Tema 04

O Classificador Bayesiano

Professora:
Ariane Machado Lima

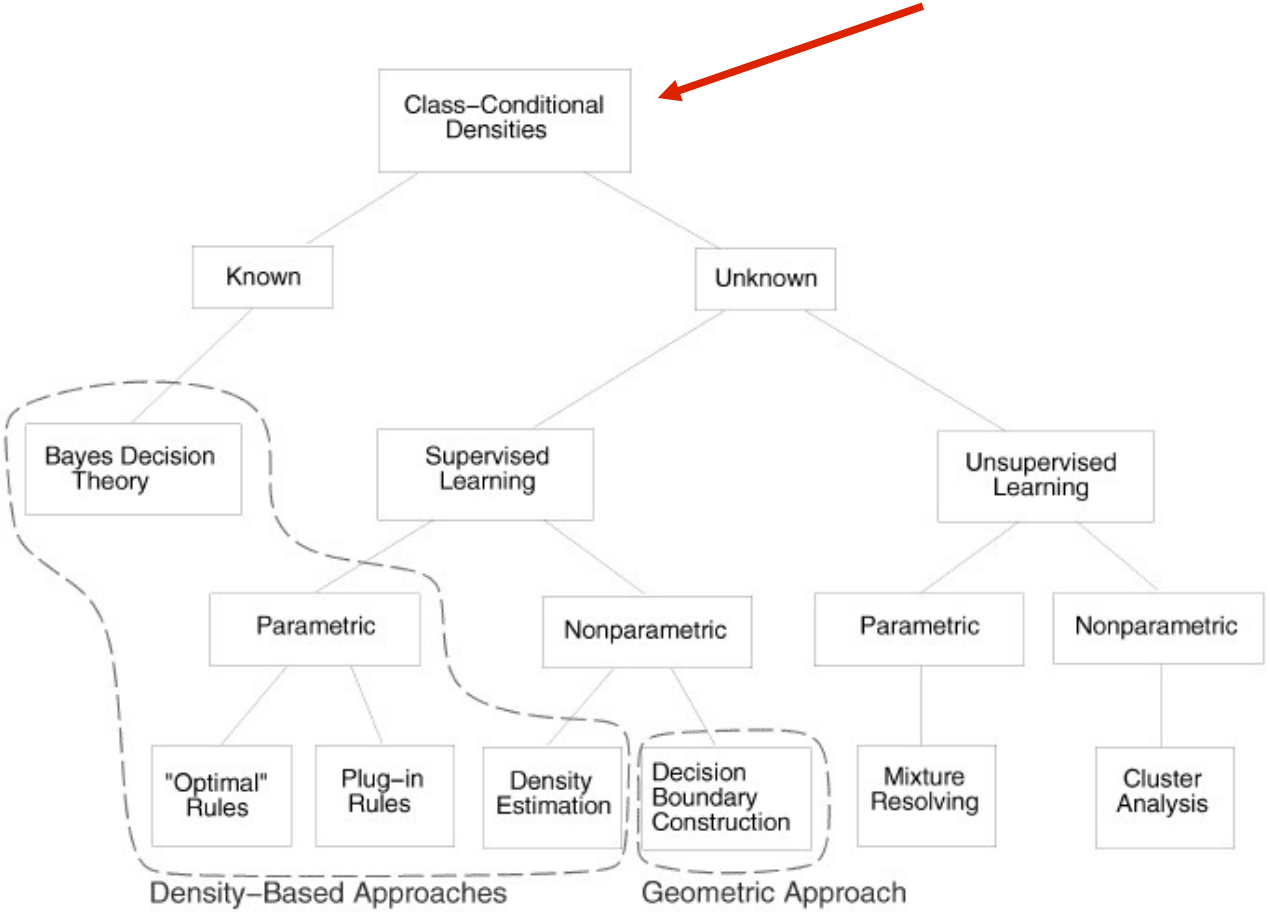
Classificação

Dado um conjunto de elementos, queremos separá-los por classes.

Algumas questões:

- você sabe quantas classes?
- você conhece exemplos dessas classes?

Técnicas de Classificação



[JAIN et al, 2000]

O que são essas densidades condicionais da classe?

- O que quer dizer densidade neste contexto?
- Função densidade de probabilidade

Revisão (rápida) de probabilidade e estatística

- X é uma variável aleatória (cujo valor pode ser visto como o resultado de um experimento) que assume valores sobre Ω (espaço amostral)
- Função de probabilidade: p(x) = P(X = x)
- Função de distribuição de probabilidade: F(x) = P(X<=x)

Revisão (rápida) de probabilidade e estatística

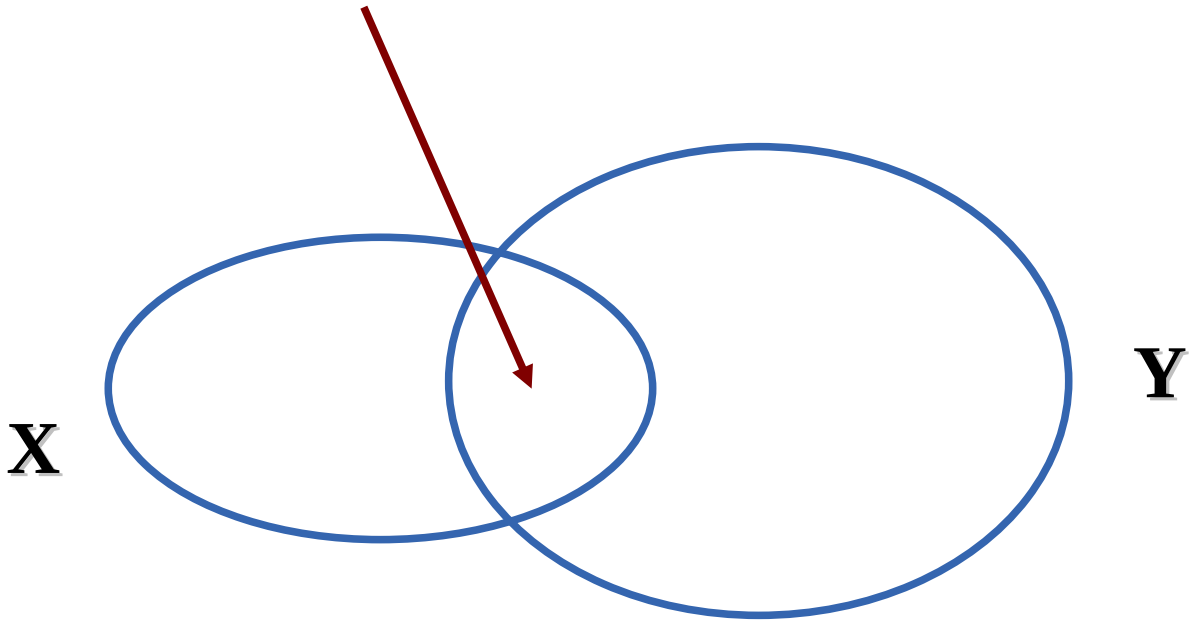
- Se X é discreta (assume apenas valores inteiros):
  - $F(x) = \sum_{i:x_i<=x} P(X=x_i)$
- Se X é contínua (pode assumir valores reais não inteiros):
  - $F(X) = \int_{-\infty}^{x} f(t)\,dt$, sendo f a função densidade de probabilidade de X
  - Ou seja, a densidade só vira probabilidade quando integrado em um intervalo

Revisão (rápida) de probabilidade e estatística

- Exemplos de distribuições discretas:
  - Bernoulli (lançamento de uma moeda)
  - Binomial (vários Bernoulli)
  - Multinomial (generalização da Binomial para k possíveis resultados)
  - Poisson (vista como uma Binomial de eventos raros)
- Exemplos de distribuições contínuas
  - Normal
  - Exponencial (intervalo entre dois sucessos consecutivos de uma Poisson)
  - Gumbel (EVD)

Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y

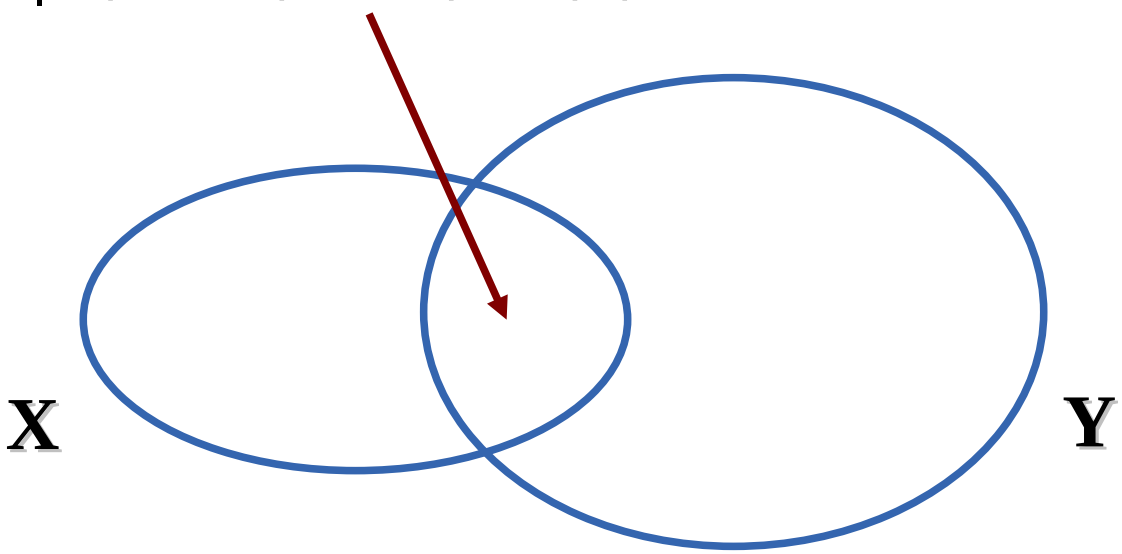


Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (a ∈ X | a ∈ Y) = P ([a ∈ X] ∩ [a ∈ Y]) / P (a ∈ Y)

Revisão (rápida) de probabilidade e estatística

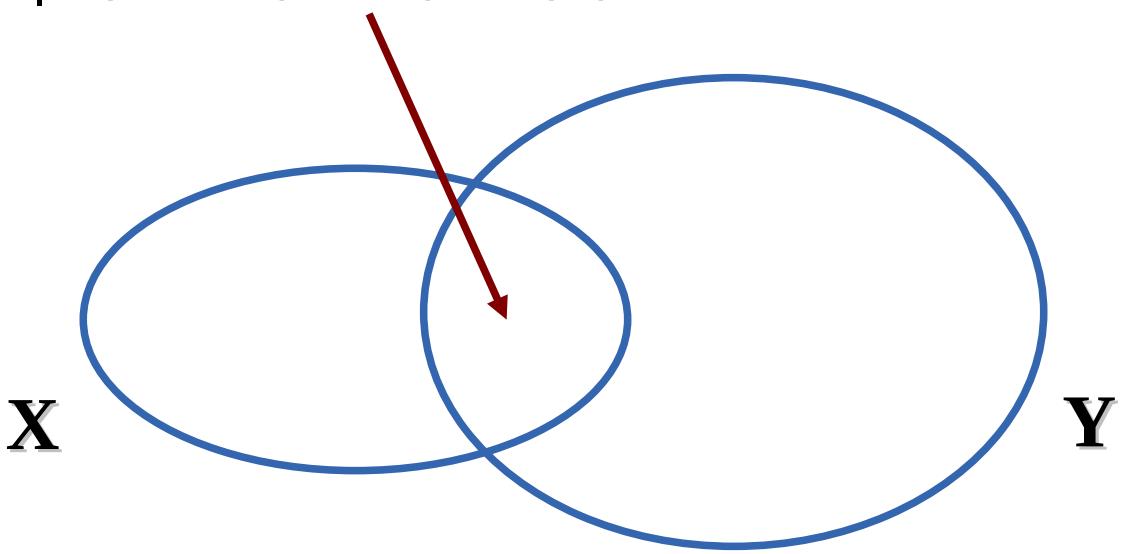
- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (X | Y) = P (X ∩ Y) / P(Y)



Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (X | Y) = P (X , Y) / P(Y)

Distribuição conjunta



Probabilidades condicionais em classificação

- X pode ser um vetor aleatório das características dos elementos que eu observo (que eu quero classificar)
- Y pode ser as classes possíveis: P(Y =yi) é a probabilidade de um elemento ser da classe yi
- X possui uma distribuição dentro de Ω (geral, “no mundo”), e possivelmente uma distribuição diferente “dentro de” (condicionada a) uma dada classe

Ex: * probabilidade de um aluno da USP ser do gênero feminino

* probabilidade de um aluno da USP do curso de SI ser do gênero feminino

* probabilidade de um aluno da USP do curso de pedagogia ser do gênero feminino

Como isso pode nos ajudar na classificação?

