Classificadores Bayesianos

Mineração de Dados

Universidade Federal do ABC

Introdução

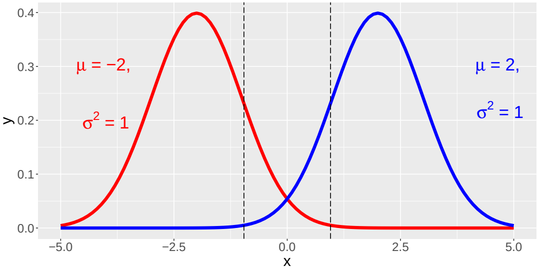
Discriminantes

Naïve Bayes

Introdução

Introdução

- Podemos pensar o problema que discutimos do ponto de vista probabilístico
  - Temos duas classes que correspondem a funções de densidade de probabilidade distintas ($p(x|c_1)$ e $p(x|c_2)$)



Introdução

- Lembrando as aulas de probabilidade
  - [Lei da Probabilidade Total]

$$p(x) = \sum_{c \in \mathcal{C}} p(x, c) = \sum_{c \in \mathcal{C}} p(x|c)p(c)$$

  - [Teorema de Bayes]

$$p(c|x) = \frac{p(c, x)}{p(x)} = \frac{p(x|c)p(c)}{p(x)} = \frac{p(x|c)p(c)}{\sum_{c \in \mathcal{C}} p(x|c)p(c)}$$

    - $posterior = likelihood\ x\ prior\ x\ evidence^{-1}$

Introdução

- Em um problema de classificação queremos encontrar

$$\arg\max_{c \in \mathcal{C}} p(c|x)$$

$$\arg\max_{c \in \mathcal{C}} \frac{p(x|c)p(c)}{\sum_{c \in \mathcal{C}} p(x|c)p(c)}$$

Introdução

- Queremos encontrar uma regra de decisão que minimiza o erro

$$p(erro|x) = \begin{cases} p(c_1|x) & \text{se escolhermos } c_2 \\ p(c_2|x) & \text{se escolhermos } c_1 \end{cases}$$

- Logo chegamos a regra de decisão de Bayes
    - Classe c_1 se $p(c_1|x) > p(c_2|x)$, e c_2 caso contrário
    - Classe c_1 se $p(x|c_1)p(c_1) > p(x|c_2)p(c_2)$, e c_2 caso contrário
- Sob essa regra, temos $p(erro|x) = \min(p(c_1|x), p(c_2|x))$
- Classificador *ótimo* de Bayes
    - Como $p(x|c_1)$ é definido?

Introdução

- Função discriminante
  - $g_i(x) > g_j(x) \forall j \neq i \rightarrow c_i$
- Formas equivalentes do ponto de vista de classificação
  - $g_i(x) = p(c_i|x) = \frac{p(x|c_i)p(c_i)}{\sum_{c_j \in C} p(x|c_j)p(c_j)}$
  - $g_i(x) = p(x|c_i)p(c_i)$
  - $g_i(x) = \log(p(x|c_i)) + \log(p(c_i))$
- Caso de duas classes (dicotomizador)
  - Uma única função discriminante
  - $g(x) = g_1(x) - g_2(x)$
  - $g(x) = \log\left(\frac{p(x|c_1)}{p(x|c_2)}\right) + \log\left(\frac{p(c_1)}{p(c_2)}\right)$

