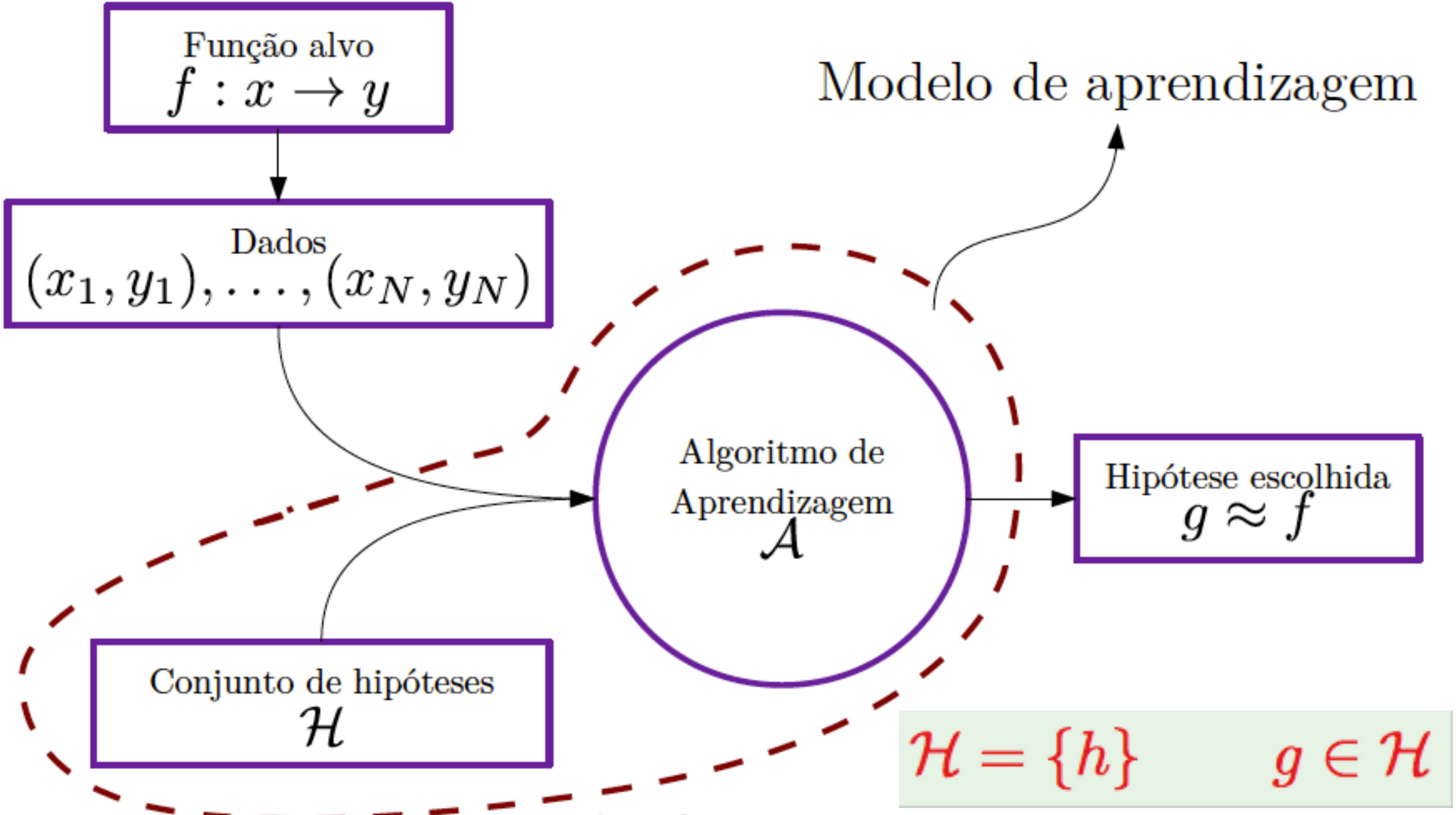