Probabilidades condicionais em classificação

- X pode ser um vetor aleatório das características dos elementos que eu observo (que eu quero classificar)
- Y pode ser as classes possíveis: P(Y =yi) é a probabilidade de um elemento ser da classe yi
- X possui uma distribuição dentro de Ω (geral, “no mundo”), e possivelmente uma distribuição diferente “dentro de” (condicionada a) uma dada classe

Ex: * probabilidade de um aluno da USP ser do gênero feminino

* probabilidade de um aluno da USP do curso de SI ser do gênero feminino

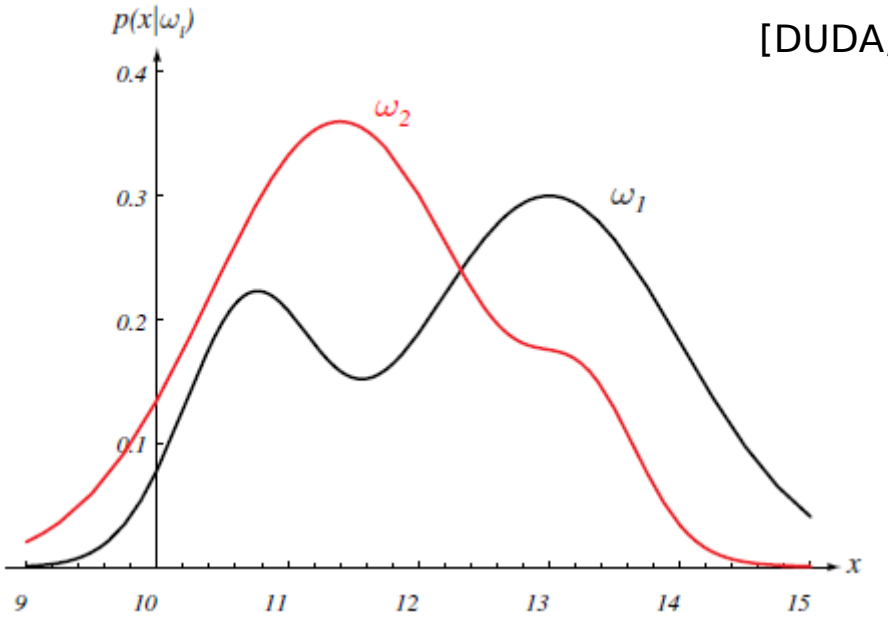
* probabilidade de um aluno da USP do curso de pedagogia ser do gênero feminino

Como isso pode nos ajudar na classificação?

Saber o gênero do aluno pode nos ajudar a prever se ele é de SI ou pedagogia

Probabilidades condicionais

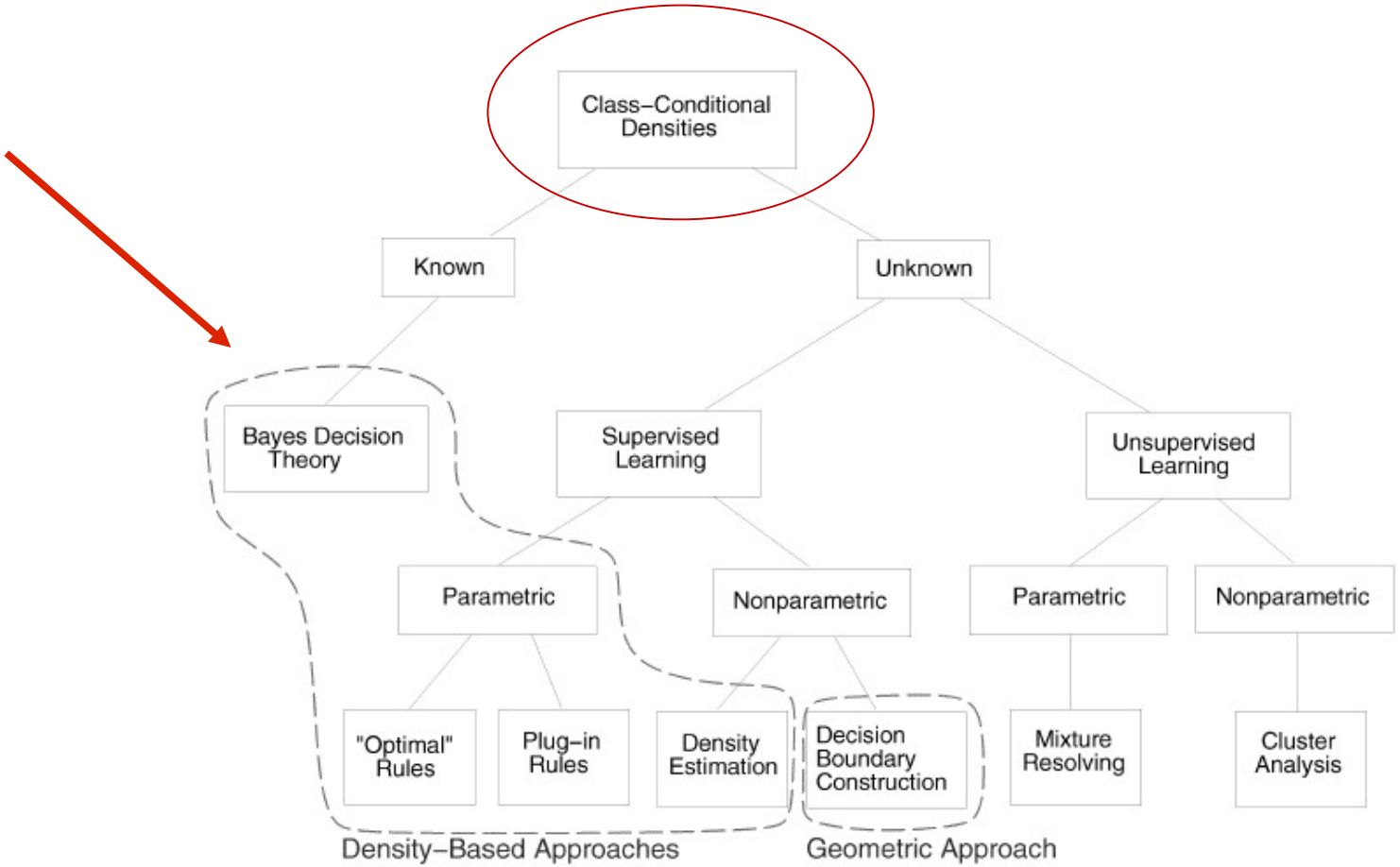
- Ex: densidade condicional da luminosidade dos peixes para as classes robalo e salmão
- P(x|c) : densidade de probabilidade condicional (à classe)
- Ex:



[DUDA, HART & STORK, 2001]

luminosidade

Técnicas de Classificação



[JAIN et al, 2000]

Teoria da Decisão Bayesiana

- Abordagem estatística para o problema de classificação / tomada de decisão
- O classificador é ótimo dentre todas as opções (minimiza o erro), porém...
- É necessário conhecer todas as probabilidades envolvidas e relevantes

Teorema de Bayes

$$P(X \mid Y) = P(X,Y) / P(Y)$$
$$P(Y \mid X) = P(Y,X) / P(X)$$

- $P(X,Y) = P(X|Y)\ P(Y)$
- $P(Y,X) = P(Y|X)\ P(X)$

- Como $P(X,Y) = P(Y,X)$,

- $P(Y|X) = P(X|Y)\ P(Y)/P(X)$

Veremos que na verdade isso tem muito mais significado...

Teoria da Decisão Bayesiana

- Classes envolvidas (ou estados da natureza)
  - Ex: robalo (c1) e salmão (c2)

- Os objetos pertencem a uma dessas classes
- Dado um objeto, ele pode pertencer aleatoriamente a c1 ou a c2
- Logo, c1 e c2 podem ser vistos como valores de uma variável aleatória c

- P(c) : probabilidade *a priori*

Regra de Decisão

- Você tem que chutar qual será o próximo peixe
- Você apenas tem a informação das probabilidades *a priori* de c1 e c2
- Qual peixe você chutaria (a priori, isto é, sem olhar o peixe)?

Regra de Decisão

- Você tem que chutar qual será o próximo peixe
- Você apenas tem a informação das probabilidades *a priori* de c1 e c2
- Qual peixe você chutaria?
  - c1 se P(c1) > P(c2)
  - c2 caso contrário

- Isto é uma regra de decisão

Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?

Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?
- Escolheríamos sempre a mesma classe, mesmo sabendo que há 2...
- Erro?

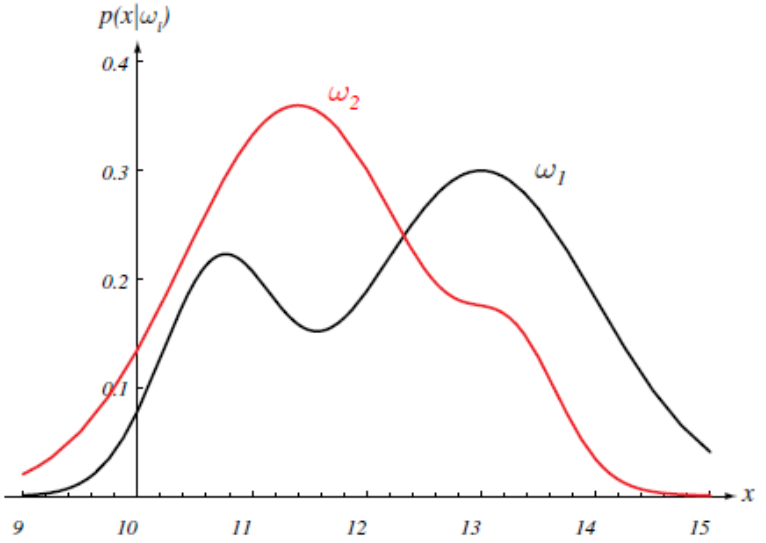
Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?
- Escolheríamos sempre a mesma classe, mesmo sabendo que há 2...
- Erro?
  - Número de peixes classificados errados
  - Aproximadamente P(c2) (se esta *priori* estiver correta)

Probabilidades condicionais

- Se eu sei mais acerca dos peixes (ex: luminosidade, tamanho, etc) – características, para CADA classe.
- P(x|c) : densidade de probabilidade condicional (à classe)
- Ex:

[DUDA, HART & STORK, 2001]



luminosidade

Revisitando Bayes

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- $P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

$P(x) = \sum j\ P(x, cj) = \sum j\ P(x|cj)\ P(cj)$

Revisitando Bayes

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- $P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

$P(x) = \sum j\ P(x, cj) = \sum j\ P(x|cj)\ P(cj)$

- posteriori = verossimilhança * *priori* / evidência

Revisitando Bayes

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- P(ci | x) = P(x|ci) P(ci) / P(x)

Atualização da sua crença (*priori*) após ver os dados = *posteriori*

$P(x) = \sum_j P(x, c_j) = \sum_j P(x|c_j) P(c_j)$

- *posteriori* = verossimilhança * *priori* / evidência

Nova regra de decisão

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

Nova regra de decisão

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

Nova regra de decisão

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

- Erro: P(erro | x) =
  - P(c1| x) se decidirmos por c2
  - P(c2| x) se decidirmos por c1