Introdução

Discriminantes

Naïve Bayes

Funções discriminantes - Caso Gaussiana Multivariada

- Abordagem paramétrica
- Vamos assumir que temos duas classes
  - $p(\mathbf{x}|c_1) = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$
  - $p(\mathbf{x}|c_2) = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$
  - $p(c_i)$ estimada pela proporção de objetos da classe i
  - $\boldsymbol{\mu}_i$ é a média (D-dimensões) da classe i
  - Σ_i é a matriz (D, D) de covariância da classe i

$$\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right]$$

$$g_i(\mathbf{x}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p(c_i))$$

Caso Gaussiana Multivariada - Classes Hiperesféricas

- $\sigma^2 = \frac{(n_1-1)\sigma_1^2+\cdots+(n_C-1)\sigma_c^2}{N-C}$ [σ_c^2 é a variância estimada nos dados da classe c, *desvio padrão combinado*]
- $|\Sigma_i| = \sigma^{2d}$ [Determinante de diag$(a_1, ..., a_n)$=$a_1 \cdot ... \cdot a_n$]
- $\Sigma^{-1} = \frac{1}{\sigma^2}\mathbf{I}$

Caso Gaussiana Multivariada - Classes Hiperesféricas

- Desconsiderando os temos iguais para todas as classes

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x} - \boldsymbol{\mu}_i)^T(\mathbf{x} - \boldsymbol{\mu}_i)}{2\sigma^2} + \ln(p(c_i))$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i) + \ln(p(c_i))$$

[$\mathbf{x}^T\mathbf{x}$ é uma constante aditiva]

$$g_i(\mathbf{x}) = \left(\frac{\boldsymbol{\mu}_i}{\sigma^2}\right)^T \mathbf{x} + \left(-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln(p(c_i))\right)$$

  - temos uma função discriminante **linear**

Caso Gaussiana Multivariada - Classes Hiperesféricas

$$g_i(\mathbf{x}) = \left(\frac{\boldsymbol{\mu}_i}{\sigma^2}\right)^T \mathbf{x} + \left(-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln(p(c_i))\right)$$

- Desconsiderando os temos iguais para todas as classes
  - se $p(c_i)$ igual para todas as classes temos o **classificador de distância mínima**
    - novo objeto classificado para a classe com média mais próxima

Caso Gaussiana Multivariada - Classes Hiperesféricas

- Exemplo
  - $\boldsymbol{\mu}_1^T = [1\ 2]\ \ \boldsymbol{\mu}_2^T = [4\ 6]\ \ \boldsymbol{\mu}_3^T = [-2\ 4]$
  - $p(c_1) = p(c_2) = \frac{1}{4} \quad p(c_3) = \frac{1}{2}$
  - $\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

$$g_i(\mathbf{x}) = \left(\frac{\boldsymbol{\mu}_i}{\sigma^2}\right)^T \mathbf{x} + \left(-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln(p(c_i))\right)$$

Caso Gaussiana Multivariada - Classes Hiperesféricas

$$g_1(\mathbf{x}) = \left(\frac{[1\ 2]}{3}\right)[x_1\ x_2]^T + \left(-\frac{5}{6} - 1,38)\right)$$

$$g_2(\mathbf{x}) = \left(\frac{[4\ 6]}{3}\right)[x_1\ x_2]^T + \left(-\frac{52}{6} - 1,38)\right)$$

$$g_3(\mathbf{x}) = \left(\frac{[-2\ 4]}{3}\right)[x_1\ x_2]^T + \left(-\frac{20}{6} - 0,69)\right)$$

Caso Gaussiana Multivariada - Classes Hiperesféricas

- Exemplo
  - Agora basta verificar casos em que $g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall j \neq i$
  - Podemos verificar as **fronteiras de decisão**
    - Regiões definidas por $g_i(\mathbf{x}) = g_j(\mathbf{x})$