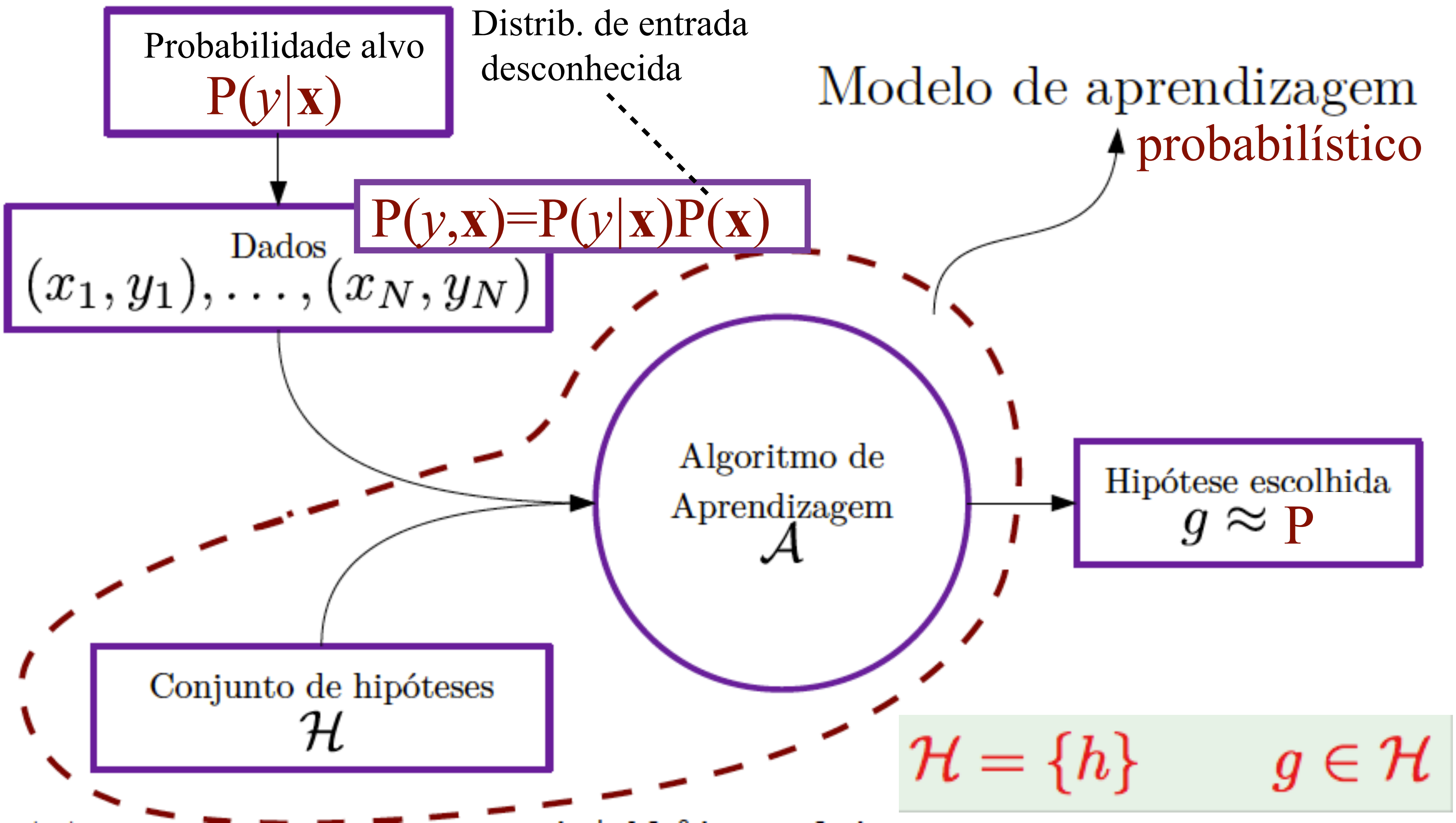
COMP0271: Inteligência Artificial