Nova regra de decisão

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

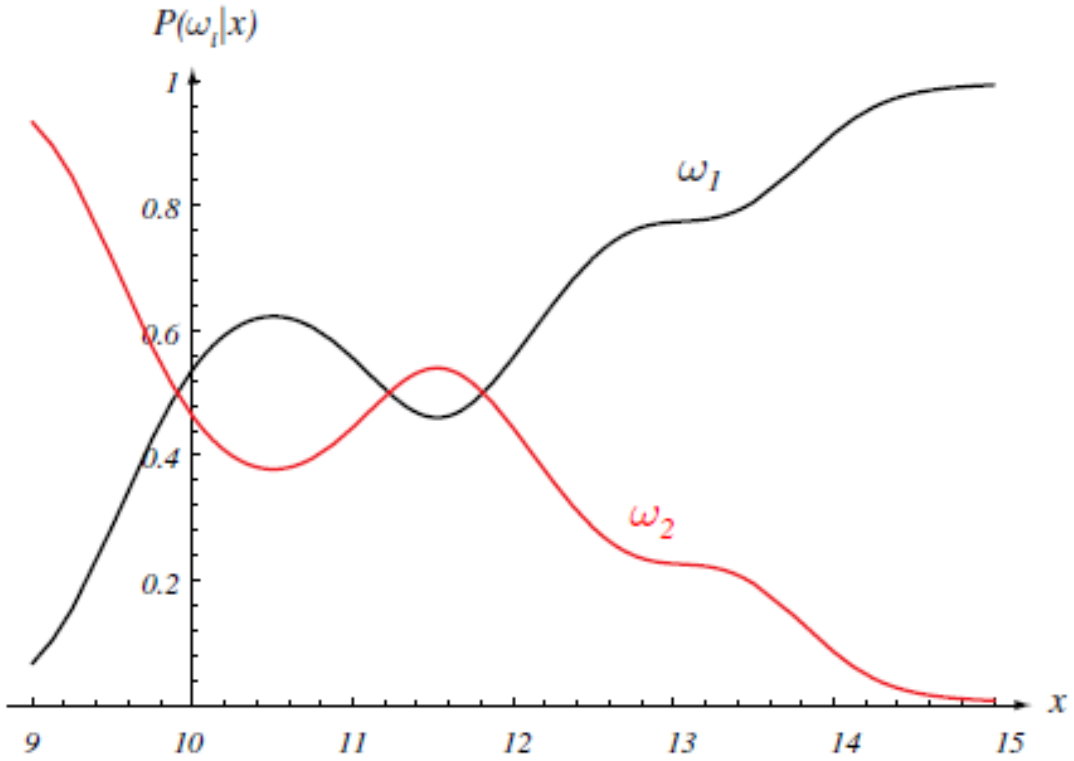
- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

Veremos que a Regra de decisão Bayesiana tem erro mínimo

- Erro: P(erro | x) =
  - P(c1| x) se decidirmos por c2
  - P(c2| x) se decidirmos por c1

Posterioris

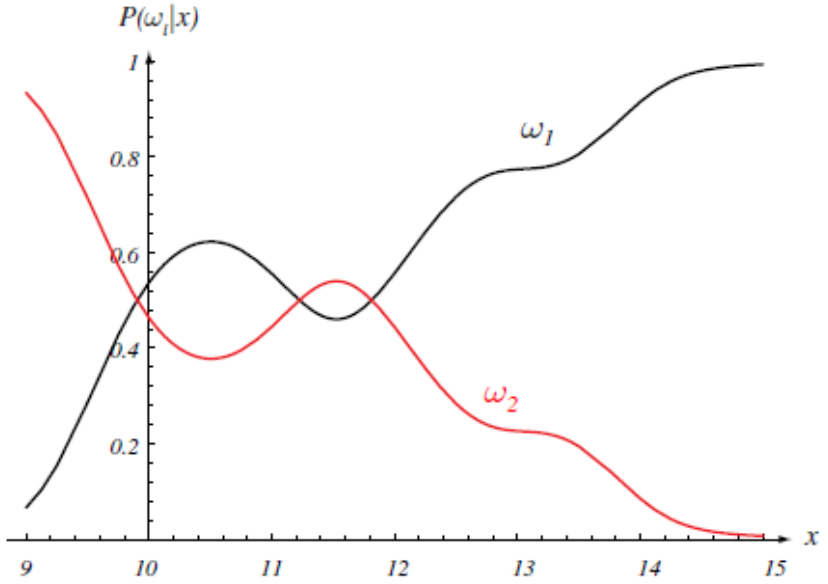
Ex: $P(\omega_1 = 2/3)$ e $P(\omega_2 = 1/3)$



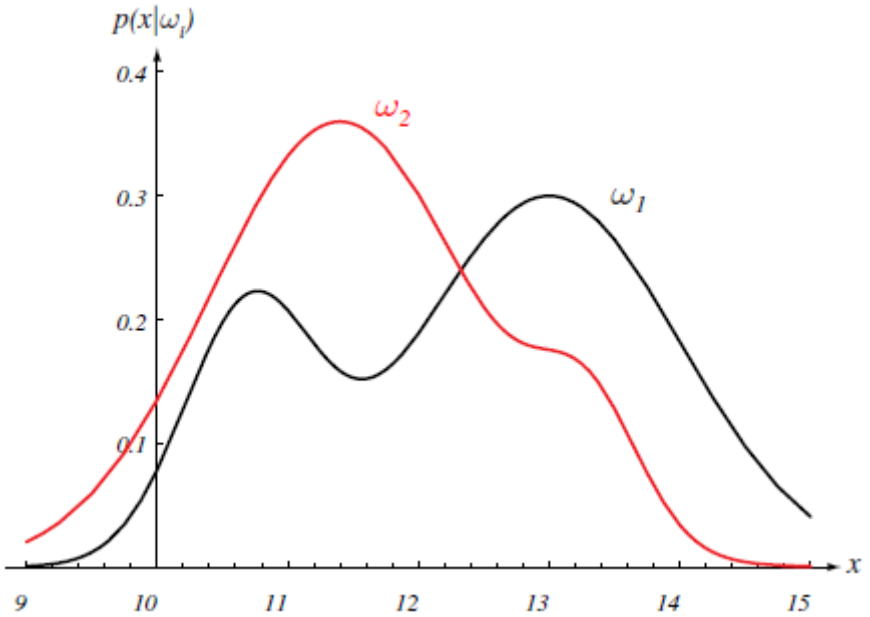
[DUDA, HART & STORK, 2001]

Ex: $P(\omega_1 = 2/3)$ e $P(\omega_2 = 1/3)$

Posterioris



Condicionais (verossimilhanças)



Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

- Se P(c1) = P(c2)?

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

- Se P(c1) = P(c2), a decisão se resume à verossimilhança

Generalizando

(Teoria da Decisão)

Generalizando...

- **x** pode ser um vetor de características
- Pode haver várias classes (c1, ..., ck)
  - Decido pela classe i se P(ci | **x**) > P(cj| **x**), i ≠ j
- Posso realizar *a* ações (ao invés de apenas classificar), α1, ..., αa
  - Por ex, ficar indeciso e não fazer nada
- Cada ação αi tem um custo para cada cj: função perda λ(αi | cj)
- Perda esperada de se tomar uma ação αi ao observar x:

$$R(\alpha i|\mathbf{x}) = \sum_{j=1..c} \lambda(\alpha i \mid cj)\, P(cj|\mathbf{x})$$

Risco condicional

Regra de Decisão Bayesiana

- $\alpha(\mathbf{x})$ é uma função que escolhe αi com base em $\mathbf{x}$
- Risco total: $R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, dx$
- Como minimizamos R?

Regra de Decisão Bayesiana

- $\alpha(\mathbf{x})$ é uma função que escolhe αi com base em $\mathbf{x}$
- Risco total: $R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, dx$
- Como minimizamos R?
- $\alpha(\mathbf{x})$ deve sempre escolher a ação αi que tem o menor risco condicional $R(\alpha i|\mathbf{x})$
  - Regra de decisão Bayesiana
- Este R mínimo (R*) é o risco de Bayes

  melhor resultado possível

Vejamos o exemplo de classificação binária

Classificação Binária

- $\lambda_{ij} = \lambda(\alpha i \mid cj)$, αi = classificar como classe ci
- $R(\alpha 1|\mathbf{x}) = \lambda_{11} P(c1|\mathbf{x}) + \lambda_{12} P(c2|\mathbf{x})$
- $R(\alpha 2|\mathbf{x}) = \lambda_{21} P(c1|\mathbf{x}) + \lambda_{22} P(c2|\mathbf{x})$

Tomamos a ação $\alpha 1$ (isto é, classificamos como pertencente à classe c1) se $R(\alpha 2|\mathbf{x}) > R(\alpha 1|\mathbf{x})$

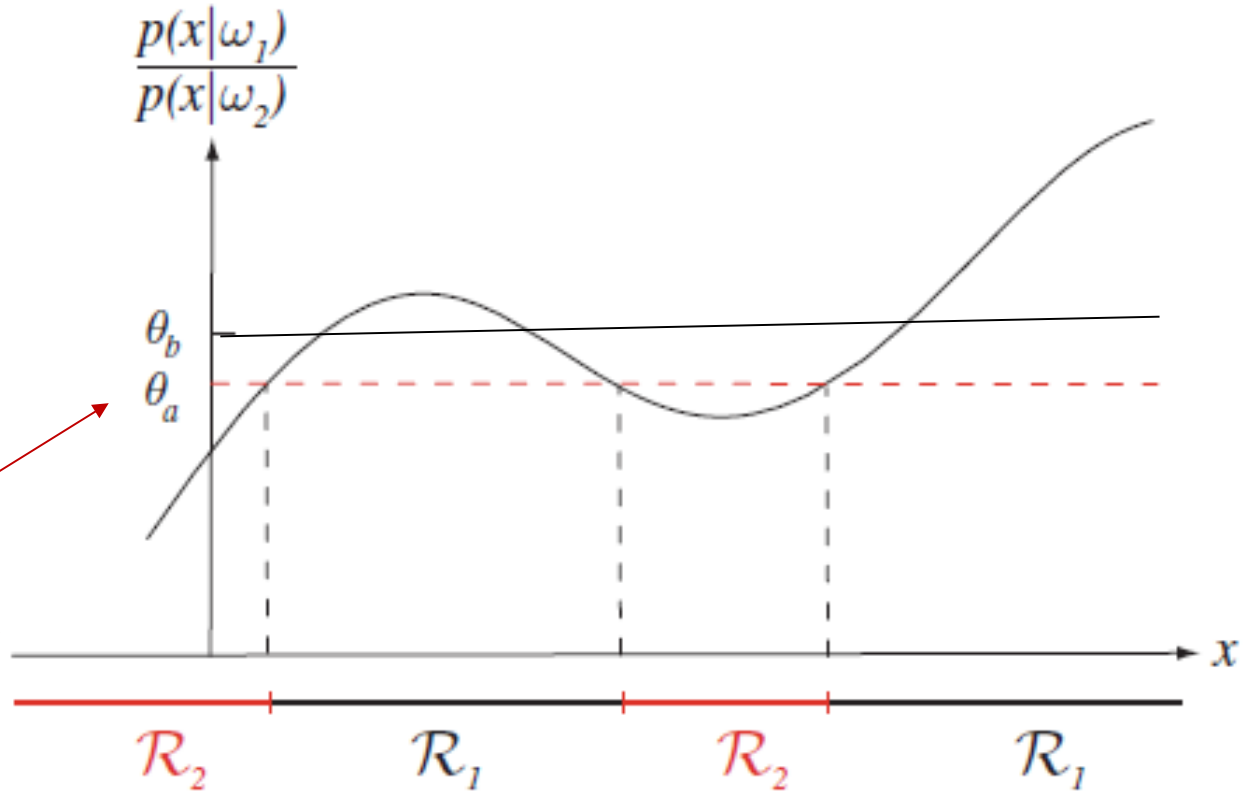
- $\lambda_{21} P(c1|\mathbf{x}) + \lambda_{22} P(c2|\mathbf{x}) > \lambda_{11} P(c1|\mathbf{x}) + \lambda_{12} P(c2|\mathbf{x})$
- $(\lambda_{21} - \lambda_{11}) P(c1|\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(c2|\mathbf{x})$
- $(\lambda_{21} - \lambda_{11}) P(\mathbf{x}|c1) P(c1) > (\lambda_{12} - \lambda_{22}) P(\mathbf{x}|c2) P(c2)$
- $$\frac{P(\mathbf{x}|c1)}{P(\mathbf{x}|c2)} > \frac{(\lambda_{12} - \lambda_{22}) P(c2)}{(\lambda_{21} - \lambda_{11}) P(c1)} = \theta$$ Likehood ratio (Razão de verossimilhança)