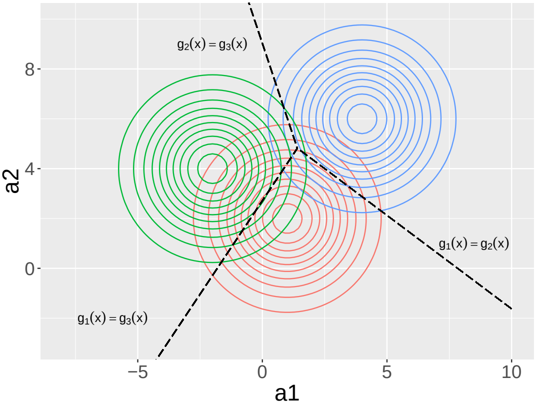
$$g_1(\mathbf{x}) = g_2(\mathbf{x}) \rightarrow -x_1 - \frac{4}{3}x_2 = -\frac{47}{6}$$

$$g_2(\mathbf{x}) = g_3(\mathbf{x}) \rightarrow 2x_1 + \frac{2}{3}x_2 = 6.02$$

$$g_1(\mathbf{x}) = g_3(\mathbf{x}) \rightarrow x_1 - \frac{2}{3}x_2 = -1.81$$

Caso Gaussiana Multivariada - Classes Hiperesféricas

- Exemplo



Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

- Caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)
  - $\Sigma = \frac{(n_1-1)\Sigma_1+\cdots+(n_C-1)\Sigma_c}{N-C}$ [Σ_c é a matriz de covariância estimada nos dados da classe c, *matriz de covariância combinada*]
  - $|\Sigma_i|$ igual para todas as classes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p(c_i))$$

  - Se $p(c_i)$ igual para todas as classes ainda temos um classificador de distância mínima
    - de acordo com a distância de **Mahalanobis**

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

- Caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p(c_i))$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x} + \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln(p(c_i))$$

[$\mathbf{x}^T \Sigma^{-1} \mathbf{x}$ constante aditiva]

$$g_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln(p(c_i))$$

- Novamente temos um caso de discriminante linear
  - método conhecido como *análise de discriminante linear* (LDA)

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

- Exemplo caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)
  - $\boldsymbol{\mu}_1^T = [1\ 2]$ $\boldsymbol{\mu}_2^T = [-1\ 5]$ $\boldsymbol{\mu}_3^T = [-2\ 4]$
  - $p(c_1) = p(c_2) = \frac{1}{4}$ $\quad p(c_3) = \frac{1}{2}$
  - $\Sigma = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 4 \end{bmatrix}$

$$g_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \mu_{\mathbf{i}}{}^T \Sigma^{-1} \mu_{\mathbf{i}} + \ln(p(c_i))$$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

- Exemplo caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)
  - Vale a pena manipular a equação para encontrar as fronteiras de decisão

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

$$\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \ln(p(c_i)) = \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + \ln(p(c_j))$$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

$$\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \ln(p(c_i)) = \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + \ln(p(c_j))$$

$$\left(\boldsymbol{\mu}_i^T\Sigma^{-1} - \boldsymbol{\mu}_j^T\Sigma^{-1}\right)\mathbf{x} = \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i - \ln(p(c_i)) - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + \ln(p(c_j))$$

$$\left(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T\right)\Sigma^{-1}\mathbf{x} = \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + \ln\left(\frac{p(c_j)}{p(c_i)}\right)$$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

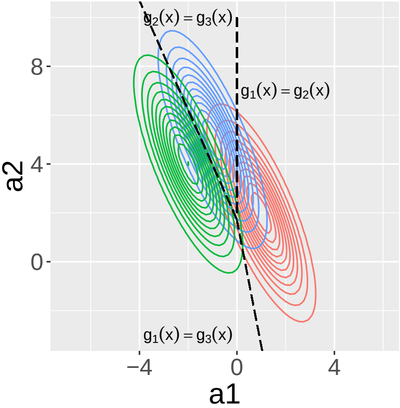
- Exemplo caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)

$$\left(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T\right)\Sigma^{-1}\mathbf{x} = \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + \ln\left(\frac{p(c_j)}{p(c_i)}\right)$$

- Caso $g_1(x) = g_2(x) \rightarrow x_1 = 0$
- Caso $g_2(x) = g_3(x) \rightarrow 3,14x_1 + 1,4x_2 = 2,41$
- Caso $g_1(x) = g_3(x) \rightarrow 5,14x_1 + 1,43x_2 = 2,41$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais I