Naïve Bayes

Professor: Hendrik Macedo

Universidade Federal de Sergipe, Brasil


Função alvo
$f : x \to y$
Modelo de aprendizagem
Dados
$(x_1, y_1), \ldots, (x_N, y_N)$
Algoritmo de Aprendizagem
$\mathcal{A}$
Hipótese escolhida
$g \approx f$
Conjunto de hipóteses
$\mathcal{H}$
$\mathcal{H} = \{h\}$
$g \in \mathcal{H}$



Probabilidade alvo
P(y|x)
Distrib. de entrada
desconhecida
Modelo de aprendizagem
probabilístico
Dados
$(x_1, y_1), \ldots, (x_N, y_N)$
P(y,x)=P(y|x)P(x)
Algoritmo de
Aprendizagem
$\mathcal{A}$
Hipótese escolhida
$g \approx$ P
Conjunto de hipóteses
$\mathcal{H}$
$\mathcal{H} = \{h\}$
$g \in \mathcal{H}$

Regra de Bayes

classe dados

$$P(c \mid d) = \frac{P(d \mid c) P(c)}{P(d)}$$

classe mais provável (*MAP - maximum a posteriori*)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

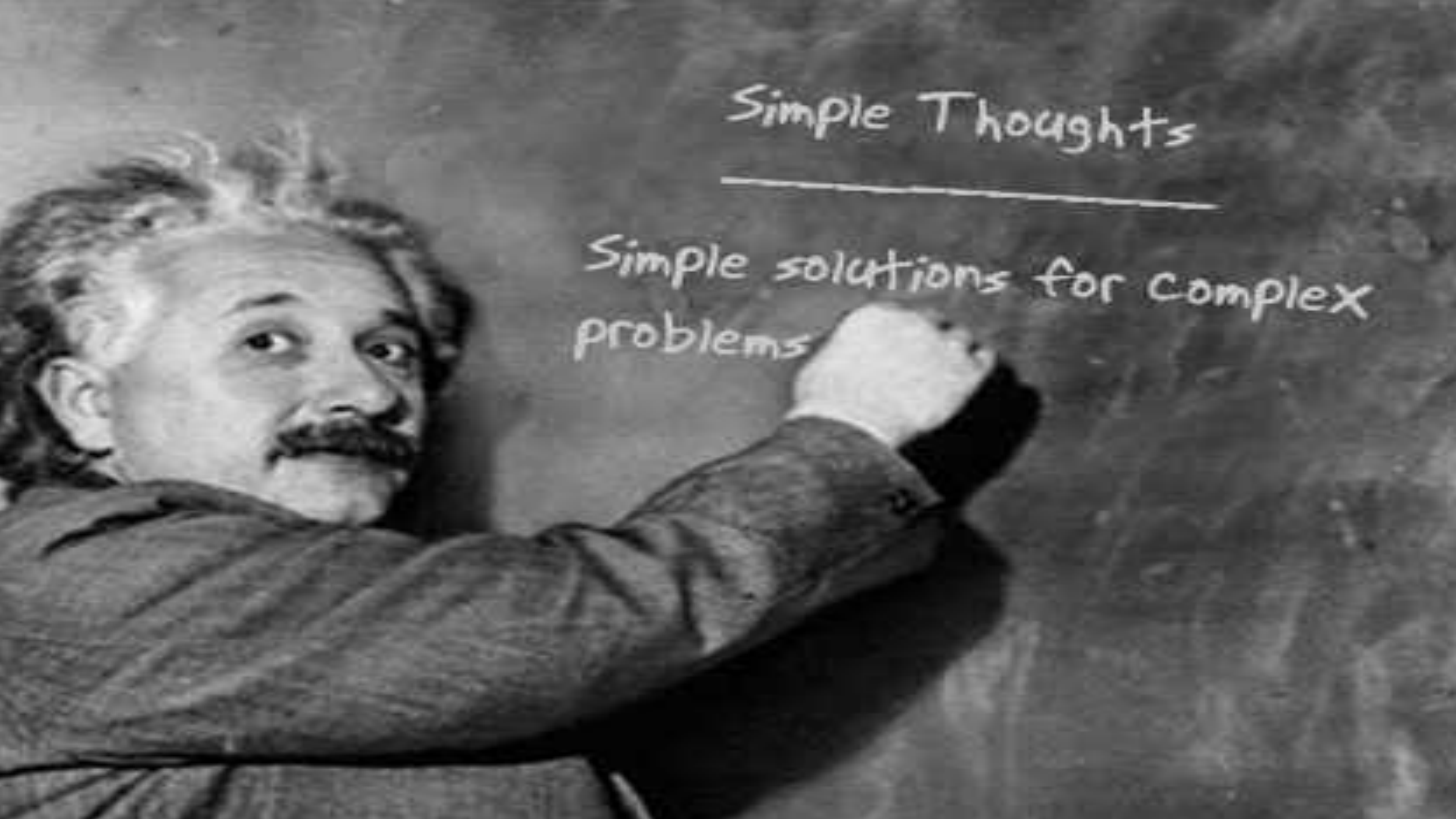
$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