Função perda zero-um

Função perda zero-um:

$$\lambda(\alpha i \mid cj) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Razão de verossimilhança



Perda 0-1

prioris iguais

Ri: Região de decisão por i

[DUDA, HART & STORK, 2001]

Razão de verossimilhança

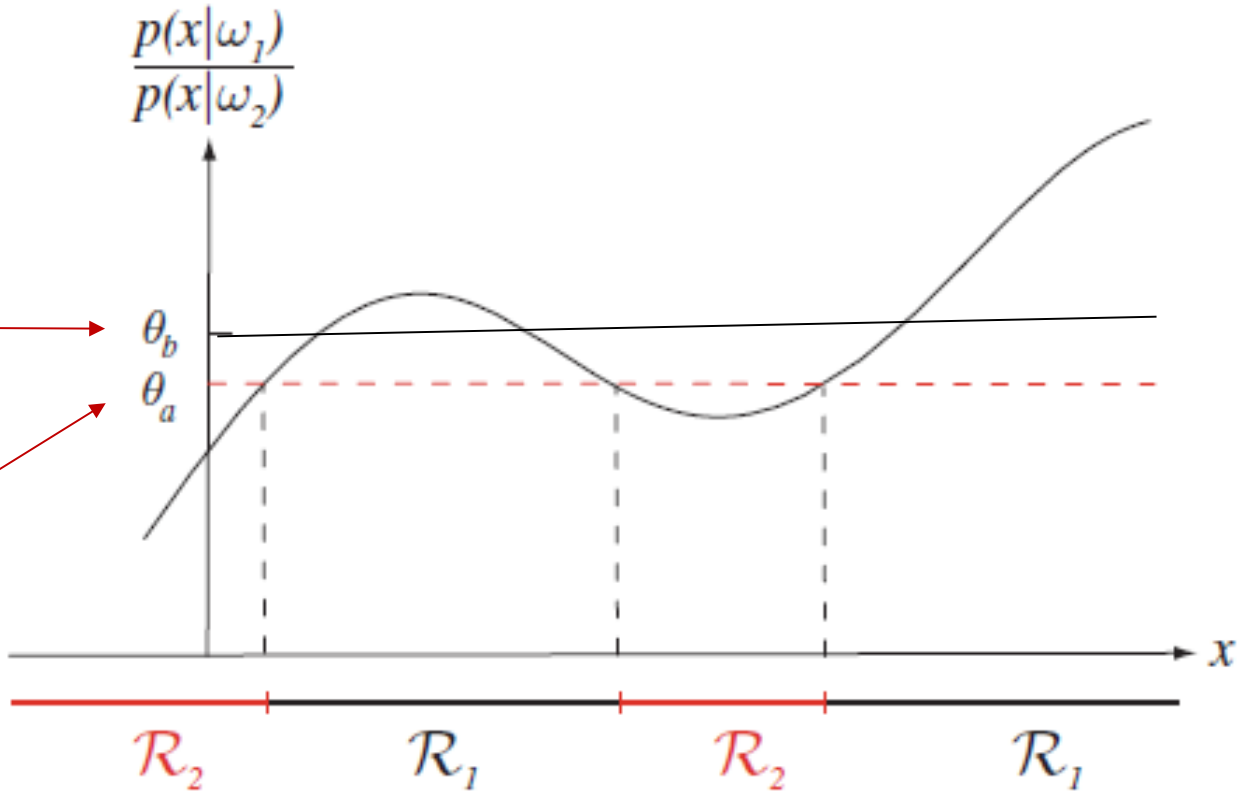
$$\frac{P(\mathbf{x}|c1)}{P(\mathbf{x}|c2)} > \frac{(\lambda_{12} - \lambda_{22})\, P(c2)}{(\lambda_{21} - \lambda_{11})\, P(c1)}$$

$\lambda_{12} > \lambda_{21}$

(R1 se torna menor)

Perda 0-1

prioris iguais



Ri: Região de decisão por i

[DUDA, HART & STORK, 2001]

Classificação de múltiplas classes – taxa de erro mínima

- Função perda zero-um:

$$\lambda(\alpha i \mid cj) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- $R(\alpha i|\mathbf{x}) = \sum_{j=1..c} \lambda(\alpha i \mid cj)\, P(cj|\mathbf{x})$

$$= \sum_{j \neq i} P(cj|\mathbf{x})$$

$$= 1 - P(ci|\mathbf{x})$$

- Decida por ci se $P(ci|\mathbf{x}) > P(cj|\mathbf{x})$, para todo $j \neq i$

Que é a regra de classificação bayesiana baseada na *posteriori*

Resumindo

- Para uma função perda genérica, não há classificador melhor que o teste bayesiano da razão de verossimilhança no sentido de minimizar o custo
- Se a sua função perda é a zero-um, isso se resumo a escolher a classe com maior *posteriori*

Problema

- Normalmente não conhecemos as probabilidades condicionais P(. |ci)
- Neste caso temos que estimá-las a partir de dados conhecidos, o que pode ser complexo
- Mesmo conhecendo todas as probabilidades necessárias, o teste pode ser complexo em termos de tempo e memória

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  -
-

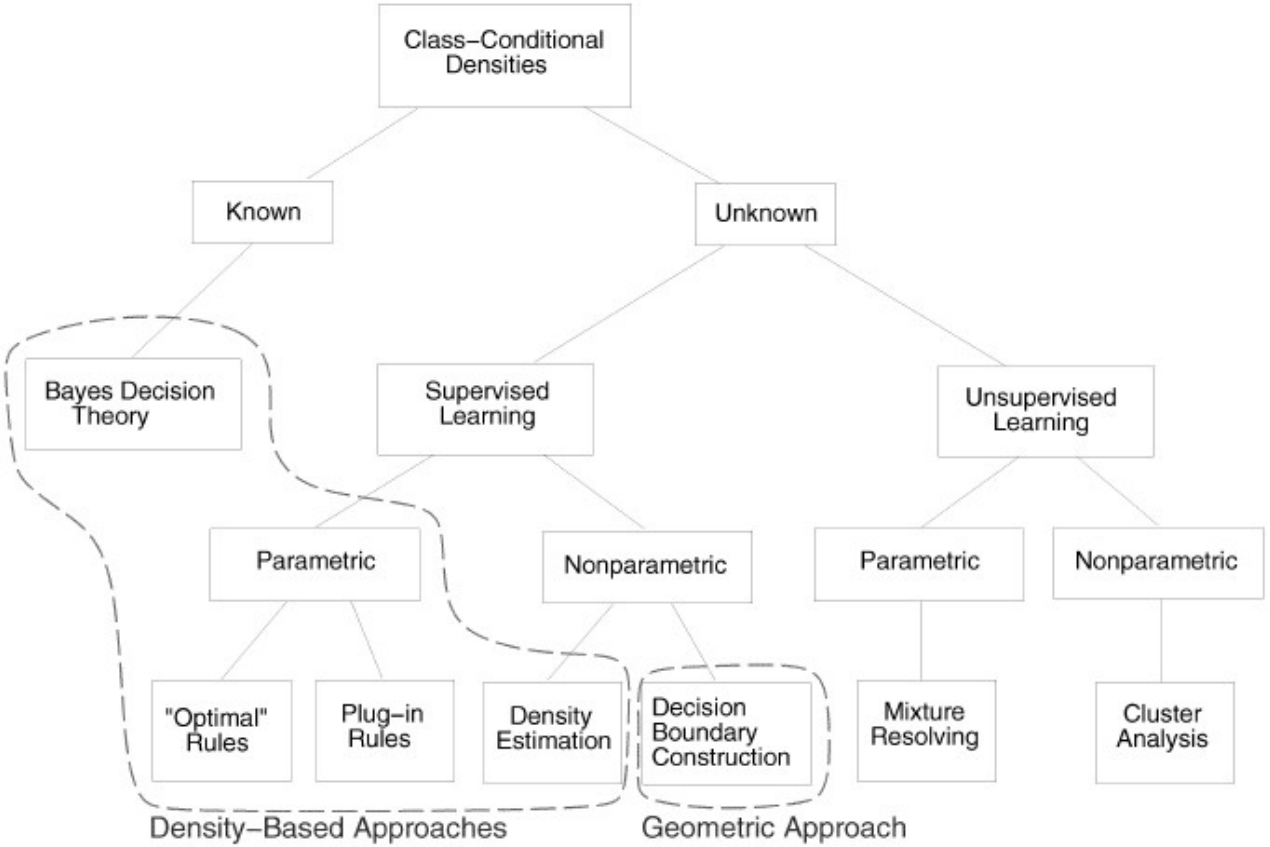
Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
-

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
- Quando não é possível escolher uma forma matemática então opta-se por um CLASSIFICADOR NÃO PARAMÉTRICO

Métodos de Classificação



[JAIN et al, 2000]

Referências até aqui

- DUDA, R.; HART, P.; STORK, D**. Pattern Classification**. John Willey, 2001 (Cap. 2.1 a 2.3)

- JAIN, A.K.; DUIN, R.P.W.; MAO, J. Statistical Pattern Recognition : A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4-37, 2000 (seções 2 e 7)

Redes Bayesianas

Teorema de Bayes

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

Teorema de Bayes

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

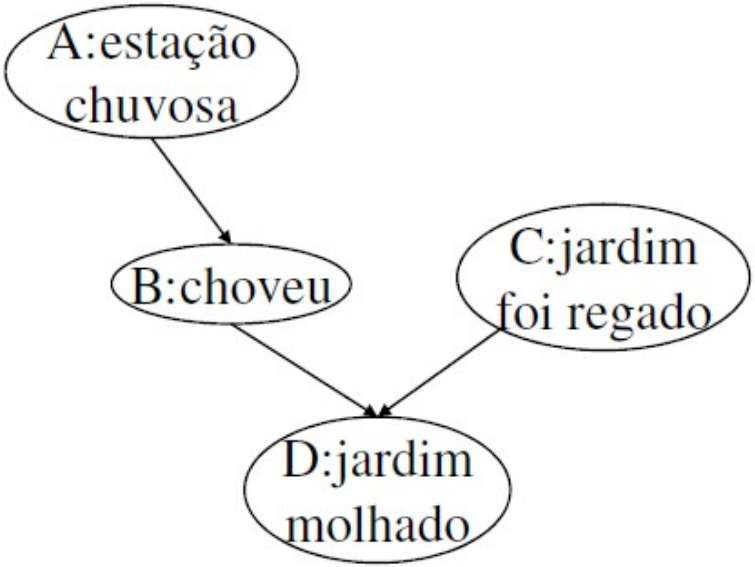
posteriori = verossimilhança * priori / evidência

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

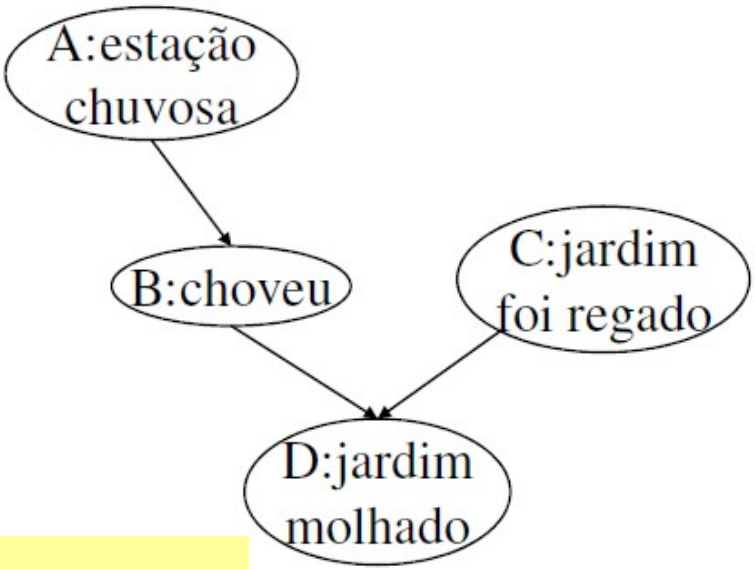
Redes Bayesianas

- Modelo gráfico para representação dessa rede
- Grafo dirigido e acíclico:



Redes Bayesianas

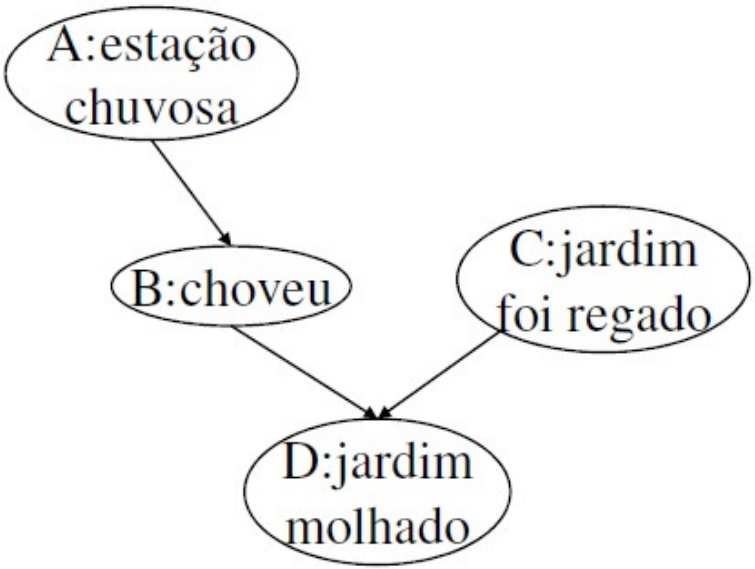
- Modelo gráfico para representação dessa rede
- Grafo dirigido e acíclico:
  - Vértices: variáveis aleatórias



- Discretas ou contínuas
- Atributos visíveis ou variáveis ocultas (ex: a classe)

Redes Bayesianas

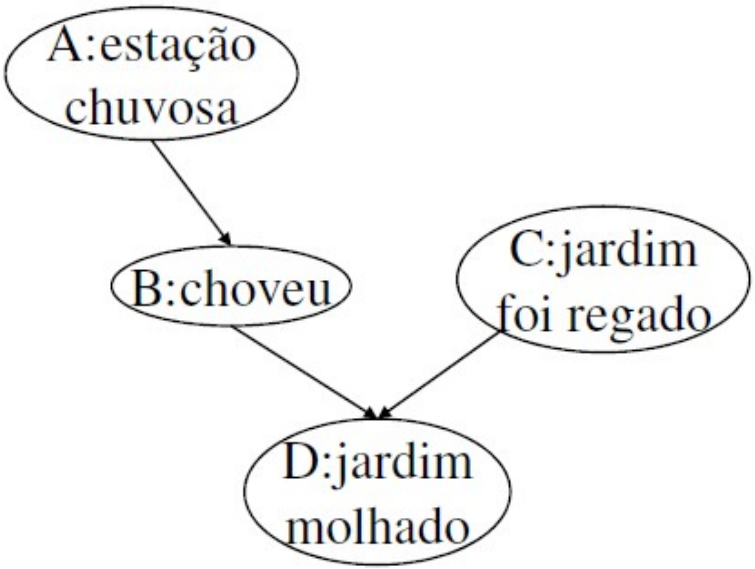
- Modelo gráfico para representação dessa rede
- Grafo dirigido e acíclico:
  - Vértices: variáveis aleatórias
  - Arestas: relações de dependência



Dados dois vértices u e v:
- (u,v) denota a aresta que sai de u e chega em v
- u é dito pai de v
- significa que v depende de u (existe uma distribuição p(v|u))

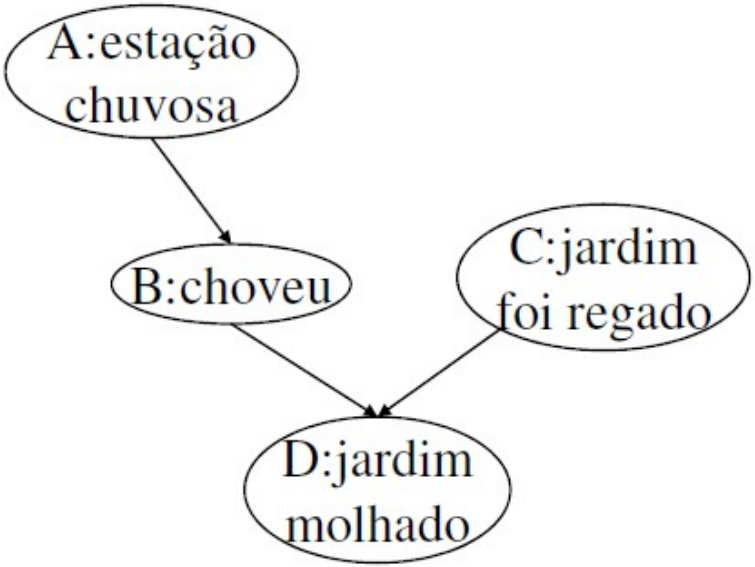
Redes Bayesianas

- 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

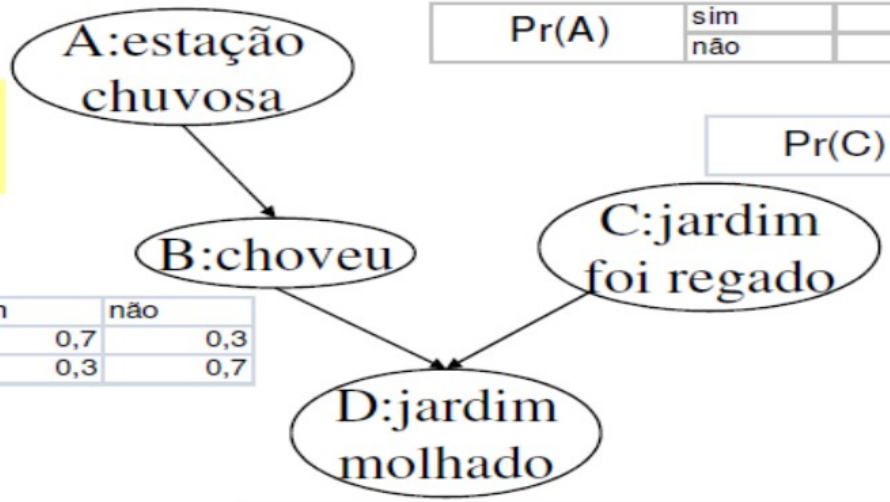
- Par (S, P) onde:
  - S é a estrutura da rede (nós x = {x_1, ..., x_n} e arestas)



Rede Bayesiana - Definição

- Par (S, P) onde:
  - S é a estrutura da rede (nós $x = \{x_1, ..., x_n\}$ 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

Uma tabela de probabilidades condicionais para cada variável



Pr(A)	sim	0,5
	não	0,5

Pr(C)	sim	0,2
	não	0,8

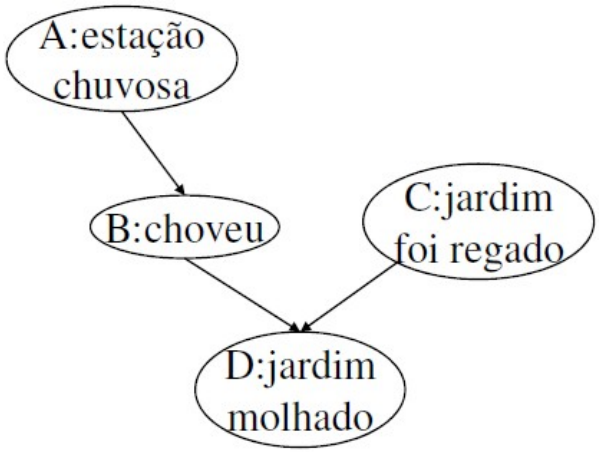
	A	sim	não
Pr(B|A)	sim	0,7	0,3
	não	0,3	0,7

	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

Rede Bayesiana - Definição

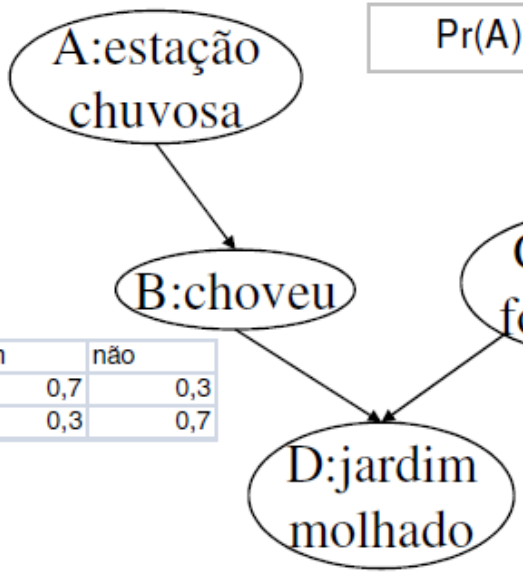
- Par (S, P) onde:
  - S é a estrutura da rede (nós $x = \{x_1, ..., x_n\}$ e arestas)
  - P é um conjunto de distribuições de probabilidades $p(x_i \mid pa(x_i))$, no qual $pa(x_i)$ são os nós pais de x_i
- Probabilidade conjunta da rede:

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



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

Exemplo (supondo que você conhece esses valores)



