


Aula 05

Métodos probabilísticos de classificação.
Naïve Bayes (cont.).
Classificação de texto.

Aprendizado Bayesiano na classificação

- Assumindo:
 - \mathbf{x} é a entrada, um vetor com valores dos atributos preditivos
 - y é a saída, um valor categórico em $\{c_1, c_2, \dots, c_m\}$
- Objetivo:
 - Encontrar a classe que maximize a probabilidade a posteriori

$$y_{MAP} = \underset{i}{\operatorname{argmax}} \underbrace{P(y_i|\mathbf{x})}$$


$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

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

Teorema de Bayes

Aula passada

Naïve Bayes

- Ideia "naïve": assumir que os valores dos atributos de um exemplo \mathbf{x} são independentes entre si dada a classe y
- Adota a suposição de independência condicional:

$$P(\mathbf{x}|y) = P(x^1, \dots, x^d|y) = \underbrace{\prod_{j=1}^d P(x^j|x^1, \dots, x^{j-1}, y)}_{\text{Regra da Cadeia (exata)}} = \underbrace{\prod_{j=1}^d P(x^j|y)}_{\text{Independência condicional}}$$

- Substituindo na fórmula do Teorema de Bayes:

$$P(y_i|\mathbf{x}) = \frac{P(\mathbf{x}|y_i)P(y_i)}{P(\mathbf{x})} \longrightarrow P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

Aula passada

Naïve Bayes

- **Todas as probabilidades utilizadas são estimadas a partir dos dados de treinamento** $\mathbf{D} = \{(\mathbf{x}_k, f(\mathbf{x}_k)), k = 1, \dots, n\}$, com cada exemplo dado por um vetor de atributos $\mathbf{x}_k = \{x_k^1, x_k^2, \dots, x_k^d\}$ formado pelo valor observado para os d atributos A_1, \dots, A_d

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

- $P(y_i)$: probabilidade da classe y_i no conjunto de treinamento
 - Calculada como a razão entre quantas instâncias pertencem à classe y_i e o número total de exemplos no conjunto D
- $P(x^j|y_i)$: probabilidade do valor de atributo A_j assumir um valor x^j dada a distribuição de classe y_i

Naïve Bayes com atributos categóricos

- Um restaurante deseja descobrir sob que condições climáticas se vende mais Feijoada ou Filé à Parmegiana.
- Objetivo: preparar melhor a *mise en place* da cozinha



- A equipe de vendas coletou os seguintes dados (*próximo slide*)

Aula passada: Naïve Bayes com atributos categóricos

	Previsão	Temperatura	Umidade	Vento	Prato
x_1	chuva	frio	normal	sim	parmegiana
x_2	chuva	moderado	alta	sim	parmegiana
x_3	sol	quente	alta	não	parmegiana
x_4	sol	quente	alta	sim	parmegiana
x_5	sol	moderado	alta	não	parmegiana
x_6	nublado	frio	normal	sim	feijoada
x_7	nublado	quente	alta	não	feijoada
x_8	nublado	quente	normal	não	feijoada
x_9	nublado	moderado	alta	sim	feijoada
x_{10}	chuva	frio	normal	não	feijoada
x_{11}	chuva	moderado	alta	não	feijoada
x_{12}	chuva	moderado	normal	não	feijoada
x_{13}	sol	frio	normal	não	feijoada
x_{14}	sol	moderado	normal	sim	feijoada

Instância de teste x_t

x_t	sol	frio	normal	sim	feijoada
-------	-----	------	--------	-----	-----------------

$$P(\text{parmegiana}) = 5/14 = 0.36$$

$$P(\text{feijoada}) = 9/14 = 0.64$$

$$\text{prod}_p = 0.0144$$

$$\text{prod}_f = 0.0158$$

8. Calcular a probabilidade a posteriori para cada classe

$$\begin{aligned} P(\text{feijoada} | x_t) &= \text{prod}_f * P(\text{feijoada}) \\ &= 0.0158 * 0.64 \\ &= \mathbf{0.011} \end{aligned}$$

$$\begin{aligned} P(\text{parmegiana} | x_t) &\neq \text{prod}_p * P(\text{parmegiana}) \\ &= 0.0144 * 0.36 \\ &= 0.005 \end{aligned}$$

A classe de maior probabilidade a posteriori é feijoada, assim $y_t = \text{feijoada}$

Aplicação de Naïve Bayes para classificação de dados estruturados

Estimativa das probabilidades com atributos categóricos,
numéricos e mistos.

Algoritmo Naïve Bayes

com atributos numéricos

- Quando os atributos são contínuos (numéricos) e o número de valores possíveis não é enumerável, há duas possibilidades principais para calcular a probabilidade condicional

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

- Discretizar os valores do atributo em uma fase de pré-processamento
- Assumir uma distribuição particular para os valores dos atributos

Algoritmo Naïve Bayes

com atributos numéricos

- Quando os atributos são contínuos (numéricos) e o número de valores possíveis não é enumerável, há duas possibilidades principais para calcular a probabilidade condicional

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

- Discretizar os valores do atributo em uma fase de pré-processamento
- Assumir uma distribuição particular para os valores dos atributos

Existem diversos métodos de discretização de atributos (quantitativos → qualitativos). Após, o algoritmo naïve Bayes para atributos categóricos pode ser aplicado.

- Baseado em algoritmo de agrupamento (não-supervisionado)
- Divisão de valores em intervalos de larguras iguais (afetado por *outliers*)
- Inspeção visual

Trabalhos anteriores propõem que o número de intervalos seja fixado em $p = \min(10, \text{número de valores diferentes para o atributo})$

Algoritmo Naïve Bayes

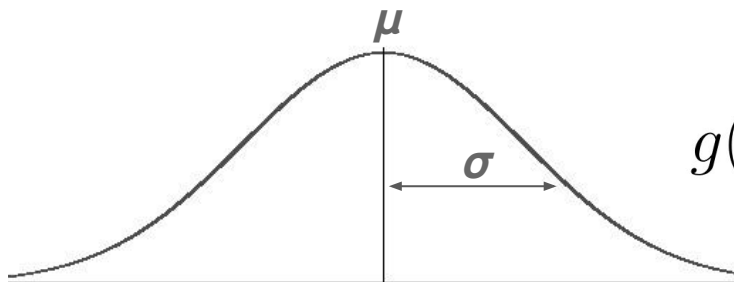
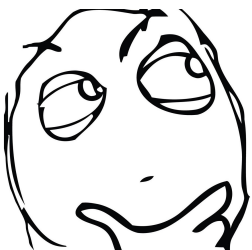
com atributos numéricos

- Quando os atributos são contínuos (numéricos) e o número de valores possíveis não é enumerável, há duas possibilidades principais para calcular a probabilidade condicional

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

- Discretizar os valores do atributo em uma fase de pré-processamento
- Assumir uma distribuição particular para os valores dos atributos

Tipicamente assume-se que os valores contínuos seguem uma distribuição Gaussiana com média μ e desvio padrão σ