classe mais provável (*MAP - maximum a posteriori*)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

difícil de aprender !


Simple Thoughts
Simple solutions for complex problems

Modelo *Naïve Bayes*

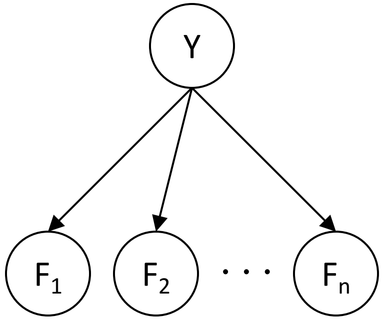
- Assume-se independência condicional
  - atributos x_i são independentes entre si dada a classe

$$P(x_1,\ldots,x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

classe mais provável (*MAP - maximum a posteriori*)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} \; P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c) \bullet P(c)$$

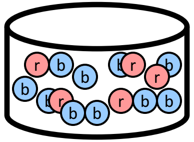
Modelo de aprendizado *Naïve Bayes*



$$P(\mathsf{Y}, \mathsf{F}_1 \ldots \mathsf{F}_n) = \quad P(\mathsf{Y}) \prod_i P(\mathsf{F}_i | \mathsf{Y})$$

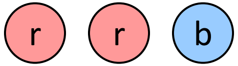
Parâmetros do modelo e denotados por $\boldsymbol{\theta}$

Aprender θ



- Para cada saida x dos dados de treinamento …

$$P_{\mathsf{ML}}(x) = \frac{\mathsf{count}(x)}{\mathsf{total\ samples}}$$



$$P_{\mathsf{ML}}(\mathsf{r}) = 2/3$$

  - Estimativa que maximiza a *verossimilhança (likelihood) dos dados*

$$L(x, \theta) = \prod_i P_\theta(x_i)$$

$$P(\mathsf{Y}, \mathsf{F}_1 \ldots \mathsf{F}_n) = P(\mathsf{Y}) \prod_i P(\mathsf{F}_i | \mathsf{Y})$$

Aprender θ

- Atributos discretos

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

instâncias total

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

vezes valor x_i aparece em instâncias com classe c_j

instâncias com classe c_j

Aprender θ

- Atributos contínuos
  - Probabilidade condicional modelada com a distribuição gaussiana

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\,\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

Exemplo: Reconhecimento de dígitos

- Características (*Features*): atributos usados para decisão sobre classe
  - Pixels: (6,8) = ON
  - Padrões de forma (Shape): NumComponents, AspectRatio, NumLoops
  - …

0

1

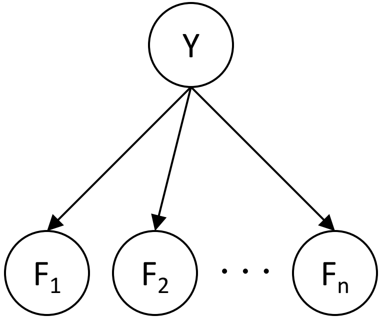
2

1

??

