

Naive Bayes & Regressão Logística

Henrique Laureano

<http://leg.ufpr.br/~henrique>

CiDWeek I, 03-07/02/2020



Naive Bayes

Primeiro, precisamos falar sobre o que é um **classificador de Bayes**.

Classificador de Bayes

Um classificador probabilístico baseado no **teorema de Bayes**.

Exemplo, _____

- » Meningite causa torcicolo 50% das vezes, $\mathbb{P}[T|M]$
- » Prob. *a priori* de um paciente estar com meningite é $1/50.000$, $\mathbb{P}[M]$
- » Probabilidade *a priori* de um paciente estar com torcicolo é $1/20$, $\mathbb{P}[T]$

Se um paciente está com torcicolo, qual a probabilidade dele estar com meningite?

$$\mathbb{P}[M|T] = \frac{\mathbb{P}[T|M] \mathbb{P}[M]}{\mathbb{P}[T]} = \frac{1/2 \times 1/50.000}{1/20} = 0.0002.$$



Classificadores Bayesianos

Considere **atributos** A_1, A_2, \dots, A_n e uma **classe** C com rótulos c_1, c_2, \dots, c_k .

O que queremos?

Predição : $C = c_1$ ou $C = c_2$ ou \dots ,

i.e., queremos o valor de C que maximiza $\mathbb{P}[C|A_1, A_2, \dots, A_n]$.

Como fazemos? Teorema de Bayes.

Calculamos a probabilidade *a posteriori* $\mathbb{P}[C|A_1, A_2, \dots, A_n]$ para todos os valores de C ,

$$\mathbb{P}[C_k|A_1, A_2, \dots, A_n] = \frac{\mathbb{P}[A_1, A_2, \dots, A_n|C_k] \mathbb{P}[C_k]}{\mathbb{P}[A_1, A_2, \dots, A_n]}.$$

E como calculamos $\mathbb{P}[A_1, A_2, \dots, A_n|C_k]$? **Naive Bayes.**



Classificador Naive Bayes

Por que *naive*?

Porque se assume **independência** entre os atributos A_i dado uma classe, i.e.,

$$\mathbb{P}[A_1, A_2, \dots, A_n | C_k] = \mathbb{P}[A_1 | C_k] \mathbb{P}[A_2 | C_k] \dots \mathbb{P}[A_n | C_k].$$

Vantagem: Grande redução do custo computacional.

Um novo ponto é classificado como C_k se $\mathbb{P}[C_k] \times \prod_{i=1}^n \mathbb{P}[A_i | C_k]$ é máximo.

i.e., _____

$$C_k = \operatorname{argmax}_{k \in \{1, \dots, K\}} \mathbb{P}[C_k] \times \prod_{i=1}^n \mathbb{P}[A_i | C_k]$$



Exemplo: Estimando probabilidades a partir dos dados

| <i>Tid</i> | Refund | Marital Status | Taxable Income | Evade |
|------------|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

$$\gg \mathbb{P}[C] = N_k/N$$

$$\gg \mathbb{P}[C = \text{No}] = 7/10$$

$$\gg \mathbb{P}[C = \text{Yes}] = 3/10$$

Atributos discretos, _____

$$\gg \mathbb{P}[A_i|C_k] = A_{ik}/N_k$$

$$\gg \mathbb{P}[\text{Status} = \text{Married}|\text{No}] = 4/7$$

$$\gg \mathbb{P}[\text{Refund} = \text{Yes}|\text{Yes}] = 0$$

» ...



E com atributos contínuos?

| Tid | Refund | Marital Status | Taxable Income | Evade |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Estimação da densidade de probabilidade

- » Se assume distribuição Normal
- » Se estima a média μ e o desvio padrão σ
- » Se estima a probabilidade condicional

$$\mathbb{P}[A_i|C_k] = \frac{\exp\left\{-\frac{(A_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right\}}{\sqrt{2\pi\sigma_{ik}^2}}$$

Exemplo, _____

$$\begin{aligned}\mathbb{P}[\text{Income} = 120|\text{No}] &= \frac{1}{\sqrt{2\pi 2975}} \exp\left\{-\frac{(120 - 110)^2}{2 \times 2975}\right\} \\ &= 0.0072.\end{aligned}$$



Classificador Naive Bayes: Exemplo

Dado o perfil: $X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{k})$

$$\begin{aligned}\mathbb{P}[X|\text{Class} = \text{No}] &= \mathbb{P}[\text{Refund} = \text{No}|\text{Class} = \text{No}] \times \\ &\quad \mathbb{P}[\text{Married}|\text{Class} = \text{No}] \times \\ &\quad \mathbb{P}[\text{Income} = 120\text{k}|\text{Class} = \text{No}] \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024.\end{aligned}$$

$$\begin{aligned}\mathbb{P}[X|\text{Class} = \text{Yes}] &= \mathbb{P}[\text{Refund} = \text{No}|\text{Class} = \text{Yes}] \times \\ &\quad \mathbb{P}[\text{Married}|\text{Class} = \text{Yes}] \times \\ &\quad \mathbb{P}[\text{Income} = 120\text{k}|\text{Class} = \text{Yes}] \\ &= 1 \times 0 \times 10^{-9} = 0.\end{aligned}$$