- Exemplo caso em que $\Sigma_i = \Sigma$ (classes hiperelipsoidais, mesma forma)



Caso Gaussiana Multivariada - Classes Hiperelipsoidais II

- Caso em que Σ_i são arbitrárias (classes hiperelipsoidais)
  - Não é possível remover muitos termos
  - Resulta em fronteiras **quadráticas**
    - método chamado de *análise de discriminante quadrática* (QDA)

$$g_i(\mathbf{x}) = -\frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln(p(c_i))$$

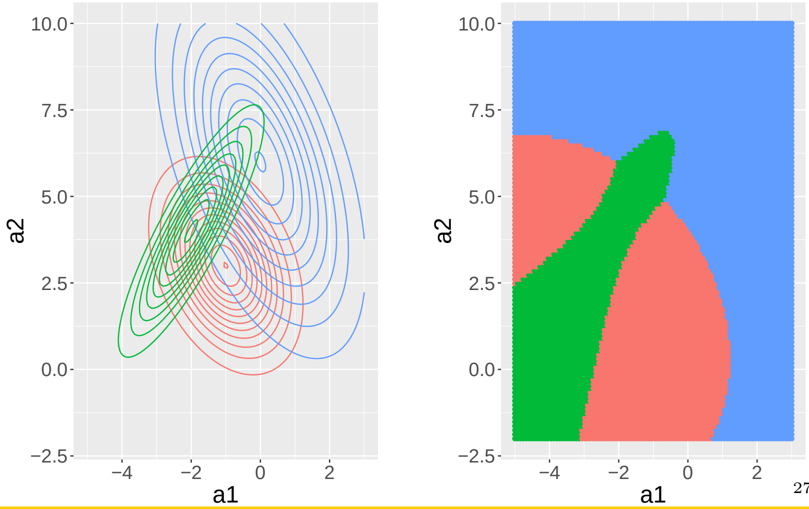
$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\Sigma_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln(|\Sigma_i|) + \ln(p(c_i))$$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais II

- Exemplo caso em que Σ_i são arbitrárias (classes hiperelipsoidais)
  - $\boldsymbol{\mu}_1^T = [-1\ 3]$ $\boldsymbol{\mu}_2^T = [0\ 6]$ $\boldsymbol{\mu}_3^T = [-2\ 4]$
  - $p(c_1) = p(c_2) = \frac{1}{4}$ $\quad p(c_3) = \frac{1}{2}$
  - $\Sigma_1 = \begin{bmatrix} 1 & -0,5 \\ -0,5 & 2 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 2 & -2 \\ -2 & 7 \end{bmatrix}$ $\Sigma_3 = \begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix}$

Caso Gaussiana Multivariada - Classes Hiperelipsoidais II

- Exemplo caso em que Σ_i são arbitrárias



Caso Gaussiana Multivariada - Classes Hiperelipsoidais II

- Temos um classificador ótimo quando as premissas são válidas
  - Normalmente a premissa de normalidade não é válida
- Apesar disso o método apresenta bons resultados
  - Mesmo no caso restrito em que se assume matrizes de covariância iguais
- Em alta dimensionalidade o custo de inverter a matriz de covariância é alto — $O(D^3)$
  - Como podemos melhorar?