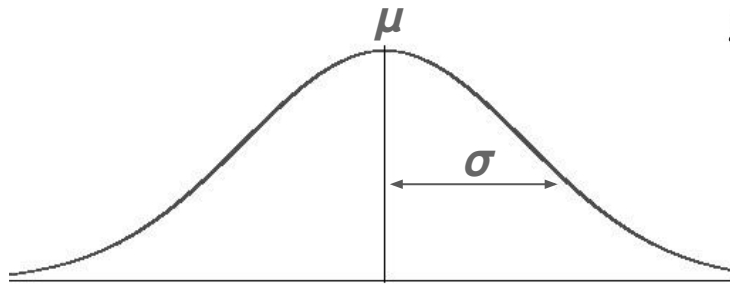
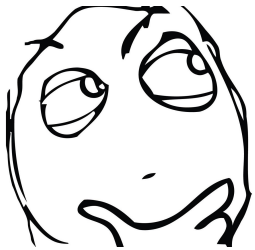


$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Algoritmo Naïve Bayes

com atributos numéricos

- Assumimos que os valores de um atributo numérico contínuo seguem uma distribuição Gaussiana



$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(x^j|y_i) = g(x^j, \mu_{y_i}, \sigma_{y_i})$$

- Estimamos a média e variância a partir dos dados de treinamento, avaliando cada atributo A_j para todos os exemplos pertencente à classe y_i

$$P(x^j|y_i) = g(x^j, \mu_{y_i}, \sigma_{y_i})$$



$$\mu_{y_i} = \frac{1}{|D_{y_i}|} \sum_{k:y_k=y_i} x_k^j$$

$$\sigma_{y_i}^2 = \frac{1}{|D_{y_i}|} \sum_{k:y_k=y_i} (x_k^j - \mu_{y_i})^2$$

Algoritmo Naïve Bayes

com atributos numéricos

1. Calcular a média e variância ($= \sigma^2$) dos valores do atributo A^j para todos os exemplos de treinamento pertencentes à classe y_i

$$P(x^j|y_i) = g(x^j, \mu_{y_i}, \sigma_{y_i})$$

$$\mu_{y_i} = \frac{1}{|D_{y_i}|} \sum_{k:y_k=y_i} x_k^j$$

$$\sigma_{y_i}^2 = \frac{1}{|D_{y_i}|} \sum_{k:y_k=y_i} (x_k^j - \mu_{y_i})^2$$

2. Ao calcular a probabilidade condicional, substituímos o valor de x^j , e da média μ_i e desvio padrão σ_i estimadas para a classe y_i na fórmula da Gaussiana

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i) \longleftarrow g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

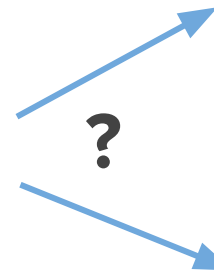
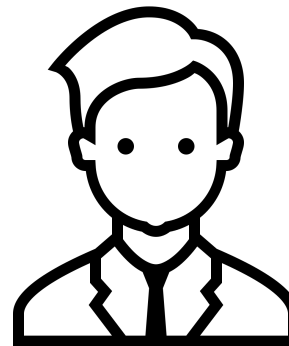
Algoritmo Naïve Bayes

com atributos numéricos

- Relembrando um exemplo anterior...



Atributos




Satisfeito


Insatisfeito

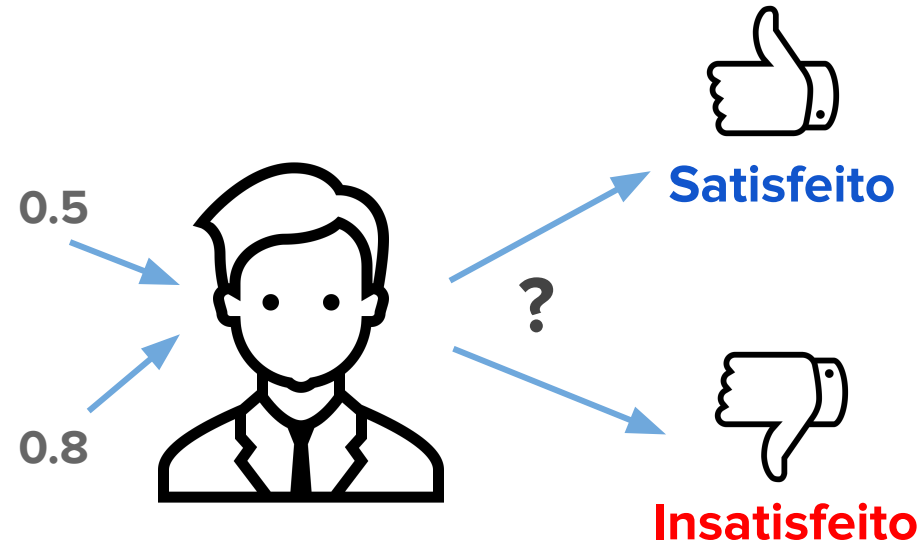
Tempo de espera
Preço
Sabor da comida
Qualidade do serviço
....

- Como realizar a classificação neste cenário com Naïve Bayes?

Algoritmo Naïve Bayes

com atributos numéricos

id	tempoEspera	preco	classe
12	1.00	0.20	INS
7	0.20	0.60	INS
11	0.50	0.10	INS
8	0.40	0.30	INS
9	0.61	0.20	INS
10	0.60	0.43	INS
6	0.60	1.00	SAT
4	0.90	0.70	SAT
5	0.80	0.30	SAT
2	0.40	0.50	SAT
1	0.70	0.40	SAT
3	0.61	0.40	SAT



Qual a reação mais provável do cliente para um restaurante cujo tempo de espera é 0.5 e o preço é 0.8 (valores normalizados) ?

Algoritmo Naïve Bayes

com atributos numéricos

id	tempoEspera	preco	classe
12	1.00	0.20	INS
7	0.20	0.60	INS
11	0.50	0.10	INS
8	0.40	0.30	INS
9	0.61	0.20	INS
10	0.60	0.43	INS
6	0.60	1.00	SAT
4	0.90	0.70	SAT
5	0.80	0.30	SAT
2	0.40	0.50	SAT
1	0.70	0.40	SAT
3	0.61	0.40	SAT

x_t	0.50	0.80	?
-------	-------------	-------------	----------

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

Média e desvio padrão dos atributos,

para cada classe:

$$\mu_{\text{tempo, INS}} = 0.550 \quad \mu_{\text{tempo, SAT}} = 0.670$$

$$\sigma_{\text{tempo, INS}} = 0.266 \quad \sigma_{\text{tempo, SAT}} = 0.174$$

TempoEspera

$$\mu_{\text{preco, INS}} = 0.305 \quad \mu_{\text{preco, SAT}} = 0.550$$

$$\sigma_{\text{preco, INS}} = 0.182 \quad \sigma_{\text{preco, SAT}} = 0.258$$

Preço

$$\mu_{y_i} = \frac{1}{|D_{y_i}|} \sum_{k: y_k = y_i} x_k^j \quad \sigma_{y_i}^2 = \frac{1}{|D_{y_i}|} \sum_{k: y_k = y_i} (x_k^j - \mu_{y_i})^2$$

Algoritmo Naïve Bayes

com atributos numéricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

Média e desvio padrão para cada atributo

\mathbf{x}_t	0.50	0.80	?
----------------	------	------	---

- $P(y_{INS}|\mathbf{x}_t) = ?$
- $P(y_{SAT}|\mathbf{x}_t) = ?$

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i) \rightarrow P(x^j|y_i) = g(x^j, \mu_{y_i}, \sigma_{y_i})$$