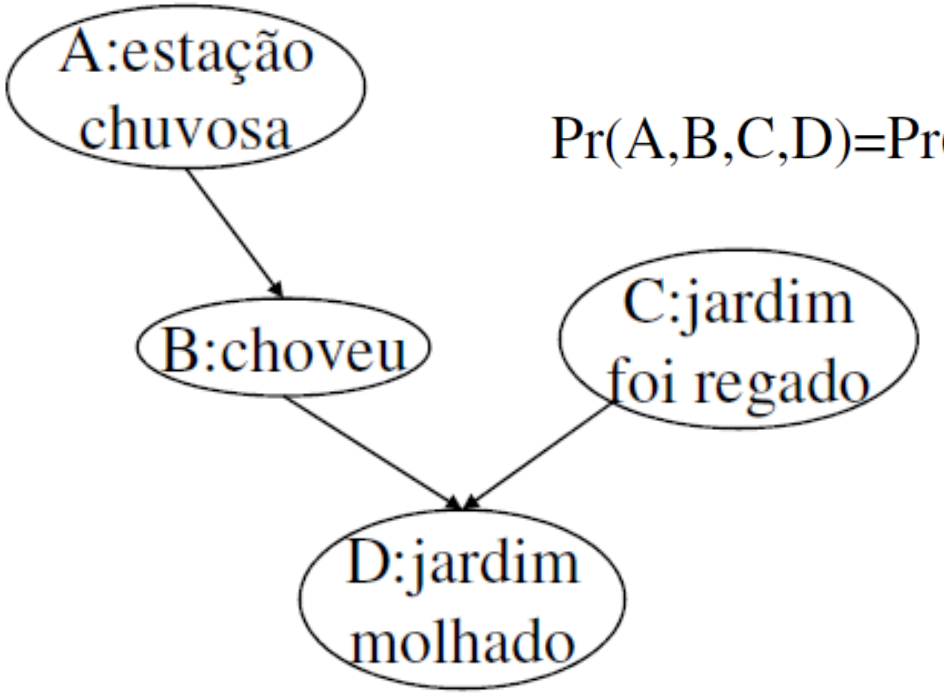
Pr(A)	sim	0,5
	não	0,5

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

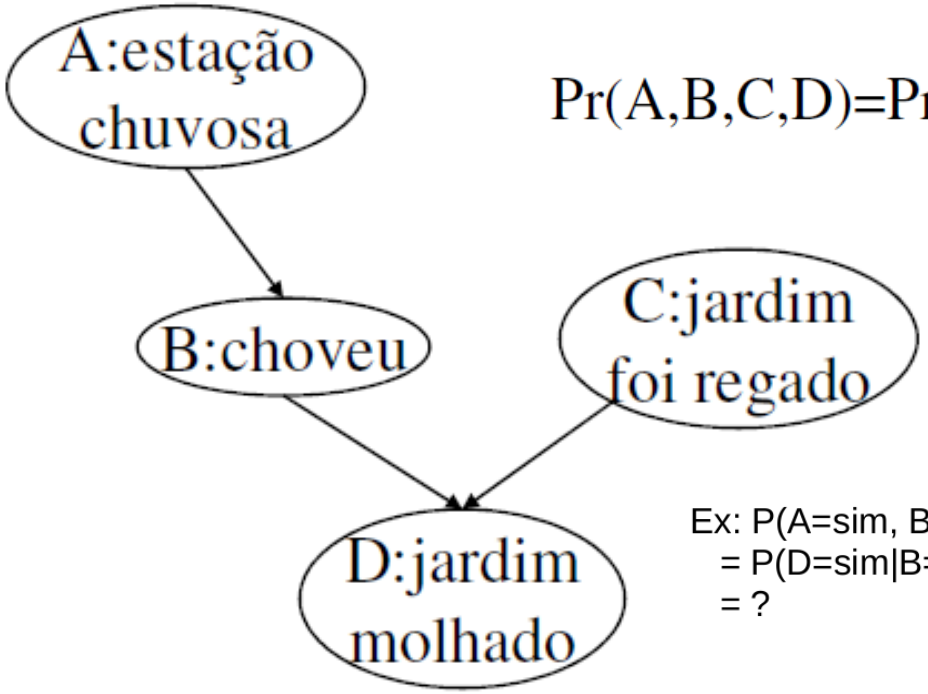
Probabilidade conjunta da rede



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

	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

Probabilidade conjunta da rede

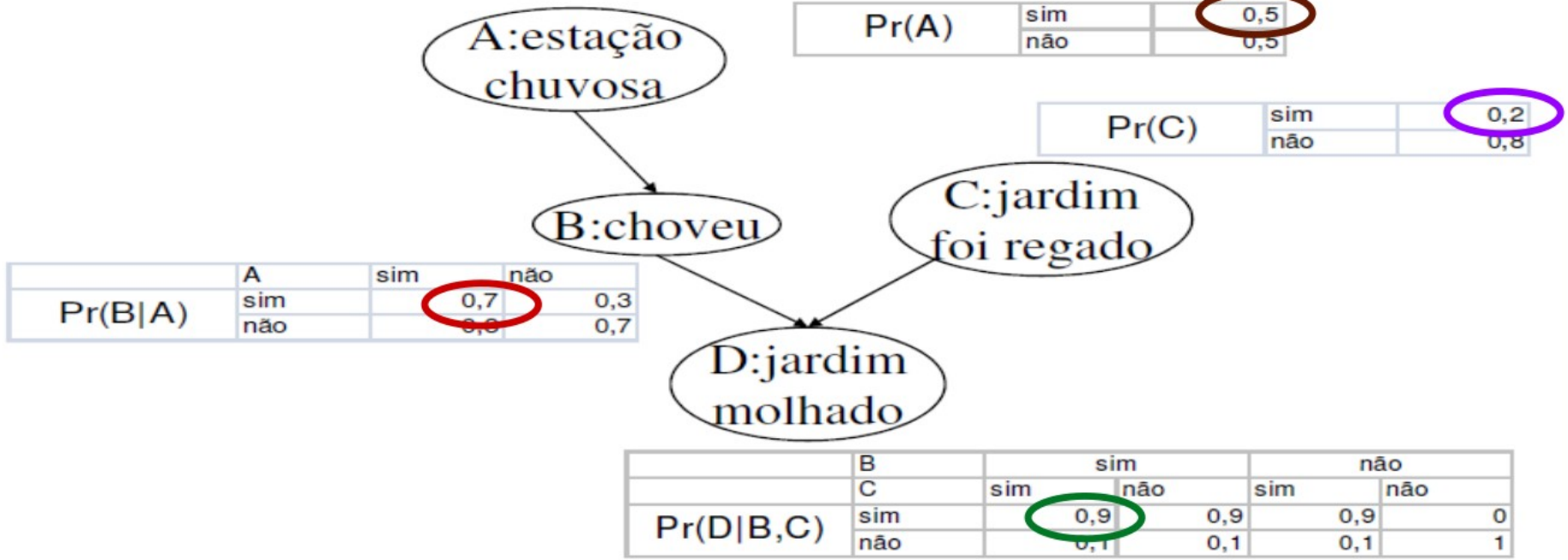


$$Pr(A,B,C,D)=Pr(D|B,C)*Pr(B|A)*Pr(C)*Pr(A)$$

Ex: P(A=sim, B=sim, C=sim, D=sim)
= P(D=sim|B=sim,C=sim)*P(B=sim|A=sim)*P(C=sim)*P(A=sim)
= ?

	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

Probabilidade conjunta da rede



Pr(A)	
sim	0,5
não	0,5

Pr(C)	
sim	0,2
não	0,8

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

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

Ex: P(A=sim, B=sim, C=sim, D=sim)
= P(D=sim|B=sim,C=sim)*P(B=sim|A=sim)*P(C=sim)*P(A=sim)
= 0,9 * 0,7 * 0,2 * 0,5 = 0,063

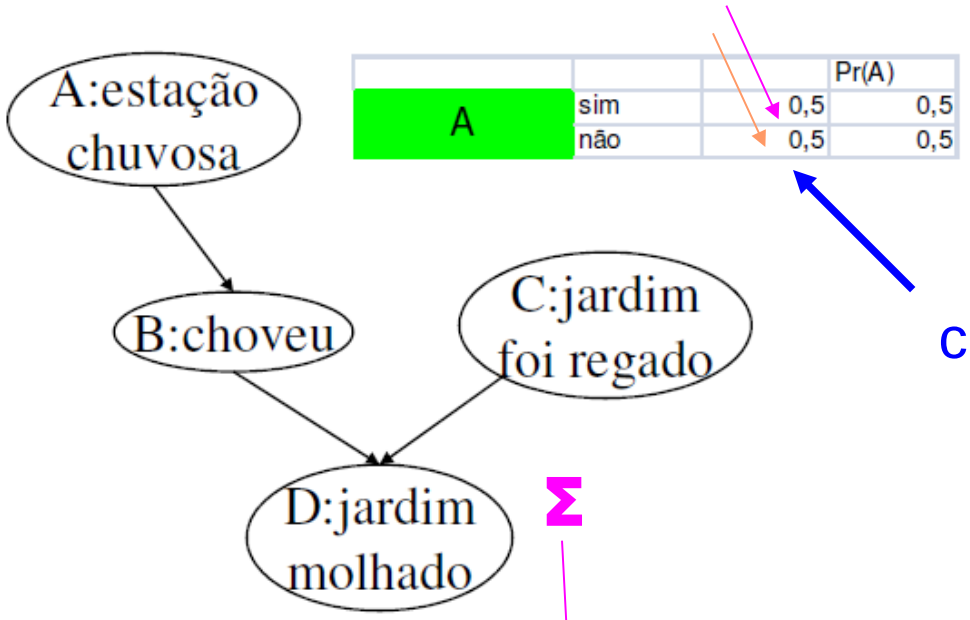
Pr(A,B,C,D)	A: sim, B: sim, C: sim	A: sim, B: sim, C: não	A: sim, B: não, C: sim	A: sim, B: não, C: não	A: não, B: sim, C: sim	A: não, B: sim, C: não	A: não, B: não, C: sim	A: não, B: não, C: não
sim	0,063	0,027	0,252	0	0,027	0,063	0,108	0
não	0,007	0,003	0,028	0,12	0,003	0,007	0,012	0,28

Redes Bayesianas: para quê servem?

Inferência em Redes Bayesianas

- Usadas para fazer inferência
- Redes bayesianas inicialmente como forma de representar conhecimento especialista sobre as incertezas envolvidas no objeto de estudo (mais à frente: como aprender a partir dos dados)
- Usadas para fazer inferência sobre o que você antes desconhecia (probabilidades não representadas DIRETAMENTE no modelo)
- 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)

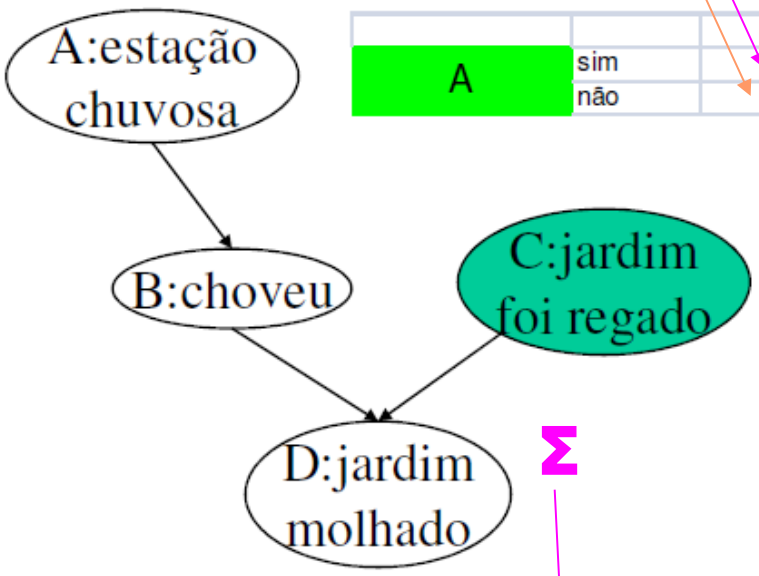


		Pr(A)
A	sim	0,5
	não	0,5

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

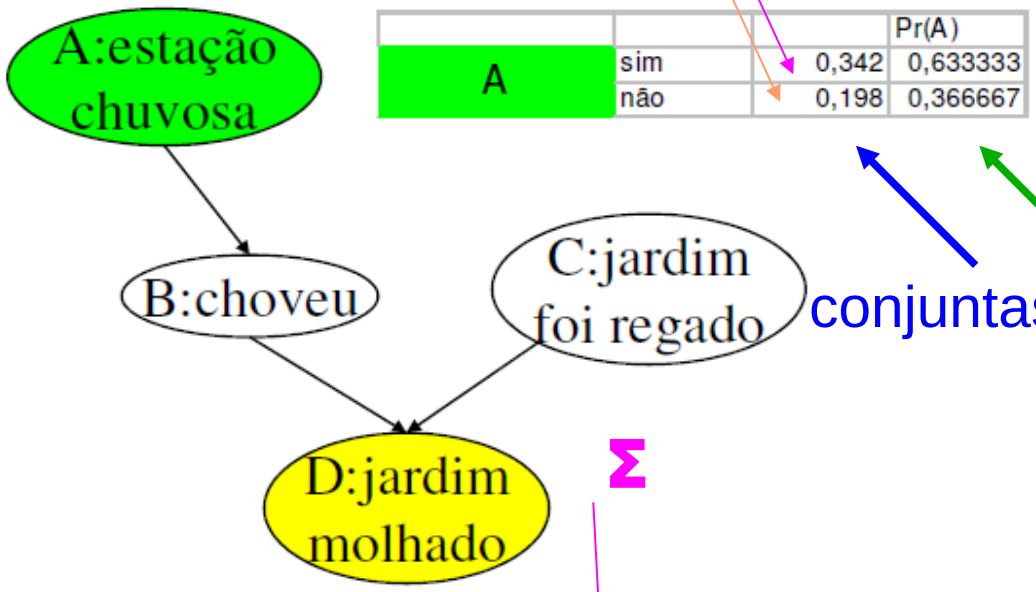
P(A|C=sim)

conjuntas Valores normalizados

Σ Σ

	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

Σ

Σ

	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

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).

ou seja, uma classificação baseada na probabilidade a *posteriori*

Aprendizado de redes bayesianas

O que aprender?

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

Aprendizado de redes bayesianas

- Pode-se criar uma rede bayesiana apenas com conhecimento especialista, ou aprendê-la a partir dos dados
- 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.