Já que $\mathbb{P}[X|\text{No}] \mathbb{P}[\text{No}] > \mathbb{P}[X|\text{Yes}] \mathbb{P}[\text{Yes}]$,

$\Rightarrow \mathbb{P}[X|\text{No}] > \mathbb{P}[X|\text{Yes}] \Rightarrow \text{Class} = \text{No}.$



Problema de probabilidade zero

Se uma das probabilidades condicionais é zero, então toda a expressão

$$\mathbb{P}[A_1, A_2, \dots, A_n | C_k] = \prod_{i=1}^n \mathbb{P}[A_i | C_k], \text{ se torna zero.}$$

Como evitamos isso?

Abordagens:

$$\text{Original : } \mathbb{P}[A_i | C_k] = \frac{N_{ik}}{N_k}, \quad \text{Laplace : } \mathbb{P}[A_i | C_k] = \frac{N_{ik} + 1}{N_k + k},$$

$$\text{Estimativa-M : } \mathbb{P}[A_i | C_k] = \frac{N_{ik} + mp}{N_k + m},$$

em que p é uma probabilidade *a priori* e m é um parâmetro.

Qual a abordagem é mais utilizada?

Correção de **Laplace** (ou estimador de Laplace).



Classificador Naive Bayes: Comentários

Vantagens, _____

- » Fácil de implementar
- » Apresenta bons resultados na maioria dos cenários
- » Robusto com *outliers* e atributos irrelevantes
- » Ignora dados faltantes durante o cálculo das probabilidades

Desvantagens, _____

- » Suposição de **independência**
- » Perda de acurácia

Como lidar com essa dependência?

Redes Bayesianas: Um modelo gráfico baseado em variáveis condicionalmente independentes.



Regressão Logística

Contexto

$$\text{Regressão : } Y = X\beta + \epsilon.$$

Logística? Quando? Quando Y é **qualitativa**.

Ideia! E se nós codificarmos Y ? _____

Assim podemos continuar usando a regressão linear usual, e.g.,
Queremos saber o que ocorreu com um paciente com base em seus sintomas

$$Y = \begin{cases} 1 & \text{se overdose de drogas} \\ 2 & \text{se ataque epilético} \end{cases}$$

Problema!

Este tipo de codificação implica num ordenamento das respostas.



Por que usar Regressão Logística?



Ok, mas e se Y tiver uma ordenação natural?
e.g., leve, moderado e severo.

Se Y for binária a regressão linear até que funciona.

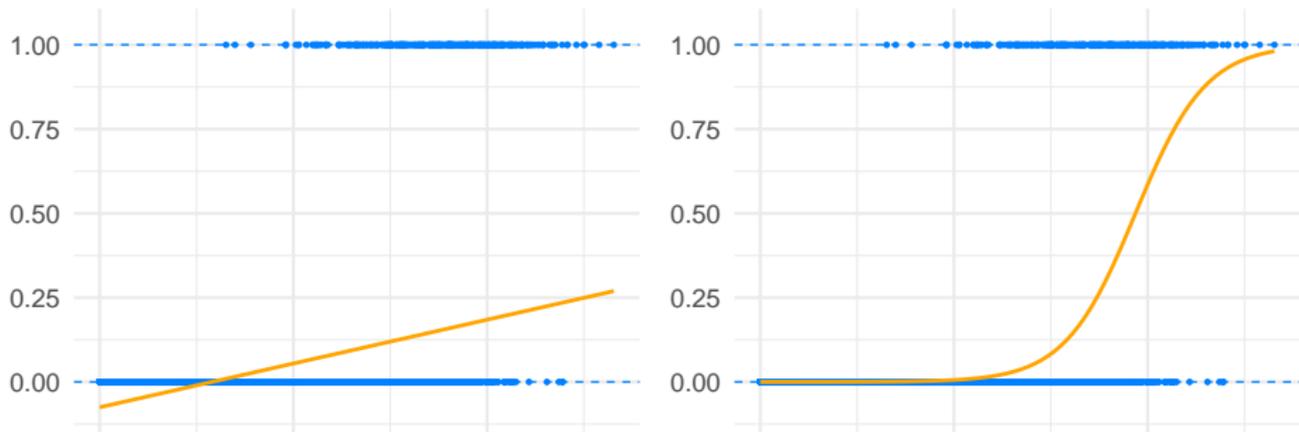
» Contudo, as previsões podem ficar fora do intervalo $[0, 1]$.

Por razões como esta que é preferível o uso de métodos de classificação próprios para variáveis qualitativas.

Métodos de classificação próprios? **Regressão Logística.**



Regressão Linear × Regressão Logística



De onde vem esta forma em S? **função logística** : $\frac{e^{X\beta}}{1 + e^{X\beta}}$.





