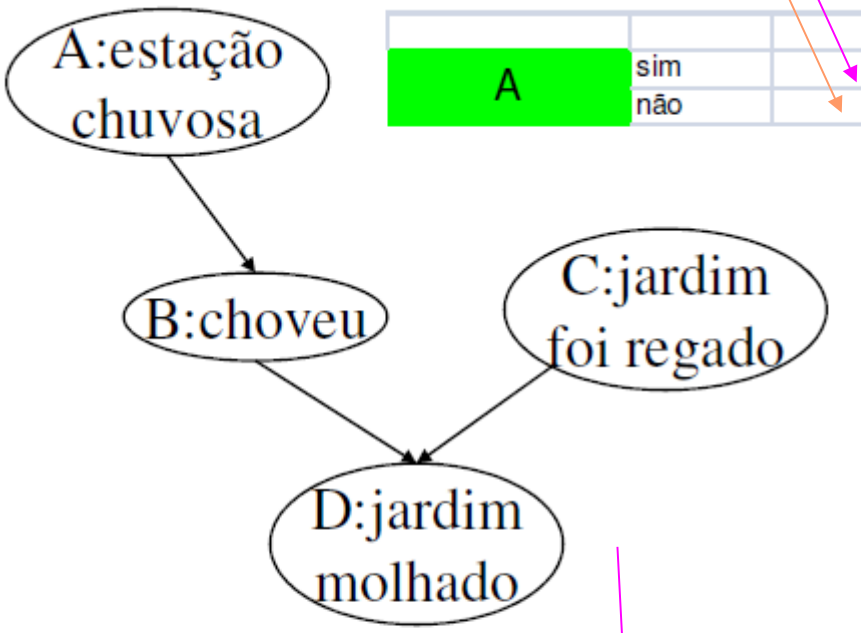
$$Pr(A,B,C,D)=Pr(D|B,C)*Pr(B|A)*Pr(C)*Pr(A)$$

	A	sim				não			
	B	sim		não		sim		não	
	C	sim	não	sim	não	sim	não	sim	não
Pr(A,B,C,D)	sim	0,063	0,027	0,252	0	0,027	0,063	0,108	0
	não	0,007	0,003	0,028	0,12	0,003	0,007	0,012	0,28

Inferência em Redes Bayesianas

- Usadas para fazer inferência
- Ex: Qual é a venda esperada para o Redoxon dadas as condições climáticas e intensidade do surto de gripe?

Inferência:
Qual a probabilidade de estarmos na estação chuvosa? (e nada mais é sabido)



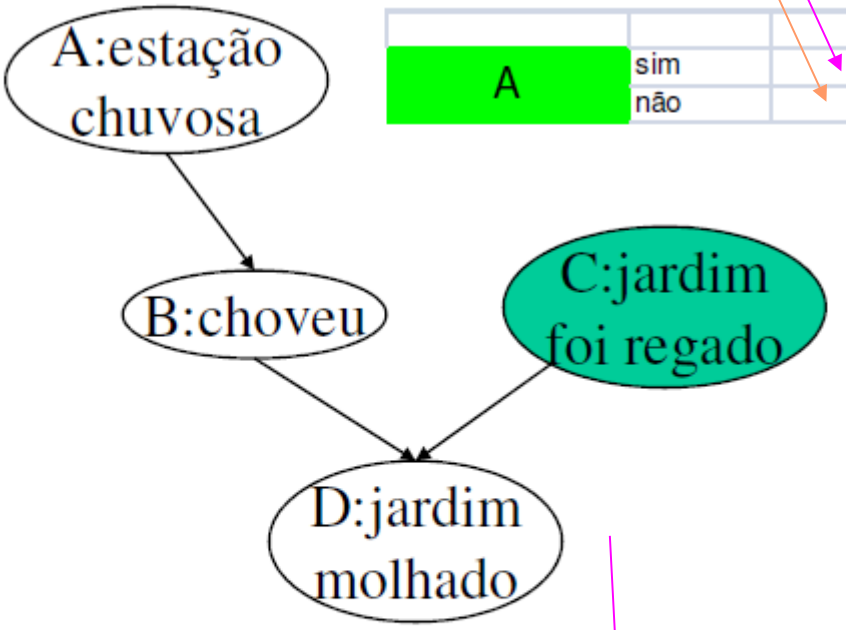
		Pr(A)
A	sim	0,5
	não	0,5

0,5

conjuntas

	A	sim				não			
	B	sim		não		sim		não	
	C	sim	não	sim	não	sim	não	sim	não
Pr(A,B,C,D)	sim	0,063	0,027	0,252	0	0,027	0,063	0,108	0
	não	0,007	0,003	0,028	0,12	0,003	0,007	0,012	0,28