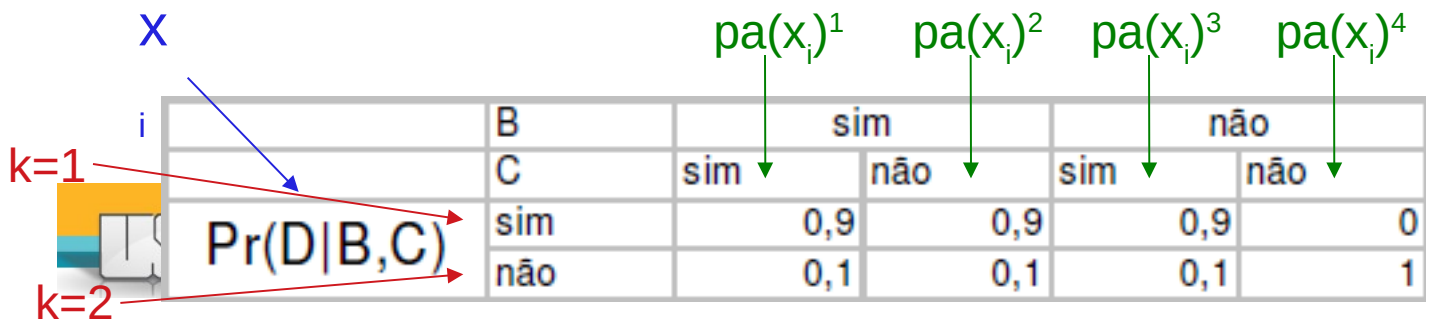
Aprendizado das probabilidades

Rede Bayesiana M = (S, **θ**), sendo S a estrutura da rede e **θ** o vetor de todos os $\boldsymbol{\theta}_i$, sendo $\boldsymbol{\theta}_i$ a tabela de probabilidades condicionais da variável xi, ou seja,

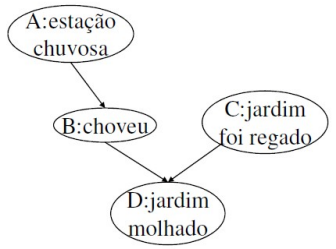
- o conjunto de todos os parâmetros que definem as probabilidades condicionais de cada variável x_i ($P(x_i \mid pa(x_i))$), ou seja,

- $\boldsymbol{\theta}_i$ 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



	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



H

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 (instância) é uma configuração sobre TODAS as variáveis de M (variáveis conhecidas e sem *missing values*)

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 | 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:

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

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

- pode dar divisão por zero! (se no denominador)
- se no numerador, a probabilidade naquele ponto será igual a zero. Boa stimativa?

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 (ex: 1 – correção Laplaciana)
  - 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*

- Utilizado não só em redes bayesianas, mas sempre que há valores ausentes
- Algoritmo iterativo: várias rodadas de dois passos: esperança e maximização:

  Valores iniciais na primeira iteração

  - Esperança: "completa-se" o dataset utilizando o θ' atual para calcular a esperança (média) para os valores ausentes
  - Maximização: usa-se o dataset "completado" para reestimar um novo valor de θ' por máxima verossimilhança (ou máxima a *posteriori*)
- Esses dois passos são intercalados até alcançar convergência ou um número máximo de iterações
- Problema: sofre de mínimos locais

Redes bayesianas 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).

ou seja, uma classificação baseada na probabilidade a *posteriori*

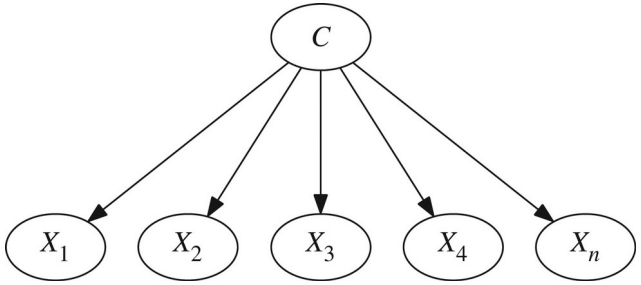
Redes bayesianas como classificadores

- 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)
  - Lembrando que ainda existe o erro de estimação das dependências
- 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 (**assume-se independência entre as características**)

Calcula-se $P(C|\mathbf{X}) = P(X_1, X_2, ..., X_n |C)P(C)$ como sendo $= \Pi_i P(X_i | C) * P(C)$

- Estrutura fixa
- Vantagem: somente os parâmetros precisam ser aprendidos (por um dos métodos de estimação de parâmetros de redes bayesianas_
- Desvantagem: assume independência das características dada a classe
- Resultados razoáveis

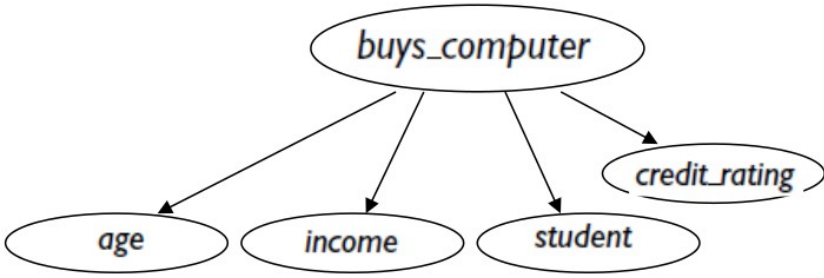
Naive Bayes Classifier - Exemplo

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

(HAN, 2012)

Naive Bayes Classifier - Exemplo

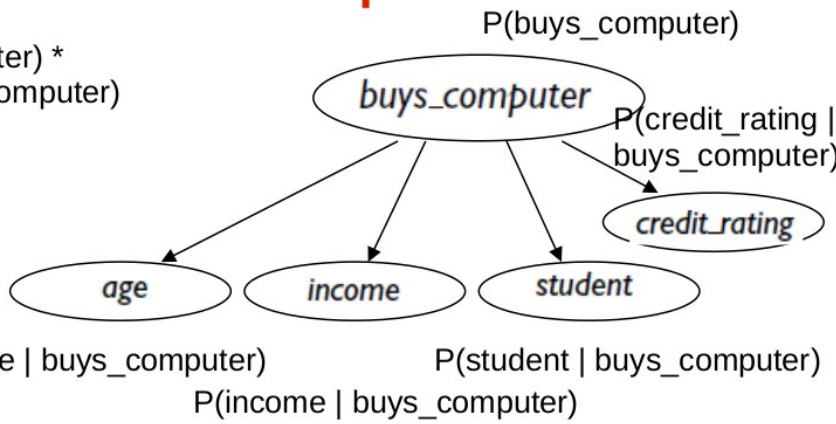


Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)

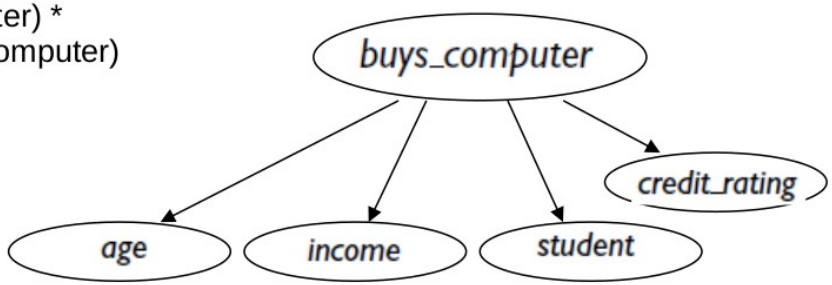


Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

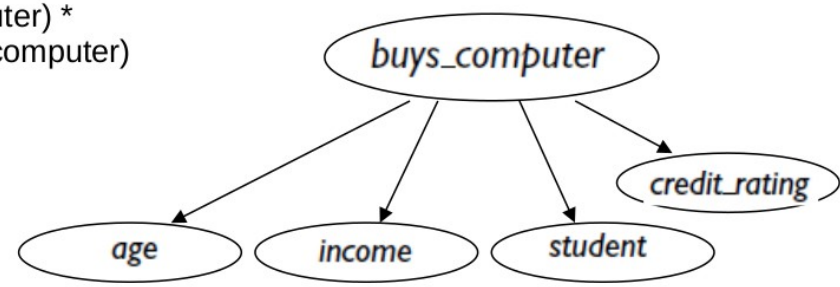
P(age | buys_computer)

	buys_computer	
age	yes	no
youth		
middle-aged		
senior		

Vamos aprender por máxima verossimilhança

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

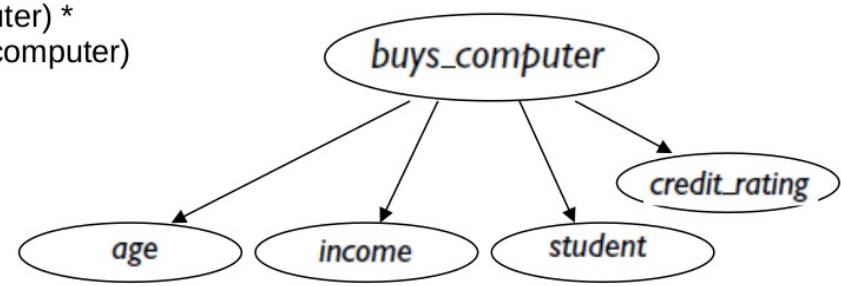
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

P(age | buys_computer)

	buys_computer	
age	yes	no
youth		
middle-aged		
senior		

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

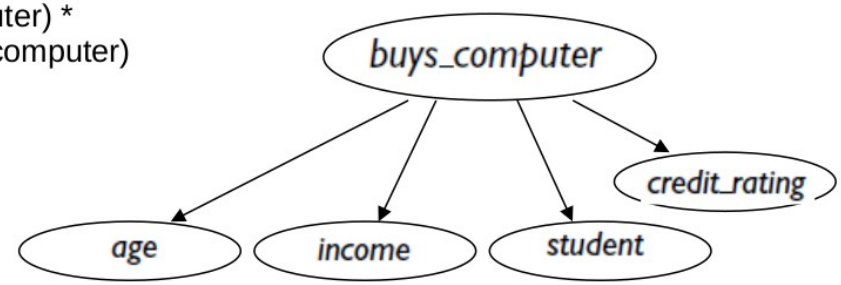
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

P(age | buys_computer)

	buys_computer	
age	yes	no
youth	2/9 = 0,22	
middle-aged	4/9 = 0,44	
senior	3/9 = 0,33	

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

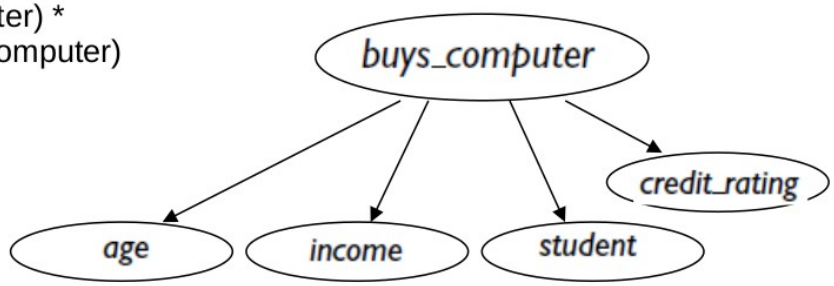
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

P(age | buys_computer)

	buys_computer	
age	yes	no
youth	2/9 = 0,22	3/5 = 0,6
middle-aged	4/9 = 0,44	0/5 = 0
senior	3/9 = 0,33	2/5 = 0,4

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

P(age | buys_computer)

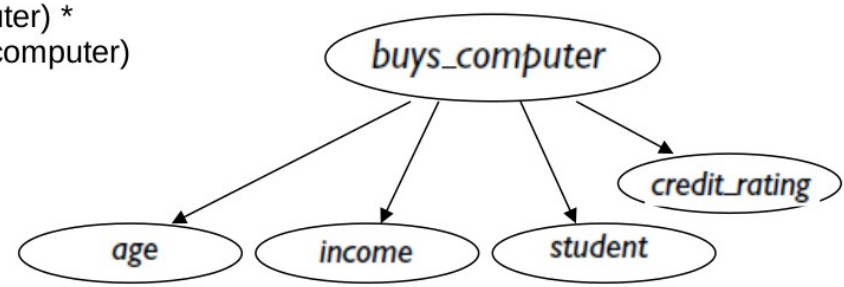
	buys_computer	
age	yes	no
youth	2/9 = 0,22	3/5 = 0,6
middle-aged	4/9 = 0,44	0/5 = 0
senior	3/9 = 0,33	2/5 = 0,4

DEU ZERO!!!
O que aconteceria com
X = (middle-aged, low, no, excellent)?

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)

= 0 !!!!!