Introdução

Discriminantes

Naïve Bayes

Naïve Bayes

- De volta ao báisco, Teorema de Bayes:
  - [Teorema de Bayes]

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})}$$

- Ao invés de assumirmos que $p(\mathbf{x}|c) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, podemos assumir que os atributos são **condicionalmente independentes dado a classe**
- O modelo obtido a partir dessa premissa é chamado de *Naïve Bayes*

Naïve Bayes

- Independência condicional
  - $p(a, b|c) = p(a|c)p(b|c)$
  - Note que não necessariamente a e b são independentes
    - Dois eventos são independentes se $p(a, b) = p(a)p(b)$
  - Exemplo:
    - Temos duas moedas e uma delas é enviesada (sempre cara)
    - Pegamos uma moeda aleatoriamente e jogamos para cima 2x
    - Seja A o evento do primeiro lançamento ser cara
    - Seja B o evento do segundo lançamento ser cara
    - Seja C o evento de termos escolhido a moeda enviesada
    - Se não soubermos nada além de que A ocorreu, a probabilidade de B aumenta
    - Se soubermos que C ocorreu, A ocorrer não influencia a probabilidade de B ocorrer

Naïve Bayes

- Independência condicional
  - No nosso contexto, os atributos são considerados independentes dado a classe
    - Premissa forte e normalmente inválida
    - Costuma funcionar bem

Naïve Bayes

$$p(c|\mathbf{x}) = \frac{p(c, x_1, x_2, \ldots, x_D)}{p(\mathbf{x})} = \frac{p(x_1|c, x_2, \ldots, x_D)p(c, x_2, \ldots, x_D)}{p(\mathbf{x})}$$

$$p(c|\mathbf{x}) = \frac{p(x_1|c, x_2, \ldots, x_D)p(x_2|c, x_3, \ldots, x_D)p(c, x_3, x_4, \ldots, x_D)}{p(\mathbf{x})}$$

$$p(c|\mathbf{x}) = \frac{p(x_1|c)p(x_2|c)\ldots p(x_D|c)p(c)}{p(\mathbf{x})} = \frac{p(c)\prod_{i=1}^{D} p(x_i|c)}{p(\mathbf{x})}$$

Naïve Bayes

- Vantagens
  - Fácil de incorporar atributos de diferentes tipos
    - Cada atributo pode ser modelado por uma distribuição específica (não necessariamente Gaussiana)
  - Estimação de parâmetros sempre feita considerando apenas 1 dimensão
  - Modelo pode ser aprendido de forma incremental
- Cuidados
  - Atributos redundantes

Naïve Bayes - Modelando os atributos

- Considere $\mathcal{X}^c$ como o conjunto de objetos da classe c
- Atributos booleanos:
  - Distribuição de Bernoulli:
    - $p(x_i|c) = \theta^{x_i}(1-\theta)^{1-x_i} \quad [x_i \in \{0,1\}]$
    - MLE de $\theta = \frac{1}{|\mathcal{X}^c|}\sum_{\mathbf{x}\in\mathcal{X}^c} x_i$ [e se todos os valores forem 1?]
- Atributos nominais:
  - Distribuição categórica:
    - $p(x_i|c) = \prod_{j=1}^{V} p_j^{x_i=j} \quad [x_i \in \{1,\ldots,V\}]$
    - MLE de $p_j = \frac{1}{|\mathcal{X}^c|}\sum_{\mathbf{x}\in\mathcal{X}^c} \mathbb{1}_{[x_i=j]}$ [e se todos os valores forem iguais?]

Naïve Bayes - Modelando os atributos

- Atributos contínuos
  - Mais comum Distribuição Normal
    - $p(x_i|c) = \mathcal{N}(\mu, \sigma^2)$
    - MLE de μ e σ^2 foram apresentados na aula de *Parzen Window*

Naïve Bayes - Exemplo

Chuta2Pes	Altura	Peso	HsTreino	Empresario	Sucesso
N	(1,5-1,6]	67	2	Fulano	N
S	(1,8-1,9]	80	8	Fulano	N
N	(1,8-1,9]	92	2	Ciclano	N
N	(1,7-1,8]	79	4	Ciclano	N
S	(1,7-1,8]	67	5	Beltrano	S
N	(1,5-1,6]	50	9	Ciclano	S
S	(1,6-1,7]	58	6	Ciclano	S
S	(1,7-1,8]	73	3	Fulano	S
N	(1,8-1,9]	90	10	Fulano	S
S	(1,5-1,6]	63	5	Ciclano	S
S	(1,7-1,8]	77	6	Beltrano	S
S	(1,7-1,8]	60	1	Beltrano	S