Inferência:
Qual a probabilidade de estarmos na estação chuvosa? (sabendo que o jardim foi regado)



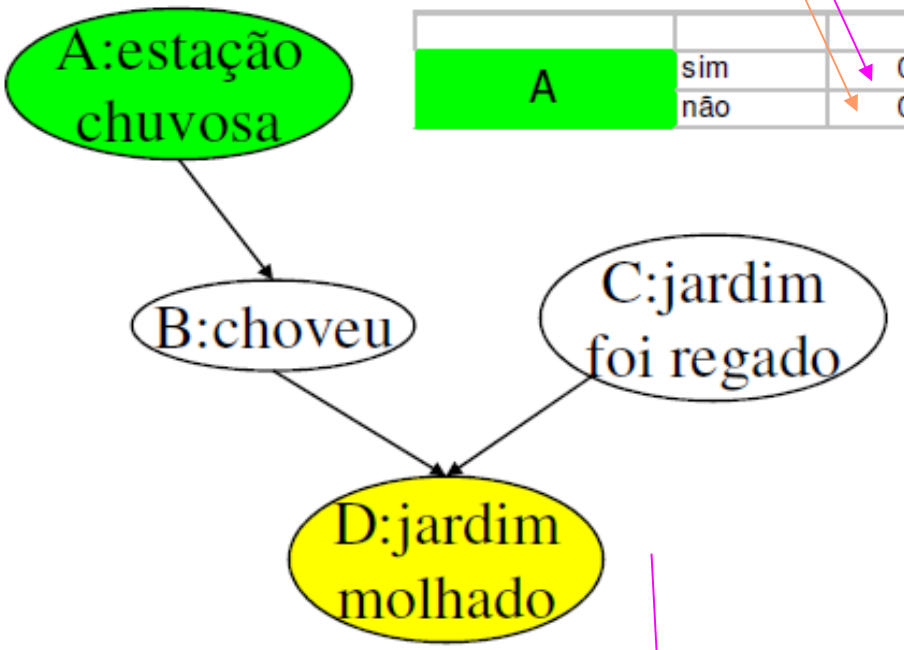
		Pr(A)
A	sim	0,7
	não	0,3

P(A|C=sim)

conjuntas Valores normalizados

	A	sim				não			
	B	sim		não		sim		não	
	C	sim	não	sim	não	sim	não	sim	não
Pr(A,B,C,D)	sim	0,063	0,027	0,252	0	0,027	0,063	0,108	0
	não	0,007	0,003	0,028	0,12	0,003	0,007	0,012	0,28

Inferência:
Qual a probabilidade de estarmos na estação chuvosa? (sabendo que o jardim está molhado)



			Pr(A)
A	sim	0,342	0,633333
	não	0,198	0,366667

P(A|D=sim)

conjuntas Valores normalizados

	A	sim				não			
	B	sim		não		sim		não	
	C	sim	não	sim	não	sim	não	sim	não
Pr(A,B,C,D)	sim	0,063	0,027	0,252	0	0,027	0,063	0,108	0
	não	0,007	0,003	0,028	0,12	0,003	0,007	0,012	0,28

Aprendizado de redes bayesianas

- Estrutura:
  - Variáveis (nós): número e quais
  - Relações de dependência (arestas)
  - Direção das arestas
- Probabilidades

Aprendizado de redes bayesianas

- Métodos automáticos e semi-automáticos para aprendizado da estrutura e das probabilidades condicionais
- Pode não se saber as relações de dependência, mas apenas ter um conjunto de dados