Ok, mas onde e como se usa essa **função logística**?

Num modelo de regressão temos $\mathbb{E}[Y|X] = \mu = g^{-1}(X\beta)$.

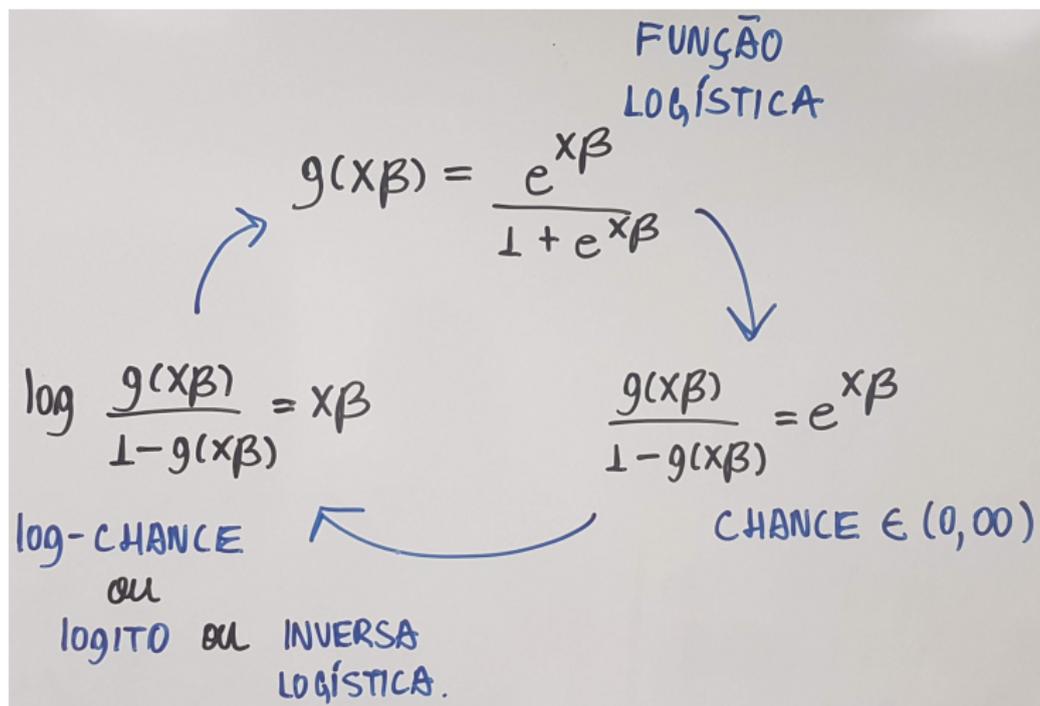
- » No caso Normal, g é uma função identidade.
O que configura a regressão linear que todos conhecemos, $Y = X\beta + \epsilon$.
- » Se assumida uma distribuição diferente da Normal para $Y|X$, g será diferente da função identidade e assim teremos o que configura os chamados GLMs.

Regressão Logística

Se Y for dicotômica e g for a função logística, então temos uma regressão logística.



Interpretação?



Interpretação? Razão de chances

Chances,

$$\text{chance}(X) = e^{X\beta} = \frac{g(X\beta)}{1 - g(X\beta)}.$$

e.g.,

$$g(X\beta) = 0.2 \Rightarrow \frac{0.2}{1 - 0.2} = \frac{1}{4}, \quad g(X\beta) = 0.9 \Rightarrow \frac{0.9}{1 - 0.9} = 9.$$

Razão de chances,

Para uma variável contínua:

$$\frac{\text{chance}(x + 1)}{\text{chance}(x)} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}.$$



Máxima verossimilhança

O quê? _____

Função de verossimilhança é o nome dado a função que precisamos maximizar.

Verossimilhança

$$L(\beta) = \prod_{i: Y_i=1} g(X_i\beta) \prod_{i': Y_{i'}=0} (1 - g(X_{i'}\beta))$$

Por quê? _____

Queremos estimativas para β , e tais estimativas são obtidas via a maximização de $L(\beta)$.

Como? _____

$L(\beta)$ é uma função "qualquer" que queremos otimizar. Dependendo da função uma solução analítica pode não existir, e aí métodos numéricos se fazem necessário.



Software: R

Naive Bayes, _____

```
library(e1071)

modelo <- naiveBayes(Y ~ x1 + x2 + x3, data = dados)
## ou
modelo <- naiveBayes(Y ~ ., data = dados)
## se Y, x1, x2 e x3 forem todas as colunas de "dados"
```

Regressão Logística, _____

```
modelo <- glm(Y ~ x1 + x2 + x3,
              family = binomial(link = "logit"), data = dados)
## ou
modelo <- glm(Y ~ ., family = binomial(link = "logit"), data = dados)
## se Y, x1, x2 e x3 forem todas as colunas de "dados"
```



THANK YOU



mamagenerator.net