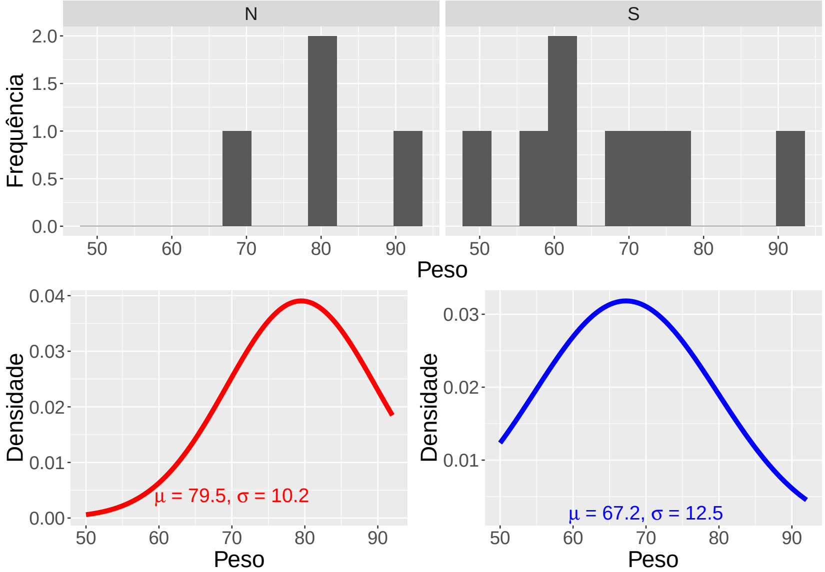
Naïve Bayes - Exemplo

Sucesso/Chuta2Pes	N	S
N	3	1
S	2	6

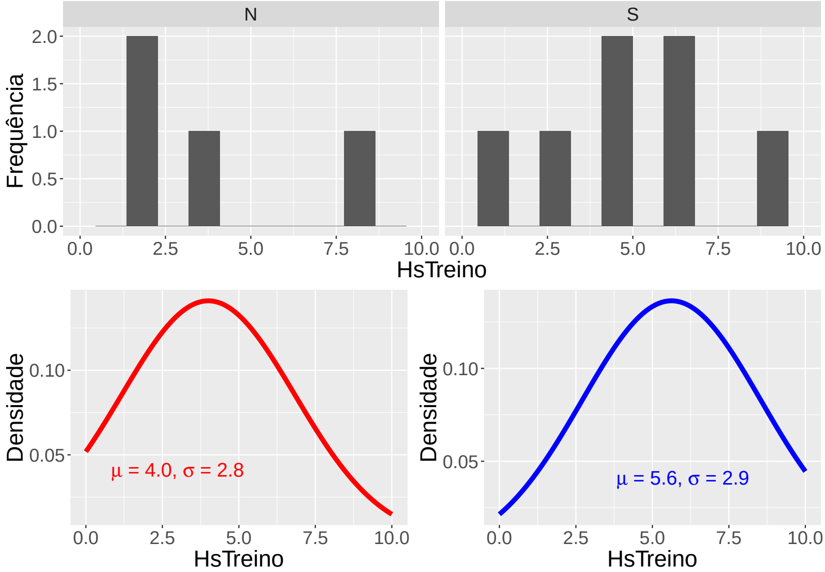
Sucesso/Empresario	Beltrano	Ciclano	Fulano
N	0	2	2
S	3	3	2