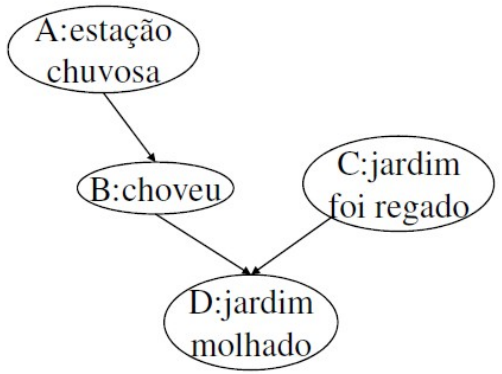
Aprendizado da estrutura

- Testes estatísticos de independência para identificar dependências e independências entre as variáveis
- Tem que saber as variáveis
- Adequar os parâmetros do teste para impedir que, dependendo do tamanho da amostra, o teste diga que as variáveis são dependentes quando na verdade não são.

Aprendizado das probabilidades

Rede Bayesiana M = (S, **θ**), sendo S a estrutura da rede e θ o vetor de todos os $\boldsymbol{\theta}_{i,}$ sendo $\boldsymbol{\theta}_{i}$:

- o conjunto de todos os parâmetros que definem as probabilidades condicionais de cada variável x_i ($P(x_i \mid pa(x_i))$), ou seja,



Aprendizado das probabilidades

Rede Bayesiana M = (S, **θ**), sendo S a estrutura da rede e θ o vetor de todos os $\boldsymbol{\theta}_i$, sendo $\boldsymbol{\theta}_i$:

- o conjunto de todos os parâmetros que definem as probabilidades condicionais de cada variável x_i ($P(x_i \mid pa(x_i))$), ou seja,

- $\boldsymbol{\theta}_i$ é o vetor (matriz) especificando, em cada posição i,j,k, $\theta_{ijk} = P(x_i^k \mid pa(x_i)^j)$, sendo x_i^k o k-ésimo valor que x_i pode assumir e $pa(x_i)^j$ a j-ésima configuração que os pais de x_i podem assumir

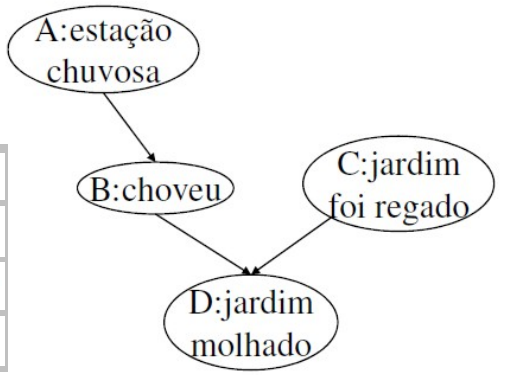
- Ex: para x_i = D, k = 1 (x_i^k = sim) ou k = 2 (x_i^k = não), $pa(x_i)$ = (B, C), $pa(x_i)^1$ = (B=sim, C=sim), $pa(x_i)^2$ = (B=sim, C=não), $pa(x_i)^3$ = (B=não, C=sim), $pa(x_i)^4$ = (B=não, C=não), $\boldsymbol{\theta}_i$ = matriz abaixo

x_i $pa(x_i)^1$ $pa(x_i)^2$ $pa(x_i)^3$ $pa(x_i)^4$

	B	sim		não	
	C	sim	não	sim	não
Pr(D|B,C)	sim	0,9	0,9	0,9	0
	não	0,1	0,1	0,1	1

k=1

k=2



Aprendizado das probabilidades

- Duas coisas são assumidas (a fim de que os parâmetros possam ser aprendidos independentemente):

  - Independência global: os parâmetros das várias variáveis (θ_i) são independentes

    - Isso significa que podemos modificar as tabelas de cada variável independentemente

  - Independência local: as incertezas dos parâmetros para as diferentes configurações de pais (θ_{ijk} para cada i) são independentes (isto é, a incerteza em P(A|b,c) é independente da incerteza em P(A|b´,c´)

    - Isso significa que os parâmetros para as duas distribuições podem ser modificados independentemente