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

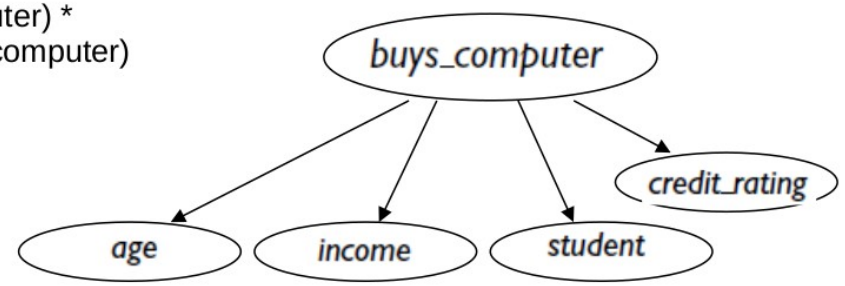
P(age | buys_computer)

	buys_computer	
age	yes	no
youth	2/9 = 0,22	3/5 = 0,6
middle-aged	4/9 = 0,44	0/5 = 0
senior	3/9 = 0,33	2/5 = 0,4

DEU ZERO!!!
O que aconteceria com
X = (middle-aged, low, no, excellent)?

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

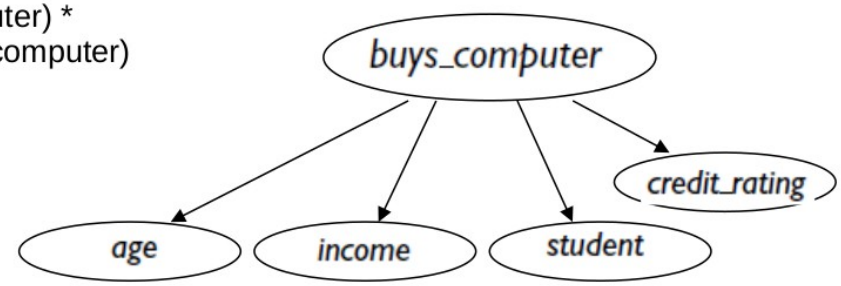
P(age | buys_computer)

	buys_computer	
age	yes	no
youth	2/9 = 0,22	3/5 = 0,6
middle-aged	4/9 = 0,44	0/5 = 0
senior	3/9 = 0,33	2/5 = 0,4

Que tal usar MAP?

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

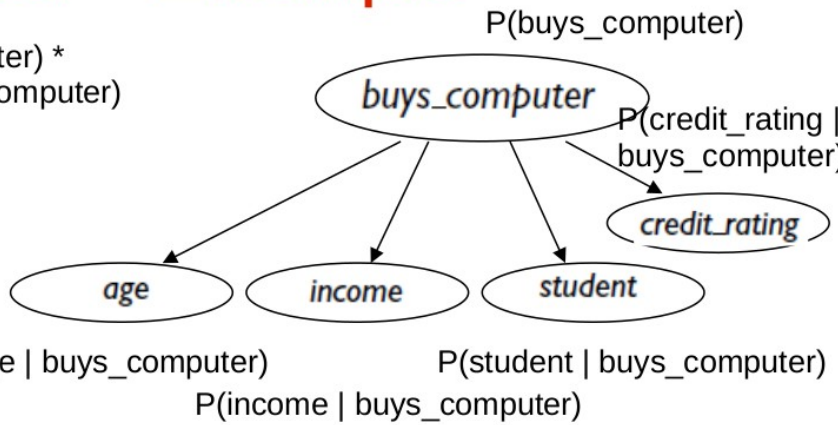
P(age | buys_computer)

	buys_computer	
age	yes	no
youth	3/12 = 0,25	4/8 = 0,5
middle-aged	5/12 = 0,42	1/8 = 0,125
senior	4/12 = 0,33	3/8 = 0,375

MAP com correção Laplaciana

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)

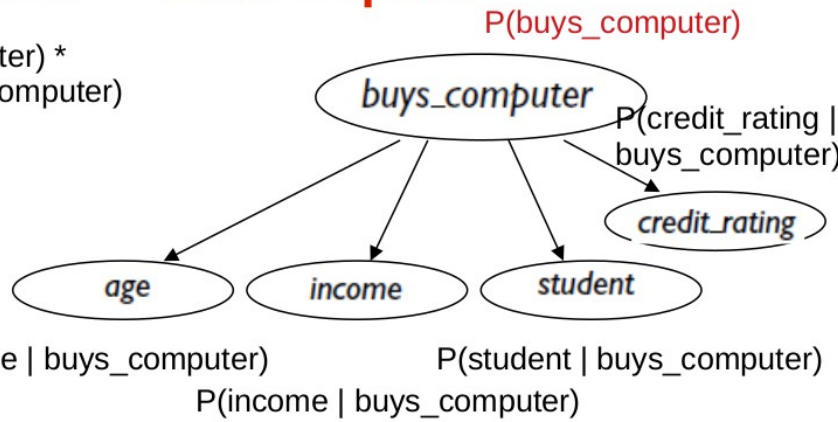


Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

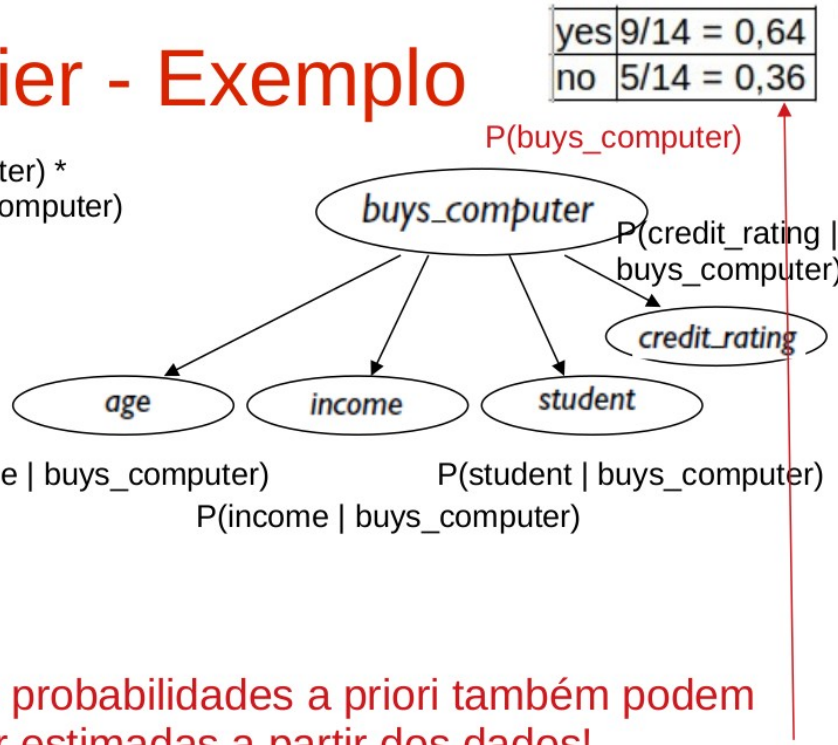
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

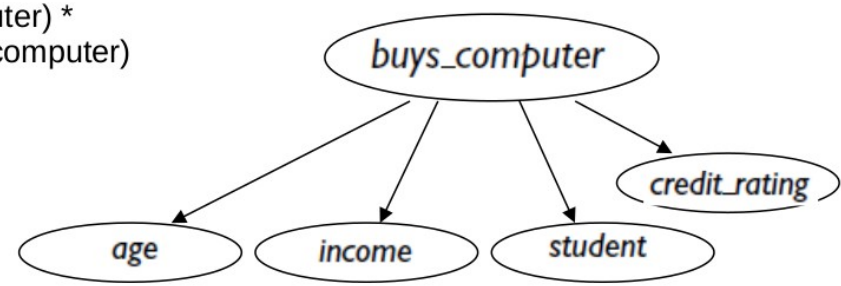
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



As probabilidades a priori também podem ser estimadas a partir dos dados!

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

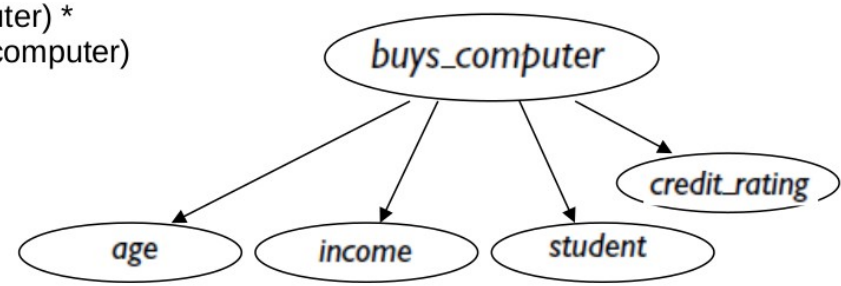
P(age | buys_computer)

	buys_computer	
age	yes	no
youth	3/12 = 0,25	4/8 = 0,5
middle-aged	5/12 = 0,42	1/8 = 0,125
senior	4/12 = 0,33	3/8 = 0,375

EXERCÍCIO: Utilizem esse mesmo algoritmo para estimar as demais probabilidades e brinquem de classificar algumas instâncias

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

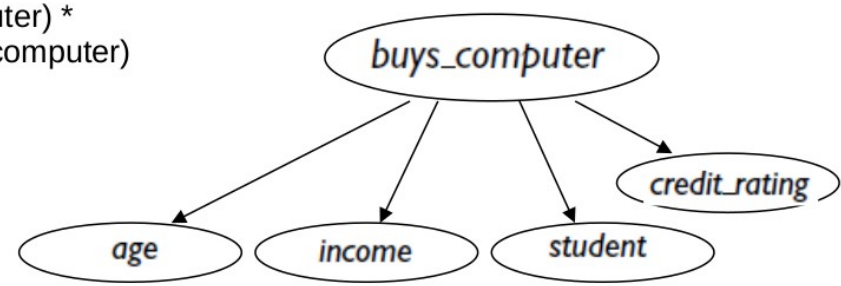
P(age | buys_computer)

	buys_computer	
age	yes	no
youth	3/12 = 0,25	4/8 = 0,5
middle-aged	5/12 = 0,42	1/8 = 0,125
senior	4/12 = 0,33	3/8 = 0,375

Obs: e se tivermos variáveis contínuas?

Naive Bayes Classifier - Exemplo

P(buys_computer | **X**) = P(age | buys_computer) * P(income | buys_computer) * P(student | buys_computer) * P(credit_rating | buys_computer) * P(buys_computer)



Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

P(age | buys_computer)

	buys_computer	
age	yes	no
youth	3/12 = 0,25	4/8 = 0,5
middle-aged	5/12 = 0,42	1/8 = 0,125
senior	4/12 = 0,33	3/8 = 0,375

Obs: e se tivermos variáveis contínuas?
Precisa estimar a função de distribuição
Ex: se for uma distribuição normal, usar os dados (daquela característica) para estimar μ e σ (para cada classe)

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)

Pontos fortes de Redes Bayesianas

- Consegue lidar com dados incompletos (*missing values*)
- Aprendizado de relações causais
- Combina dados e conhecimento *a priori*
- Boa forma de evitar *overfitting* (ao usar estimação MAP)
- Classificação multiclasse é natural

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 ®
- Naive Bayes em Python:

https://scikit-learn.org/stable/modules/naive_bayes.html

Naive Bayes no Python

- Sklearn:
  - CategoricalNB para variáveis categóricas
  - GaussianNB para variáveis numéricas
  - MultinomialNM para classificação de texto

- Para datasets mistos:

https://pypi.org/project/mixed-naive-bayes/

Atividade 4 (para dia 15/9)

Utilizar o classificador Naive Bayes utilizando:

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

Para cada um deles:

- dividir o dataset (já reduzido por PCA ou selecionador) em 80% para treino e 20% para teste (mantendo as proporção das classes em cada parte)
- Usar os 80% para treinar o Naive Bayes e testar nos 20%
- Calcular a acurácia (% de acerto) na amostra de teste

Atividade 4 (para dia 15/9)

Obs:

1- Usar o Naive Bayes adequado para o tipo de variáveis do seu dataset (completo ou reduzido)

2- Ex: Independente do tipo de variáveis do seu dataset original, o dataset reduzido por PCA terá sempre variáveis contínuas!!!!

Referências

HAN, J.; KAMBER, M.; PEI, J. Data mining: Concepts and Techniques. 3rd edition. Elsevier. 2012

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

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

Slides de aula do Prof. Fabio Nakano