Sucesso/Altura	(1,5-1,6]	(1,6-1,7]	(1,7-1,8]	(1,8-1,9]
N	1	0	1	2
S	2	1	4	1

Naïve Bayes - Exemplo



Naïve Bayes - Exemplo



Naïve Bayes - Exemplo

- $p(Sucesso = S) = 8/12 \;\; p(Sucesso = N) = 4/12$
- Peso
  - $p(Peso = z|Sucesso = S) = \mathcal{N}(z|67.25, 157.07)$*
  - $p(Peso = z|Sucesso = N) = \mathcal{N}(z|79.50, 104.33)$*
- HsTreino
  - $p(HsTreino = z|Sucesso = S) = \mathcal{N}(z|5.62, 8.55)$*
  - $p(HsTreino = z|Sucesso = N) = \mathcal{N}(z|4.00, 8.00)$*
- Chuta2Pes
  - $p(Chuta2Pes = 1|Sucesso = S) = 6/8$
  - $p(Chuta2Pes = 1|Sucesso = N) = 1/4$
- Empresário
  - $p(Empresario = Beltrano|Sucesso = S) = 3/8$
  - $p(Empresario = Ciclano|Sucesso = S) = 3/8$
  - $p(Empresario = Beltrano|Sucesso = N) = 0/4$
  - $p(Empresario = Ciclano|Sucesso = N) = 2/4$

Naïve Bayes - Exemplo

- Altura
  - $p(Altura = (1,5 - 1,6]|Sucesso = S) = 2/8$
  - $p(Altura = (1,6 - 1,7]|Sucesso = S) = 1/8$
  - $p(Altura = (1,7 - 1,8]|Sucesso = S) = 4/8$
  - $p(Altura = (1,8 - 1,9]|Sucesso = S) = 1/8$
  - $p(Altura = (1,5 - 1,6]|Sucesso = N) = 1/4$
  - $p(Altura = (1,6 - 1,7]|Sucesso = N) = 0/4$
  - $p(Altura = (1,7 - 1,8]|Sucesso = N) = 1/4$
  - $p(Altura = (1,8 - 1,9]|Sucesso = N) = 2/4$

Naïve Bayes - Exemplo

- Modelo pronto
- Podemos fazer a classificação de novos dados
- Considere $\mathbf{x}_t$ o seguinte objeto de teste:

Altura	Peso	HsTreino	Chuta2Pes	Empresario
(1,7-1,8]	73	6	S	Beltrano

Naïve Bayes - Exemplo

- $p(Sucesso = S) = 0.67$
- $p(Sucesso = N) = 0.33$
- $p(Altura = (1,7 - 1,8]|Sucesso = S) = 0.50$
- $p(Altura = (1,7 - 1,8]|Sucesso = N) = 0.25$
- $p(Peso = 73|Sucesso = S) = 0.029$
- $p(Peso = 73|Sucesso = N) = 0.032$
- $p(HsTreino = 6|Sucesso = S) = 0.135$
- $p(HsTreino = 6|Sucesso = N) = 0.110$
- $p(Chuta2Pes = 1|Sucesso = S) = 0.75$
- $p(Chuta2Pes = 1|Sucesso = N) = 0.25$
- $p(Empresario = Beltrano|Sucesso = S) = 0.38$
- $p(Empresario = Beltrano|Sucesso = N) = 0.00$

Naïve Bayes - Exemplo

$$p(Sucesso = S|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 0.67 \cdot 0.50 \cdot 0.029 \cdot 0.135 \cdot 0.75 \cdot 0.38$$

$$p(Sucesso = N|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 0.33 \cdot 0.25 \cdot 0.032 \cdot 0.110 \cdot 0.25 \cdot 0.00$$

Naïve Bayes - Laplace Smoothing

- O zero observado em $p(Empresario = Beltrano|Sucesso = N)$ torna nula a probabilidade da classe ser "N"
  - Definitivamente um problema
  - Valor não visto no treinamento
- Para evitar esse problema em atributos discretos utilizamos *Laplace smoothing*
  - Adicionamos uma probabilidade pequena de um valor não visto ocorrer