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu_{\text{tempo, INS}} = 0.550 \quad \mu_{\text{tempo, SAT}} = 0.670$$

$$\sigma_{\text{tempo, INS}} = 0.266 \quad \sigma_{\text{tempo, SAT}} = 0.174$$

TempoEspera

$$\mu_{\text{preco, INS}} = 0.310 \quad \mu_{\text{preco, SAT}} = 0.550$$

$$\sigma_{\text{preco, INS}} = 0.182 \quad \sigma_{\text{preco, SAT}} = 0.258$$

Preço

Algoritmo Naïve Bayes

com atributos numéricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

Média e desvio padrão para cada atributo

x_t	0.50	0.80	?
-------	------	------	---

$$\mu_{\text{tempo, INS}} = 0.550 \quad \mu_{\text{tempo, SAT}} = 0.670$$

$$\sigma_{\text{tempo, INS}} = 0.266 \quad \sigma_{\text{tempo, SAT}} = 0.174$$

TempoEspera

$$\mu_{\text{preco, INS}} = 0.310 \quad \mu_{\text{preco, SAT}} = 0.550$$

$$\sigma_{\text{preco, INS}} = 0.182 \quad \sigma_{\text{preco, SAT}} = 0.258$$

Preço

- $P(y_{INS} | \mathbf{x}_t) = ?$
- $P(y_{SAT} | \mathbf{x}_t) = ?$

$$P(y_{INS} | \mathbf{x}_t) = P(y_{INS}) P(x_t^{\text{tempo}} | y_{INS}) P(x_t^{\text{preco}} | y_{INS})$$

$$P(y_{INS} | \mathbf{x}_t) = P(y_{INS}) g(x_t^{\text{tempo}} | \mu_{\text{tempo, INS}}, \sigma_{\text{tempo, INS}}) g(x_t^{\text{preco}} | \mu_{\text{preco, INS}}, \sigma_{\text{preco, INS}})$$

$$P(y_{INS} | \mathbf{x}_t) = 0.5 * g(0.5, 0.550, 0.266) * g(0.8, 0.310, 0.182)$$

$$P(y_{INS} | \mathbf{x}_t) = 0.5 * 1.467 * 0.055$$

$$P(y_{INS} | \mathbf{x}_t) = 0.040$$

Algoritmo Naïve Bayes

com atributos numéricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

Média e desvio padrão para cada atributo

x_t	0.50	0.80	?
-------	------	------	---

- $P(y_{INS} | \mathbf{x}_t) = 0.040$
- $P(y_{SAT} | \mathbf{x}_t) = 0.693$

$$\mu_{\text{tempo}, INS} = 0.550 \quad \mu_{\text{tempo}, SAT} = 0.670$$

$$\sigma_{\text{tempo}, INS} = 0.266 \quad \sigma_{\text{tempo}, SAT} = 0.174$$

TempoEspera

$$\mu_{\text{preco}, INS} = 0.310 \quad \mu_{\text{preco}, SAT} = 0.550$$

$$\sigma_{\text{preco}, INS} = 0.182 \quad \sigma_{\text{preco}, SAT} = 0.258$$

Preço

A classe SATISFEITO é a mais provável para esta entrada

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

- Supondo um conjunto de dados D com d atributos, sendo d_c do tipo categóricos e d_n do tipo numérico (contínuo), tal que $d_c + d_n = d$
- A probabilidade a posteriori pode ser estimada como:

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^{d_c} P(x^j|y_i) \prod_{j=1}^{d_n} P(x^j|y_i)$$

- Suposição da independência condicional adotada pelo algoritmo facilita a análise para tipos heterogêneos de atributos:
 - calcula-se separadamente as probabilidades condicionais para cada conjunto de atributos dada uma classe y_i , agrupados pelo seu tipo (numérico ou categórico), e multiplica-se os seus resultados, juntamente com a probabilidade a priori de y_i

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

tempoEspera	preco	servico	classe
1.00	0.20	bom	INS
0.20	0.60	regular	INS
0.50	0.10	regular	INS
0.40	0.30	ruim	INS
0.61	0.20	ruim	INS
0.60	0.43	regular	INS
0.60	1.00	bom	SAT
0.90	0.70	bom	SAT
0.80	0.30	regular	SAT
0.40	0.50	ruim	SAT
0.70	0.40	regular	SAT
0.61	0.40	bom	SAT

x_t	0.50	0.80	regular	?
-------------------------	-------------	-------------	----------------	----------

Qual a classe predita pelo algoritmo Naïve Bayes para x_t ?

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

$$\mu_{\text{tempo}, INS} = 0.550 \quad \mu_{\text{tempo}, SAT} = 0.670$$

$$\sigma_{\text{tempo}, INS} = 0.266 \quad \sigma_{\text{tempo}, SAT} = 0.174$$

$$\mu_{\text{preco}, INS} = 0.310 \quad \mu_{\text{preco}, SAT} = 0.550$$

$$\sigma_{\text{preco}, INS} = 0.182 \quad \sigma_{\text{preco}, SAT} = 0.258$$

- $P(y_{INS} | \mathbf{x}_t) = ?$

- $P(y_{SAT} | \mathbf{x}_t) = ?$

$$P(y_{INS} | \mathbf{x}_t) = P(y_{INS}) P(x_t^{\text{tempo}} | y_{INS}) P(x_t^{\text{preco}} | y_{INS}) P(x_t^{\text{serv}} | y_{INS})$$

tempoEspera	preco	servico	classe
1.00	0.20	bom	INS
0.20	0.60	regular	INS
0.50	0.10	regular	INS
0.40	0.30	ruim	INS
0.61	0.20	ruim	INS
0.60	0.43	regular	INS
0.60	1.00	bom	SAT
0.90	0.70	bom	SAT
0.80	0.30	regular	SAT
0.40	0.50	ruim	SAT
0.70	0.40	regular	SAT
0.61	0.40	bom	SAT

\mathbf{x}_t	0.50	0.80	regular	?
----------------	------	------	---------	---

TempoEspera

Preço

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

$$\mu_{\text{tempo}, INS} = 0.550 \quad \mu_{\text{tempo}, SAT} = 0.670$$

$$\sigma_{\text{tempo}, INS} = 0.266 \quad \sigma_{\text{tempo}, SAT} = 0.174$$

$$\mu_{\text{preco}, INS} = 0.310 \quad \mu_{\text{preco}, SAT} = 0.550$$

$$\sigma_{\text{preco}, INS} = 0.182 \quad \sigma_{\text{preco}, SAT} = 0.258$$

- $P(y_{INS} | \mathbf{x}_t) = ?$

- $P(y_{SAT} | \mathbf{x}_t) = ?$

$$P(y_{INS} | \mathbf{x}_t) = P(y_{INS}) \underbrace{P(x_t^{\text{tempo}} | y_{INS})}_{\text{OK! (exemplo anterior)}} \underbrace{P(x_t^{\text{preco}} | y_{INS})}_{?} P(x_t^{\text{serv}} | y_{INS})$$

OK! (exemplo anterior)

?

$$P(\text{regular} | INS) = ?$$

$$P(\text{regular} | SAT) = ?$$

TempoEspera	tempoEspera	preco	servico	classe
	1.00	0.20	bom	INS
	0.20	0.60	regular	INS
	0.50	0.10	regular	INS
	0.40	0.30	ruim	INS
	0.61	0.20	ruim	INS
	0.60	0.43	regular	INS
	0.60	1.00	bom	SAT
	0.90	0.70	bom	SAT
	0.80	0.30	regular	SAT
	0.40	0.50	ruim	SAT
	0.70	0.40	regular	SAT
	0.61	0.40	bom	SAT
x_t	0.50	0.80	regular	?