Aprendizado das probabilidades

- Dados completos
  - Estimação por máxima verossimilhança (ML – *Maximum likelihood*)
  - Estimação bayesiana (MAP – *Maximum a posteriori*)
- Dados incompletos
  - Estimação aproximada (algoritmo EM)

Aprendizado das probabilidades - Dados completos

Seja D um dataset de casos:

- - ele é completo se cada caso é uma configuração sobre TODAS as variáveis de M

Estimação por Máxima Verossimilhança

- Para cada caso d ϵ D, P(d | M) é a verossimilhança de M dado d, ou L(M | d)
- Assumindo que os casos são independentes (dado M):

$$L(M \mid D) = \Pi_{d \in D} P(d \mid M), \text{ ou}$$

$$LL(M \mid D) = \Sigma_{d \in D} \log P(d \mid M)$$

LL : log-likelihood

- Estimação de θ por máxima verossimilhança: cálculo dos valores que maximizam LL(M | D)

Estimação por Máxima Verossimilhança

- Uso das frequências como estimativas
- Exemplo:

$$P(A=a \mid B=b, C=c) = \#(A=a, B=b, C=c) / \#(B=b, C=c)$$

Estimação por Máxima Verossimilhança

- Uso das frequências como estimativas
- Exemplo:

  P(A=a | B=b, C=c) = # (A=a, B=b, C=c) / # (B=b, C=c)

- Possível problema:

Estimação por Máxima Verossimilhança

- Uso das frequências como estimativas
- Exemplo:

  $P(A=a \mid B=b, C=c) = \#(A=a, B=b, C=c) / \#(B=b, C=c)$

- Possível problema: e se algumas contagens for igual a zero?

Estimação por Máxima Verossimilhança

- Uso das frequências como estimativas
- Exemplo:

  P(A=a | B=b, C=c) = # (A=a, B=b, C=c) / # (B=b, C=c)

- Possível problema: e se algumas contagens for igual a zero?

  - a probabilidade naquele ponto será igual a zero. Boa estimativa?

Estimação por Máxima Verossimilhança

- Uso das frequências como estimativas
- Exemplo:

  P(A=a | B=b, C=c) = # (A=a, B=b, C=c) / # (B=b, C=c)

- Possível problema: e se algumas contagens for igual a zero?

  - a probabilidade naquele ponto será igual a zero. Boa estimativa?

  - provavelmente não, principalmente se a amostra for pequena

Estimação Bayesiana (Máxima a posteriori - MAP)

- Pseudocontadores (Dirichlet)
  - Inicialização dos contadores com algum valor diferente de zero
  - Valor inicial pode ser igual para todos ou não
  - Distribuição dos valores iniciais: distribuição de Dirichlet

Aprendizado das probabilidades - Dados incompletos

- Dados incompletos
  - *Missing values*: acidentais (ex: sensor), intencionais, etc
  - Não observáveis (variáveis latentes ou escondidas)
- Opções:

Aprendizado das probabilidades - Dados incompletos

- Dados incompletos
  - *Missing values*: acidentais (ex: sensor), intencionais, etc
  - Não observáveis (variáveis latentes ou escondidas)
- Opções:
  - Jogar fora os casos com *missing values*. Problemas:
    -
    -
  - Tentar trabalhar com eles

Aprendizado das probabilidades - Dados incompletos

- Dados incompletos
  - *Missing values*: acidentais (ex: sensor), intencionais, etc
  - Não observáveis (variáveis latentes ou escondidas)
- Opções:
  - Jogar fora os casos com *missing values*. Problemas:
    - Amostra final pode ficar pequena
    - Amostra final pode ficar enviesada
  - Tentar trabalhar com eles (estimação aproximada: algoritmo EM)

Algoritmo EM - *Expectation-Maximization*

- Várias rodadas de dois passos: esperança e maximização:
  - Esperança: “completa-se” o dataset utilizando o θ' atual para calcular a esperança (média) para os *missing values*
  - Maximização: usa-se o dataset “completado” para reestimar um novo valor de θ' por máxima verossimilhança
- Esses dois passos são intercalados até alcançar convergência ou um número máximo de iterações

Redes bayesianas como classificadores

- Como redes bayesianas podem ser utilizadas como classificadores?