Naïve Bayes - Laplace Smoothing

- Para evitar esse problema em atributos discretos utilizamos *Laplace smoothing*
  - Adicionamos uma probabilidade pequena de um valor não visto ocorrer
  - Neste caso, temos:

$$p_j = \frac{1 + \sum_{\mathbf{x} \in \mathcal{X}^c} \mathbb{1}_{[x_i = j]}}{|\mathcal{X}^c| + V}$$

V é o número de valores possíveis do atributo

Naïve Bayes - Exemplo

- Corrigindo nossa tabela do atributo Empresario temos:
  - $p(Empresario = Beltrano|Sucesso = S) = 4/11$
  - $p(Empresario = Ciclano|Sucesso = S) = 4/11$
  - $p(Empresario = Beltrano|Sucesso = N) = 1/7$
  - $p(Empresario = Ciclano|Sucesso = N) = 3/7$

Naïve Bayes - Exemplo

$$p(Sucesso = S|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 0.67 \cdot 0.50 \cdot 0.03 \cdot 0.14 \cdot 0.75 \cdot 0.36$$

$$p(Sucesso = S|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 3.5239281 \times 10^{-4}$$

$$p(Sucesso = N|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 0.33 \cdot 0.25 \cdot 0.03 \cdot 0.11 \cdot 0.25 \cdot 0.14$$

$$p(Sucesso = N|\mathbf{x}_t) = \frac{1}{p(\mathbf{x}_t)} 1.0428371 \times 10^{-5}$$

Naïve Bayes - Exemplo

- $p(Sucesso = S|\mathbf{x}_t) > p(Sucesso = N|\mathbf{x}_t)$
  - Já sabemos qual seria a classe predita
  - Não temos as probabilidades de cada classe! Não normalizamos por $p(\mathbf{x}_t)$

$$\begin{aligned} p(\mathbf{x}_t) &= p(\mathbf{x}_t|Sucesso = S)p(Sucesso = S) \\ &+ p(\mathbf{x}_t|Sucesso = N)p(Sucesso = N) \end{aligned}$$

Naïve Bayes - Exemplo

- Logo, já temos todos os valores para obter as probabilidades

$$p(Sucesso = S|\mathbf{x}_t) = \frac{0.000352}{(0.000010 + 0.000352)} = 0.9712575$$

$$p(Sucesso = N|\mathbf{x}_t) = \frac{0.000010}{(0.000010 + 0.000352)} = 0.0287425$$

Naïve Bayes (NB)

- Estimar parâmetros para atributos contínuos com poucos objetos pode ser um problema
  - Qual característica do NB perdemos se usarmos *Parzen Window* para estimar densidade?
- É comum considerar a versão discretizada de atributos contínuos
  - Pode-se utilizar uma das técnicas discutidas na aula anterior (discretização por largura fixa ou frequência fixa)
  - Na próxima aula falaremos de outro método de discretização que considera os rótulos de classe
- Sabe-se que as estimativas de probabilidade obtidas pelo NB não são muito boas
  - Independente disso, o classificador possui ótimos resultados em diversas aplicações

Naïve Bayes (NB)

- A fronteira de decisão obtida pelo NB depende das distribuições consideradas
  - No caso da distribuição Gaussiana a fronteira obtida é quadrática (relação com QDA)
- É possível incrementar o modelo considerando relação entre alguns atributos
  - Aumentamos o número de parâmetros
  - Podemos obter modelos mais realistas
- Classificadores desse tipo são conhecidos como Redes Bayesianas
  - Sendo o NB uma instância desse conjunto

Referências

R. Duda, P. Hart e D. Stork. Pattern Classification. **Seção 2.1, 2.4 e 2.6**

Slides Profa. Olga Veksler
http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture4.pdf

C. Bishop. Pattern Recognition and Machine Learning. **Seção 2.1, 2.2, 2.3**

T. Hastie, R. Tibshirani e J. Friedman. The Elements of Statistical Learning. **Seção 4.3**