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

$$\mu_{\text{tempo}, INS} = 0.550 \quad \mu_{\text{tempo}, SAT} = 0.670$$

$$\sigma_{\text{tempo}, INS} = 0.266 \quad \sigma_{\text{tempo}, SAT} = 0.174$$

$$\mu_{\text{preco}, INS} = 0.310 \quad \mu_{\text{preco}, SAT} = 0.550$$

$$\sigma_{\text{preco}, INS} = 0.182 \quad \sigma_{\text{preco}, SAT} = 0.258$$

- 5 exemplos possuem $x_t^{\text{serv}} = \text{regular}$
- $P(\text{regular} \mid INS) = 3/6 = 0.50$
- $P(\text{regular} \mid SAT) = 2/6 = 0.33$

TempoEspera

Preço

tempoEspera	preco	servico	classe
1.00	0.20	bom	INS
0.20	0.60	regular	INS
0.50	0.10	regular	INS
0.40	0.30	ruim	INS
0.61	0.20	ruim	INS
0.60	0.43	regular	INS
0.60	1.00	bom	SAT
0.90	0.70	bom	SAT
0.80	0.30	regular	SAT
0.40	0.50	ruim	SAT
0.70	0.40	regular	SAT
0.61	0.40	bom	SAT

x_t	0.50	0.80	regular	?
-------	------	------	---------	---

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

$$\mu_{\text{tempo}, INS} = 0.550 \quad \mu_{\text{tempo}, SAT} = 0.670$$

$$\sigma_{\text{tempo}, INS} = 0.266 \quad \sigma_{\text{tempo}, SAT} = 0.174$$

$$\mu_{\text{preco}, INS} = 0.310 \quad \mu_{\text{preco}, SAT} = 0.550$$

$$\sigma_{\text{preco}, INS} = 0.182 \quad \sigma_{\text{preco}, SAT} = 0.258$$

- 5 exemplos possuem $x_t^{\text{serv}} = \text{regular}$
- $P(\text{regular} | INS) = 3/6 = 0.50$
- $P(\text{regular} | SAT) = 2/6 = 0.33$

TempoEspera

Preço

tempoEspera	preco	servico	classe
1.00	0.20	bom	INS
0.20	0.60	regular	INS
0.50	0.10	regular	INS
0.40	0.30	ruim	INS
0.61	0.20	ruim	INS
0.60	0.43	regular	INS
0.60	1.00	bom	SAT
0.90	0.70	bom	SAT
0.80	0.30	regular	SAT
0.40	0.50	ruim	SAT
0.70	0.40	regular	SAT
0.61	0.40	bom	SAT

x_t	0.50	0.80	regular	?
-------	------	------	---------	---

$$P(y_{INS} | \mathbf{x}_t) = P(y_{INS}) P(x_t^{\text{tempo}} | y_{INS}) P(x_t^{\text{preco}} | y_{INS}) P(x_t^{\text{serv}} | y_{INS})$$

$$P(y_{INS} | \mathbf{x}_t) = 0.5 * 1.467 * 0.055 * \mathbf{0.50}$$

$$P(y_{INS} | \mathbf{x}_t) = 0.020$$

Algoritmo Naïve Bayes

com atributos numéricos e categóricos

Probabilidade a priori de cada classe

$$P(INS) = 6/12 = 0.5$$

$$P(SAT) = 6/12 = 0.5$$

$$\mu_{\text{tempo, INS}} = 0.550 \quad \mu_{\text{tempo, SAT}} = 0.670$$

$$\sigma_{\text{tempo, INS}} = 0.266 \quad \sigma_{\text{tempo, SAT}} = 0.174$$

$$\mu_{\text{preco, INS}} = 0.310 \quad \mu_{\text{preco, SAT}} = 0.550$$

$$\sigma_{\text{preco, INS}} = 0.182 \quad \sigma_{\text{preco, SAT}} = 0.258$$

- 5 exemplos possuem $x_t^{\text{serv}} = \text{regular}$
- $P(\text{regular} | INS) = 3/6 = 0.50$
- $P(\text{regular} | SAT) = 2/6 = 0.33$

TempoEspera

Preço

tempoEspera	preco	servico	classe
1.00	0.20	bom	INS
0.20	0.60	regular	INS
0.50	0.10	regular	INS
0.40	0.30	ruim	INS
0.61	0.20	ruim	INS
0.60	0.43	regular	INS
0.60	1.00	bom	SAT
0.90	0.70	bom	SAT
0.80	0.30	regular	SAT
0.40	0.50	ruim	SAT
0.70	0.40	regular	SAT
0.61	0.40	bom	SAT

x_t	0.50	0.80	regular	?
-------	------	------	---------	---

$$P(y_{INS} | x_t) = 0.020$$

$$P(y_{SAT} | x_t) = 0.229$$

Repetindo para a classe SAT, temos:

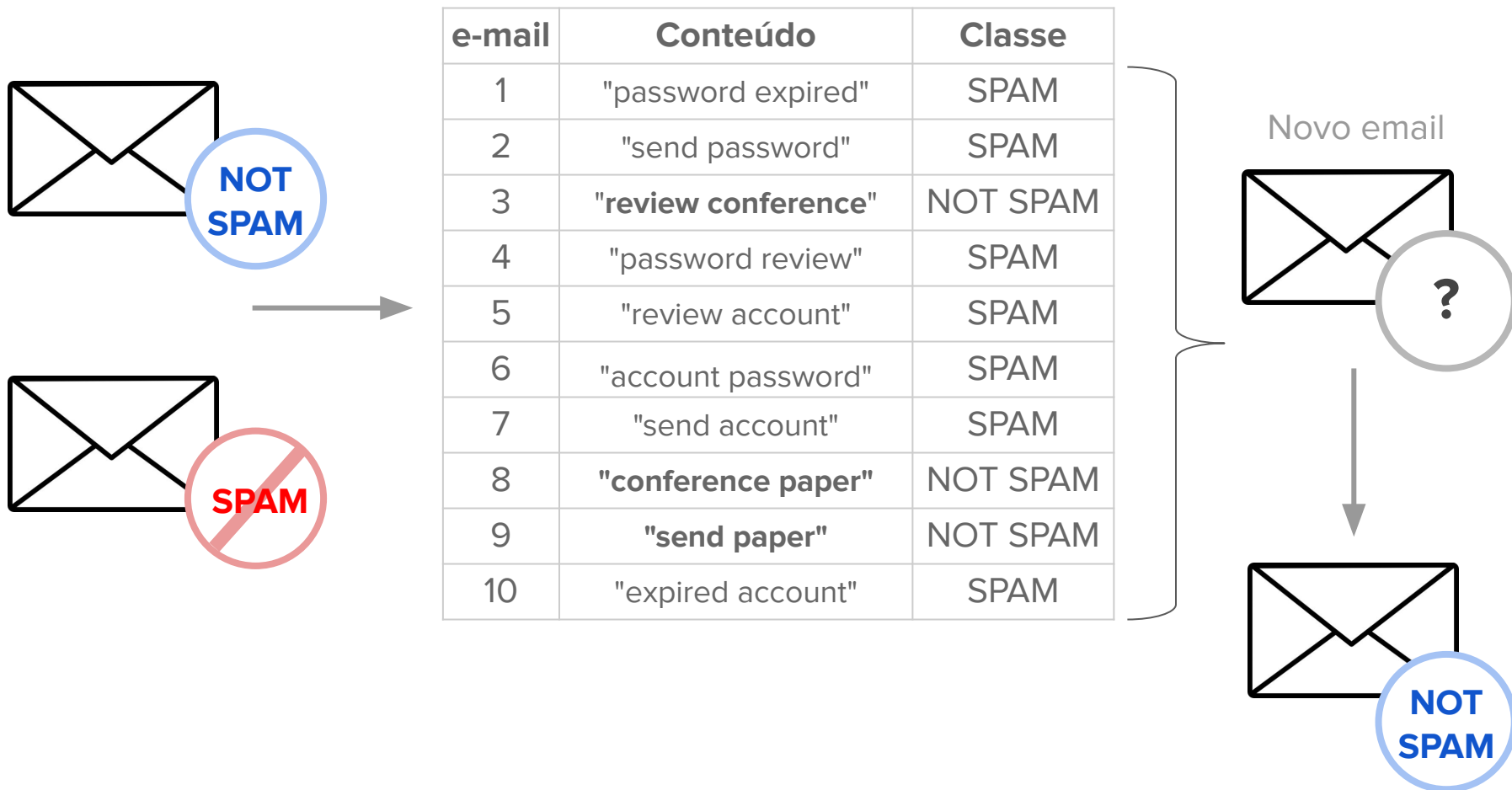
$P(y_{SAT} | x_t) > P(y_{INS} | x_t)$, logo
classe predita é "Satisfeito"