Redes bayesianas como classificadores

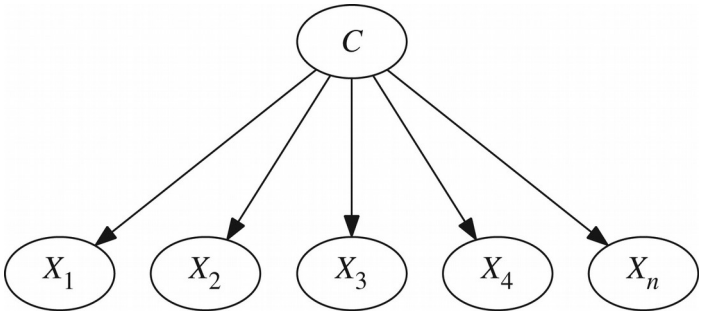
- Como redes bayesianas podem ser utilizadas como classificadores?

  - A classe é uma das variáveis
  - Cada característica é uma variável
  - A variável da classe depende das variáveis das características
  - Você classifica para a classe c_i com maior probabilidade $P(C = c_j \mid x1, x2, \ldots, xn)$ para todo j (na qual as variáveis x1, x2, …, xn são pais de C, não necessariamente exatamente todas as características, dependendo da estrutura da rede).

Redes bayesianas como classificadores

- Problema dessa estratégia?
- Se o número de configurações possíveis é grande, o erro de estimação será grande (ou se a amostra fosse adequadamente grande, seria custoso estimar os parâmetros)
- Alternativas:
  - Naive Bayes Classifier (NBC)
  - Tree Augmented NBC (TAN)

Naive Bayes Classifier



- Variável classe não tem pais
- Cada variável característica

tem apenas um pai: a variável classe

- Estrutura fixa
- Vantagem: parâmetros estimados por um dos métodos de estimação de parâmetros de redes bayesianas
- Desvantagem: assume independência das características dada a classe
- Resultados razoáveis

Tree Augmented Naive Bayes Classifier

- Extensão do NBC: cada variável característica pode ter no máximo mais uma variável característica como pai
- A estrutura não é dada
- A estrutura pode ser aprendida de forma a ter, juntamente com parâmetros probabilísticos ótimos, máxima verossimilhança
- Página 271 de (JENSEN & NIELSEN, 2007)

Características desta metodologia

- Supervisionado ou não-supervisionado?

Características desta metodologia

- Supervisionado ou não-supervisionado?
  - Supervisionado! Você precisa saber a classificação para treinar a rede!

Características desta metodologia

- Supervisionado ou não-supervisionado?
  - Supervisionado! Você precisa saber a classificação para treinar a rede!

- Paramétrico ou não-paramétrico?

Características desta metodologia

- Supervisionado ou não-supervisionado?
  - Supervisionado! Você precisa saber a classificação para treinar a rede!

- Paramétrico ou não-paramétrico?
  - Paramétrico! Normalmente assume-se uma distribuição multinomial dos dados e uma distribuição de Dirichlet como *priori* (e *posteriori*)

Software para redes bayesianas

- JavaBayes – pacote gráfico para inferência em redes bayesianas (Fábio Cozman, da EPUSP): http://www.pmr.poli.usp.br/ltd/Software/javabayes/
- R (pacote DEAL)
- Naive Bayes – pacote e1071 (R)

Trabalho (para 7 de novembro)

Realizar validação cruzada para testar naiveBayes utilizando:

- todas a características
- apenas com os componentes principais
- apenas com as características selecionadas pelo selecionador 1 (e opcionalmente o selecionador 2)

Em cada um, calcular os valores médios de revocação (sensibilidade), precisão e acurácia.

Apresentar os resultados em uma tabela (métricas nas colunas, com/sem seleção de características nas linhas) e discuti-los

Referências

JENSEN, F. V.; NIELSEN, T. D. **Bayesian networks and decision graphs**. Springer. 2nd ed. 2007. cap 6 e 8.

HECKERMAN, D. A Tutorial on learning with bayesian networks. **Technical Report** MSR-TR-95-06. Microsoft Research (disponível no COL)

Slides de aula do Prof. Fabio Nakano