Exemplo: Reconhecimento de dígitos

- Possível abordagem:
  - Uma feature (variável) F_{ij} por cada posição da grid <i,j>
  - Valores: on / off
  - Cada exemplo (input) → vetor de características (feature vector).



$$\rightarrow \langle F_{0,0} = 0 \;\; F_{0,1} = 0 \;\; F_{0,2} = 1 \;\; F_{0,3} = 1 \;\; F_{0,4} = 0 \;\; \ldots F_{15,15} = 0 \rangle$$

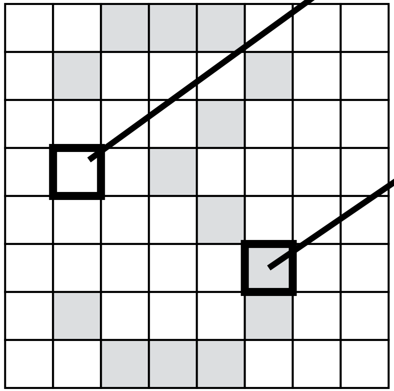
- Modelo Naïve Bayes:

$$P(Y|F_{0,0} \ldots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

Exemplo: Reconhecimento de dígitos

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

Valores advém das contagens dos dados de treinamento

Exemplo: Filtro de spam

- Modelo: $P(Y, W_1 \ldots W_n) = P(Y) \prod_i P(W_i|Y)$
- Parâmetros

$P(Y)$

ham :	0.66
spam:	0.33

$P(W|\text{spam})$

the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

$P(W|\text{ham})$

the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

Valores advém das contagens dos dados de treinamento

Exemplo “brinquedo” do livro

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$P(\text{Play}=Yes) = 9/14$ $P(\text{Play}=No) = 5/14$

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

Aplicação do modelo

$\mathbf{x}$'=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

P(Outlook=*Sunny*|Play=*Yes*) = 2/9
P(Temperature=*Cool*|Play=*Yes*) = 3/9
P(Huminity=*High*|Play=*Yes*) = 3/9
P(Wind=*Strong*|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*No*) = 5/14

P(*Yes*|$\mathbf{x}$'): P(*Sunny*|Yes)P(*Cool*|*Yes*)P(*High*|Yes)P(*Strong*|*Yes*)P(Play=*Yes*) = 0.0053

P(*No*|$\mathbf{x}$'): P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)P(Play=*No*) = 0.0206

$$P(Yes|\mathbf{x}') < P(No|\mathbf{x}') \equiv \ln \frac{P(Yes|\mathbf{x}')}{P(No|\mathbf{x}')} < 0 \quad \Rightarrow \quad \boxed{C_{MAP} = \text{“}No\text{”}}$$

Problema: prob condicional = ZERO

- Se nenhum exemplo contém um determinado valor de atributo a,

  P(X=a | cj) = 0 => $\hat{P}(x_1|\ c_i)\cdots\hat{P}(a_{jk}|\ c_i)\cdots\hat{P}(x_n|\ c_i) = 0$

  - Opções de remédio:
    - *Laplace Smoothing* $P_{LAP,k}(x|y) = \dfrac{c(x,y)+k}{c(y)+k|X|}$
    - *Linear Interpolation* $P_{LIN}(x|y) = \alpha\hat{P}(x|y) + (1.0-\alpha)\hat{P}(x)$

  - Pesquisem!

Modelo *Naïve Bayes*

- Muito rápido
- Robusto contra características irrelevantes
- Bom em domínios com muitas características importantes e bem distribuídas
- É “ótimo” se a suposição de independência é de fato verdadeira, mas funciona bem mesmo que não o seja!