Aplicação de Naïve Bayes para classificação de texto

Estimativa das probabilidades com
Modelo Bernoulli e Modelo Multinomial

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"



Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

e-mail	Conteúdo	Classe
1	"password expired"	SPAM
2	"send password"	SPAM
3	"review conference"	NOT SPAM
4	"password review"	SPAM
5	"review account"	SPAM
6	"account password"	SPAM
7	"send account"	SPAM
8	"conference paper"	NOT SPAM
9	"send paper"	NOT SPAM
10	"expired account"	SPAM
...

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

7 palavras **únicas**: { password, expired, send, review, conference, account, paper }

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} &P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} &P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * \mathbf{120/700} \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * 120/700 * \mathbf{95/700} \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * 120/700 * 95/700 * \mathbf{250/700} \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * 120/700 * 95/700 * 250/700 \\ &= \mathbf{0.0058} \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

$$\begin{aligned} P(\text{SPAM}|\text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review}|\text{SPAM}) \\ &\quad * P(\text{conference}|\text{SPAM}) * P(\text{paper}|\text{SPAM}) \\ &= 700/1000 * 120/700 * 95/700 * 250/700 \\ &= \mathbf{0.0058} \end{aligned}$$

$$\begin{aligned} P(\text{NOT SPAM}|\text{"review conference paper"}) \\ &= P(\text{NOT SPAM}) * P(\text{review}|\text{NOT SPAM}) \\ &\quad * P(\text{conference}|\text{NOT SPAM}) * P(\text{paper}|\text{NOT SPAM}) \\ &= 300/1000 * 87/300 * 150/300 * 125/300 \\ &= \mathbf{0.0181} \end{aligned}$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34
paper	250	125
...

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34

$$\begin{aligned} P(\text{SPAM} | \text{"review conference paper"}) \\ &= P(\text{SPAM}) * P(\text{review} | \text{SPAM}) \\ &\quad * P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM}) \\ &= 700/1000 * 120/700 * 95/700 * 250/700 \\ &= \mathbf{0.0058} \end{aligned}$$

A classe predita será "NOT SPAM", visto que ela maximiza $P(y_i|\mathbf{x})$

$$\begin{aligned} P(\text{NOT SPAM} | \text{"review conference paper"}) \\ &= P(\text{NOT SPAM}) * P(\text{review} | \text{NOT SPAM}) \\ &\quad * P(\text{conference} | \text{NOT SPAM}) * P(\text{paper} | \text{NOT SPAM}) \\ &= 300/1000 * 87/300 * 150/300 * 125/300 \\ &= \mathbf{0.0181} \end{aligned}$$

Contagem do número de e-mails em cada classe com ocorrência da palavra

Retomando um exemplo simples...

"Supondo um novo e-mail recebido, com conteúdo *"review conference paper"*, é mais provável que ele seja SPAM ou NOT SPAM?"

$$P(y_i|\mathbf{x}) = P(y_i) \prod_{j=1}^d P(x^j|y_i)$$

700 e-mails SPAM
300 e-mails NOT SPAM

palavra	# SPAM	#NOT SPAM
password	190	50
expired	210	30
send	290	180
review	120	87
conference	95	150
account	80	34

$P(\text{SPAM} | \text{"review conference paper"})$
= $P(\text{SPAM}) * P(\text{review} | \text{SPAM})$
 * $P(\text{conference} | \text{SPAM}) * P(\text{paper} | \text{SPAM})$
= $700/1000 * 120/700 * 95/700 * 250/700$
= **0.0058**

A classe predita será "NOT SPAM", visto que ela maximiza $P(y_i|\mathbf{x})$

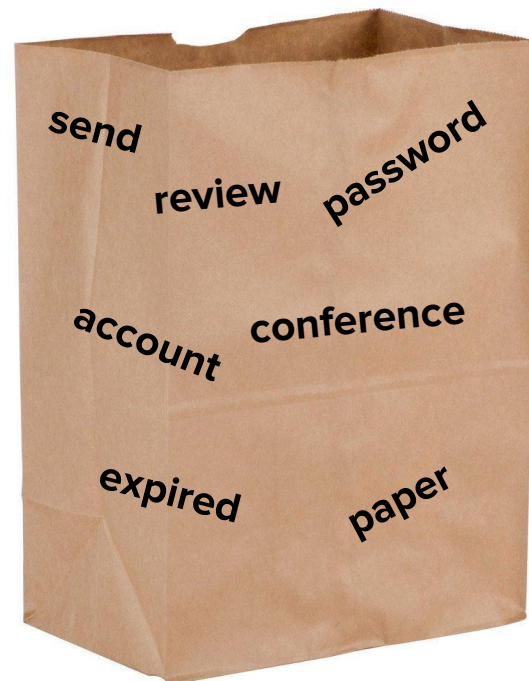
O racional do algoritmo para classificação de texto é o mesmo:
como estimar as probabilidades a partir de um conjunto de documentos?

= **0.0181**

Classificação de texto

- Documentos de texto são usualmente representados através de um modelo simples de **bag of words** (BoW):
 - armazena que palavras estão contidas no documento (e por vezes as respectivas frequências), ignorando informações como a posição e ordem entre palavras

e-mail	Conteúdo
1	"password expired"
2	"send password"
3	"review conference"
4	"password review"
5	"review account"
6	"account password"
7	"send account"
8	"conference paper"
9	"send paper"
10	"expired account"



Classificação de texto

- Formalmente... este modelo é representado por um **vetor de atributos**, cujos componentes são as palavras no *BoW*
 - o comprimento do vetor corresponde ao tamanho do vocabulário
- Considere o vocabulário:
 $\mathbf{V} = \{\text{password, expired, send, review, conference, account, paper}\}$
 - $|\mathbf{V}| = 7$ palavras (*tamanho do vocabulário*)
- Considere um novo documento com conteúdo "*password expired, review account password*"

Como representar este novo documento através de um vetor de atributos?

Representação de documentos de texto

- Supondo o vocabulário $\mathbf{V} = \{\text{password, expired, send, review, conference, account, paper}\}$ e um novo documento com conteúdo "***password expired, review account password***"
- 2 modelos possíveis para representação:
 - **Modelo Bernoulli:** um documento é representado por um vetor de atributos com valores binários, onde o valor 1 indica que a palavra está presente no documento e 0 que não está presente

$$\mathbf{b} = (1,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	1	1	0	1	0	1	0

- **Modelo Multinomial:** um documento é representado por um vetor de atributos com valores inteiros, indicando a frequência de cada palavra no documento

$$\mathbf{m} = (2,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	2	1	0	1	0	1	0

Classificação de texto com Naïve Bayes:

Modelo Bernoulli

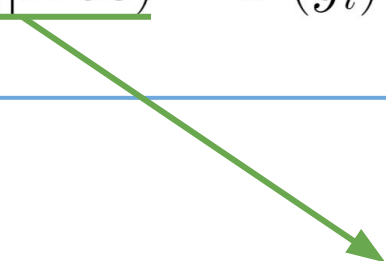
- Cada documento é representado por um vetor de atributos binário:

$$\mathbf{b} = (1,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	1	1	0	1	0	1	0

- Para cada classe y_i , considerando um novo documento Doc , calculamos:

$$\underline{P(y_i|Doc)} = P(y_i) \prod_{j=1}^{|V|} [b_j P(w^j|y_i) + (1 - b_j)(1 - P(w^j|y_i))]$$


$$P(y_i|Doc) = P(y_i|\mathbf{b})$$

Classificação de texto com Naïve Bayes:

Modelo Bernoulli

- Cada documento é representado por um vetor de atributos binário:

$$\mathbf{b} = (1,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	1	1	0	1	0	1	0

- Para cada classe y_i , considerando um novo documento Doc , representado pelo seu vetor \mathbf{b} , calculamos:

$$P(y_i|\mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j|y_i) + (1-b^j)(1-P(w^j|y_i))]$$

Onde:

- $|V|$ é o tamanho do vocabulário (comprimento do vetor \mathbf{b})
- b^j é o valor no vetor \mathbf{b} correspondente à palavra j (0 ou 1)
- $P(w^j|y_i)$ é a probabilidade da palavra w^j **ocorrer** em documentos da classe y_i
 - $1 - P(w^j|y_i)$: probabilidade da palavra w^j **não ocorrer** em documentos da classe y_i

Classificação de texto com Naïve Bayes:

Modelo Bernoulli

- Cada documento é representado por um vetor de atributos binário:

$$\mathbf{b} = (1,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	1	1	0	1	0	1	0

- Para cada classe y_i , considerando um novo documento Doc , representado pelo seu vetor \mathbf{b} , calculamos:

$$P(y_i|\mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j|y_i) + (1-b^j)(1-P(w^j|y_i))]$$

Se a palavra w^j está presente,
 $b^j = 1$ e a probabilidade
considerada é $P(w^j|y_i)$

Se a palavra w^j não está presente, $b^j = 0$ e a probabilidade
considerada é $1 - P(w^j|y_i)$

Classificação de texto com Naïve Bayes: Modelo Bernoulli

- Cada documento é representado por um vetor de atributos binário,

Como estimar as probabilidades a priori e condicional?

Sempre a partir dos dados de treinamento!

$$P(y_i|\mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j|y_i) + (1-b^j)(1-P(w^j|y_i))]$$

Se a palavra w^j está presente,
 $b^j = 1$ e a probabilidade
considerada é $P(w^j|y_i)$

Se a palavra w^j não está presente, $b^j = 0$ e a probabilidade
considerada é $1 - P(w^j|y_i)$

Classificação de texto com Naïve Bayes:

Modelo Bernoulli

- Para estimar as probabilidades envolvidas, defina a partir dos dados de treinamento:
 - **V**: o vocabulário, cujo número de palavras determina o tamanho de vetor de atributos
 - **N**: o número total de documentos
 - **N_i**: o número total de documentos pertencente à classe y_i , para y_i em $\{c_1, c_2, \dots, c_m\}$
 - **n_i(w^j)**: o número de documentos da classe y_i nos quais a palavra w^j está presente, para $j = 1 \dots |V|$ e y_i em $\{c_1, c_2, \dots, c_m\}$
- E calcule:

$$P(w^j | y_i) = \frac{n_i(w^j)}{N_i}$$

Probabilidade Condicional

$$P(y_i) = \frac{N_i}{N}$$

Probabilidade a Priori

Classificação com Modelo Bernoulli: Exemplo

- Considere o problema de classificar um determinado documento como pertencente à categoria "Esportes" ou "Informática" utilizando o Modelo Bernoulli.
- A análise dos dados de treinamento resultou no seguinte vocabulário:
 - $|V| = 8$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

Classificação com Modelo Bernoulli: Exemplo

- Os vetores de atributos para cada classe, são:

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Número de documentos pertencentes à classe

Número de palavras no vocabulário (comum a todas as classes)

Classificação com Modelo Bernoulli: Exemplo

- Os vetores de atributos para cada classe, são:

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \quad \left. \vphantom{\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}} \right\} \text{6 documentos}$$

$$\mathbf{B}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \quad \left. \vphantom{\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}} \right\} \text{5 documentos}$$

Classificação com Modelo Bernoulli: Exemplo

- Os vetores de atributos para cada classe, são:

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Análise do *Doc3*
pertencente à classe
"Esportes"

$$\mathbf{B}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Análise do *Doc2*
pertencente à classe
"Informática"

Análise da ocorrência da palavra w_3 nos documentos

Classificação com Modelo Bernoulli: Exemplo

- Estimando as probabilidades:

$$P(y_i|\mathbf{b}) = \underline{P(y_i)} \prod_{j=1}^{|V|} [b^j P(w^j|y_i) + (1-b^j)(1-P(w^j|y_i))]$$

$$P(y_i) = \frac{N_i}{N}$$

Número de documentos da classe y_i
Número total de documentos

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{B}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Bernoulli: Exemplo

- Estimando as probabilidades:

$$P(y_i|\mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [\underline{b^j P(w^j|y_i)} + (1 - b^j)(1 - \underline{P(w^j|y_i)})]$$

$$P(w^j|y_i) = \frac{n_i(w^j)}{N_i}$$

Número de documentos da classe y_i que contêm a palavra

Número de documentos da classe y_i

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	3/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
...

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{B}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Bernoulli: Exemplo

- Estimando as probabilidades:

$$P(y_i|\mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [\underline{b^j P(w^j|y_i)} + (1 - b^j)(1 - \underline{P(w^j|y_i)})]$$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	3/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
w_5	3	3/6	1	1/5
w_6	4	4/6	1	1/5
w_7	4	4/6	3	3/5
w_8	4	4/6	1	1/5

$$\mathbf{B}_{\text{ESP}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{B}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Bernoulli: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Bernoulli) para um novo documento com vetor de atributos $\mathbf{b} = (0, 1, 1, 0, 1, 0, 1, 0)$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | \mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j | y_i) + (1 - b^j)(1 - P(w^j | y_i))]$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	3/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
w_5	3	3/6	1	1/5
w_6	4	4/6	1	1/5
w_7	4	4/6	3	3/5
w_8	4	4/6	1	1/5

$$P(\text{ESP} | \mathbf{b}) = 6/11 * ((1-0) * 3/6 * 1 * 1/6 * 1 * 2/6 * (1-0) * 3/6 * 1 * 3/6 * (1-0) * 2/6 * 1 * 4/6 * (1-0) * 2/6)$$

$$P(\text{ESP} | \mathbf{b}) \approx 2.8 \times 10^{-4}$$

Classificação com Modelo Bernoulli: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Bernoulli) para um novo documento com vetor de atributos $\mathbf{b} = (0, 1, 1, 0, 1, 0, 1, 0)$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | \mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j | y_i) + (1 - b^j)(1 - P(w^j | y_i))]$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	3/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
w_5	3	3/6	1	1/5
w_6	4	4/6	1	1/5
w_7	4	4/6	3	3/5
w_8	4	4/6	1	1/5

$$P(\text{INF} | \mathbf{b}) = 5/11 * ((1-0) * 4/5 * 1 * 3/5 * 1 * 3/5 * (1-0) * 4/5 * 1 * 1/5 * (1-0) * 4/5 * 1 * 3/5 * (1-0) * 4/5)$$

$$P(\text{INF} | \mathbf{b}) \approx 8.0 \times 10^{-3}$$

Classificação com Modelo Bernoulli: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Bernoulli) para um novo documento com vetor de atributos $\mathbf{b} = (0, 1, 1, 0, 1, 0, 1, 0)$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | \mathbf{b}) = P(y_i) \prod_{j=1}^{|V|} [b^j P(w^j | y_i) + (1 - b^j)(1 - P(w^j | y_i))]$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	3	3/6	1	1/5
w_2	1	1/6	3	3/5
w_3	2	2/6	3	3/5
w_4	3	3/6	1	1/5
w_5	3	3/6	1	1/5

$$P(\text{ESP} | \mathbf{b}) \approx 2.8 \times 10^{-4}$$

$$P(\text{INF} | \mathbf{b}) \approx 8.0 \times 10^{-3}$$

$p(\text{INF} | \mathbf{b}) > p(\text{ESP} | \mathbf{b})$: a classe predita será "Informática" - classe mais provável dado vetor \mathbf{b}

Classificação de texto com Naïve Bayes:

Modelo Multinomial

- Cada documento é representado por um vetor de atributos inteiro, indicando a frequência das palavras no documento:

$$\mathbf{m} = (2,1,0,1,0,1,0)$$

	password	expired	send	review	conference	account	paper
doc1	2	1	0	1	0	1	0

- Para cada classe y_i , considerando um novo documento Doc , calculamos:

$$P(y_i|Doc) = P(y_i) \prod_{j=1}^{|V|} P(w^j|y_i)$$

Onde:

- $|V|$ é o tamanho do vocabulário (comprimento do vetor \mathbf{m})
- $P(w^j|y_i)$ é a frequência relativa da palavra w^j em documentos da classe y_i
 - valor do vetor \mathbf{m} na posição j

Classificação de texto com Naïve Bayes:

Modelo Multinomial

- Para estimar as probabilidades envolvidas, defina a partir dos dados de treinamento:
 - **V**: o vocabulário, cujo número de palavras determina o tamanho de vetor de atributos
 - **N**: o número total de documentos
 - **N_i**: o número total de documentos pertencente à classe y_i , para y_i em $\{c_1, c_2, \dots, c_m\}$
 - **n_i(w^j)**: frequência da palavra w^j dentre todos os documentos da classe y_i para $j = 1 \dots |V|$ e y_i em $\{c_1, c_2, \dots, c_m\}$
- E calcule:

$$P(w^j | y_i) = \frac{n_i(w^j)}{\sum_{s=1}^{|V|} n_i(w^s)}$$

Probabilidade Condicional

$$P(y_i) = \frac{N_i}{N}$$

Probabilidade a Priori

Classificação de texto com Naïve Bayes:

Modelo Multinomial

- Diferente do modelo Bernoulli, o modelo Multinomial assume que palavras que não ocorrem no documento a ser classificado (i.e., $n_i(w^j) = 0$) não são relevantes para o cálculo da probabilidade a posteriori
- Assim, podemos simplificar:

$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{|V|} P(w^j | y_i)$$



$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

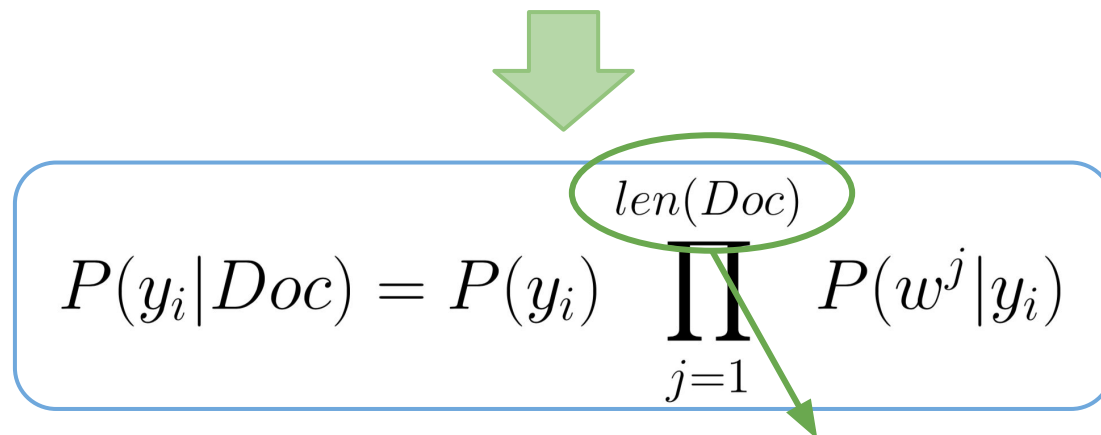
Comprimento do documento *Doc*, em função do número de palavras (únicas) contidas no texto

Classificação de texto com Naïve Bayes:

Modelo Multinomial

- Diferente do modelo Bernoulli, o modelo Multinomial assume que palavras que não ocorrem no documento a ser classificado (i.e., $n_i(w^j) = 0$) não são relevantes para o cálculo da probabilidade a posteriori
- Assim, podemos simplificar:

Na prática, apenas palavras que ocorrem no documento a ser classificado precisam ser avaliadas


$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

Comprimento do documento Doc , em função do número de palavras contidas no texto

Classificação com Modelo Multinomial: Exemplo

- Considere o mesmo problema anterior, de classificar um determinado documento como pertencente à categoria "Esportes" ou "Informática", agora utilizando o Modelo Multinomial
- Vocabulário no conjunto de treinamento ($|V| = 8$):

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

Classificação com Modelo Multinomial: Exemplo

- Os vetores de atributos para cada classe são:

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$\mathbf{M}_{\text{ESP}} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix} \quad \left. \vphantom{\begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix}} \right\} \text{6 documentos}$$

$$\mathbf{M}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \quad \left. \vphantom{\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}} \right\} \text{5 documentos}$$

Classificação com Modelo Multinomial: Exemplo

- Estimando as probabilidades:

$$P(y_i | Doc) = \underline{P(y_i)} \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(y_i) = \frac{N_i}{N}$$

Número de documentos da classe y_i
Número total de palavras em cada classe

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$\mathbf{M}_{\text{ESP}} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix}$$

$$\mathbf{M}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Multinomial: Exemplo

- Estimando as probabilidades:

$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} \underline{P(w^j | y_i)}$$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(w^j | y_i) = \frac{n_i(w^j)}{\sum_{s=1}^{|V|} n_i(w^s)}$$

Frequência da palavra na classe y_i
 Soma da frequência de todas as palavras em y_i

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
...

$$\mathbf{M}_{\text{ESP}} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix}$$

$$\mathbf{M}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Multinomial: Exemplo

- Estimando as probabilidades:

$$P(y_i|Doc) = P(y_i) \prod_{j=1}^{len(Doc)} \underline{P(w^j|y_i)}$$

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
w_5	4	4/36	1	1/16
w_6	6	6/36	2	2/16
w_7	7	7/36	3	3/16
w_8	6	6/36	1	1/16

$$\mathbf{M}_{\text{ESP}} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{pmatrix}$$

$$\mathbf{M}_{\text{INF}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Classificação com Modelo Multinomial: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Multinomial) para um novo documento contendo as palavras $Doc = \{w^3, w^5, w^2, w^7\}$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
w_5	4	4/36	1	1/16
w_6	6	6/36	2	2/16
w_7	7	7/36	3	3/16
w_8	6	6/36	1	1/16

$$\begin{aligned} P(\text{ESP} | Doc) &= 6/11 * P(w^3|\text{ESP}) * P(w^5|\text{ESP}) * \\ &\quad P(w^2|\text{ESP}) * P(w^7|\text{ESP}) \\ &= 6/11 * 2/36 * 4/36 * 1/36 * 7/36 \end{aligned}$$

$$P(\text{ESP} | Doc) \approx 1.82 \times 10^{-5}$$

Classificação com Modelo Multinomial: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Multinomial) para um novo documento contendo as palavras $Doc = \{w^3, w^5, w^2, w^7\}$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
w_5	4	4/36	1	1/16
w_6	6	6/36	2	2/16
w_7	7	7/36	3	3/16
w_8	6	6/36	1	1/16

$$\begin{aligned} P(\text{INF} | Doc) &= 5/11 * P(w^3|\text{ESP}) * P(w^5|\text{ESP}) * \\ &\quad P(w^2|\text{ESP}) * P(w^7|\text{ESP}) \\ &= 5/11 * 3/16 * 1/16 * 4/16 * 3/16 \end{aligned}$$

$$P(\text{INF} | Doc) \approx 2.5 \times 10^{-4}$$

Classificação com Modelo Multinomial: Exemplo

- Qual a classe predita pelo Naïve Bayes (modelo Multinomial) para um novo documento contendo as palavras $Doc = \{w^3, w^5, w^2, w^7\}$?

$$V = \begin{bmatrix} w_1 = \text{goal}, \\ w_2 = \text{tutor}, \\ w_3 = \text{variance}, \\ w_4 = \text{speed}, \\ w_5 = \text{drink}, \\ w_6 = \text{defence}, \\ w_7 = \text{performance}, \\ w_8 = \text{field} \end{bmatrix}$$

$$P(\text{ESP}) = 6/11$$

$$P(\text{INF}) = 5/11$$

$$P(y_i | Doc) = P(y_i) \prod_{j=1}^{len(Doc)} P(w^j | y_i)$$

	$n_{\text{ESP}}(w)$	$P(w \text{ESP})$	$n_{\text{INF}}(w)$	$P(w \text{INF})$
w_1	5	5/36	1	1/16
w_2	1	1/36	4	4/16
w_3	2	2/36	3	3/16
w_4	5	5/36	1	1/16
w_5	4	4/36	1	1/16

$$P(\text{ESP} | Doc) \approx 1.82 \times 10^{-5}$$

$$P(\text{INF} | Doc) \approx 2.5 \times 10^{-4}$$

$p(\text{INF} | Doc) > p(\text{ESP} | Doc)$: a classe predita será "Informática" - classe mais provável dado Doc

Correção de Laplace

Na classificação de texto

- Assim como na classificação de dados estruturados, o problema da frequência zero deve ser tratado
- Palavras que não ocorrem em documentos de uma determinada classe y_i causarão a probabilidade a posteriori desta classe a se igualar a zero!
- Solução: assumir uma ocorrência extra de cada palavra no vocabulário durante o cálculo das probabilidades condicionais:

$$P(w^j|y_i) = \frac{n_i(w^j) + 1}{N_i + |V|}$$

Modelo Bernoulli

$$P(w^j|y_i) = \frac{n_i(w^j) + 1}{\sum_{s=1}^{|V|} n_i(w^s) + |V|}$$

Modelo Multinomial

Sugestão de exercício:



- Considerando os vetores de atributos e probabilidades estimadas dos dois exemplos anteriores, como seriam classificadas as instâncias...
 - $\mathbf{b} = (1, 0, 0, 0, 1, 1, 1, 1)$, pelo modelo Bernoulli ?
 - $Doc_1 = \{w^1, w^2, w^5, w^6, w^8\}$, pelo modelo Multinomial?

Considerações finais...

Considerações finais

- A fórmula do Naïve Bayes também pode ser expressa como uma soma, tomando-se o logaritmo de todos os termos:

$$\log(P(y_i|\mathbf{x})) \propto \log(P(y_i)) + \sum_{j=1}^d \log(P(x^j|y_i))$$

- O uso de logaritmos e da operação soma reduz as chances de underflow no cálculo das probabilidades
- Não interfere no algoritmo: classe cujo valor associado é máximo é retornada como saída

Vantagens e desvantagens do algoritmo

- Algoritmo de fácil implementação e simples funcionamento: estima todas as probabilidades a partir dos dados de treinamento, modelando cada classe
- A suposição de independência condicional não interfere negativamente em seu desempenho: o naïve Bayes possui boa capacidade preditiva mesmo para grandes bases de palavras, e em domínios em que há clara dependência entre os atributos
- É robusto à presença de ruídos e atributos irrelevantes
- No entanto...
 - não lida bem com frases ou termos compostos: “Chicago bulls” tem significado próprio, e distinto de “Chicago” e “bulls”
 - Ignora a ordem em que as palavras aparecem.

Classificação de texto:

Modelo Bernoulli vs Multinomial

- O modelo Multinomial leva em consideração a frequência das palavras, ainda que não considere a interdependência entre elas
- O modelo Bernoulli tende a funcionar melhor para documentos curtos, enquanto o modelo Multinomial é mais recomendado para documentos longos: apenas estima probabilidade condicional das palavras presentes no documento analisado
- Pré-processamento de texto com métodos de linguagem de natural é desejável, por exemplo:
 - Lematização (*stemming*): “desflexionar” as palavras
 - Fishing, fisher, fished → fish
 - Remoção de "stop words" (palavras muito comuns, como "the") ...
- Modelo multinomial é mais robusto a "stop words":
 - Palavra "the" está presente em praticamente todo documento, Modelo Bernoulli estima $P(\text{"The"}|y_i) = 1$
 - Modelo Multinomial estima a frequência relativa, ex: $P(\text{"The"}|y_i) = 0.05$

Até agora...

- Kit de ferramentas de aprendizado supervisionado:
 - KNN
 - Árvores de Decisão
 - Naïve Bayes
- Próxima aula:
 - Avaliação de modelos

