Regressão Logística

ESTAT0016 – Tópicos Especiais em Estatística (Introdução à Apredizagem de Máquina)

Prof. Dr. Sadraque E.F. Lucena

Regressão Logística

- A regressão logística é um modelo usado para uma variável resposta que possui duas categorias.
- Ou seja, o modelo de regressão logística pode ser compreendido como um modelo de classificação.
- Exemplos:
  - Prever a presença de uma doença em um paciente com base em fatores como idade, histórico familiar e resultados de testes.
  - Prever se um cliente farão ou não uma compra após receber um e-mail promocional.
  - Prever se um cliente irá pagar um impréstimo com base no histórico de crédito, renda, e outros fatores financeiros.

Formulação

- Considere y_i como uma variável aleatória de Bernoulli com a seguinte distribuição de probabilidade:

$$\begin{cases} y_i = 1, & \text{com } P(y_i = 1) = \pi_i, \qquad \text{(resultado positivo)} \\ y_i = 0, & \text{com } P(y_i = 0) = 1 - \pi_i \quad \text{(resultado negativo)} \end{cases}$$

- Considere também um conjunto de variáveis explicativas $\underset{\sim}{x} = x_1, x_2, \ldots, x_p$.
- A regressão logística foca na modelagem da probabilidade de ocorrer $y_i = 1$, dados os valores das variáveis explicativas $\underset{\sim}{x} = x_{1i}, x_{2i}, \ldots, x_{pi}$, ou seja, queremos modelar

$$P(y_i | \underset{\sim}{x}) = \pi_i$$

Formulação

- Um modelo de regressão compreende três elementos fundamentais: a resposta, os preditores e os coeficientes.
- A abordagem ideal para atingirmos nosso objetivo seria empregar um modelo do tipo

$$\pi_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

- No entanto, é importante observar que:
  - No lado esquerdo da equação, temos que $\pi_i = P(y_i = 1 | \underset{\sim}{x}) \in [0, 1]$
  - Enquanto que no lado direito, os valores estão em $(-\infty, +\infty)$
  - Isso pode resultar em estimativas negativas da probabilidade ou valores superiores a 1, o que torna a igualdade inválida.

Formulação

- Para resolver esse problema definimos então o modelo de regressão logístico definido como

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

- A função $\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ é conhecida como função logística.
  - A razão de probabilidades $\frac{\pi_i}{1-\pi_i}$ é usada porque ela varia no intervalo $[0, +\infty)$.
  - Ao aplicar o logaritmo natural, a função logit transforma a razão de probabilidades para o intervalo $(-\infty, +\infty)$, o que é conveniente para a modelagem linear.
  - Desta forma, nos dois lados da equação temos componentes que podem assumir valores no intervalo $(-\infty, +\infty)$.
- O termo η_i é chamado preditor linear.

UNIVERSIDADE FEDERAL DE SERGIPE

Seleção de variáveis

- No modelo de regerssão logística os parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ são estimados por máxima verossimilhança.
- Dentre os vários modelos possíveis, selecionamos aquele com menor Critério de Informação de Akaike (AIC) ou Critério de Informação Bayesiano (BIC).
  - Podemos usar a função `stepAIC()` no R para selecionar as variáveis.
- Após o uso de AIC, escolhemos as variáveis significativas no modelo usamos:
  - Teste da razão de verossimilhanças;
  - Teste de Wald;
  - Teste escore.
- Apenas as variáveis com p-valor < 0.05 devem permanecer no modelo.

Predição do modelo

- Como o modelo de regressão logística é dado por

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{1}$$

- Podemos calcular a probabilidade de ocorrer um resultado positivo, $\pi_i = P(y_i = 1 | \underset{\sim}{x})$, dados os valores de $x_{1i}, \ldots, x_{pi}$ da seguinte forma:

$$\begin{aligned}\pi_i &= \frac{exp(\eta_i)}{1 + exp(\eta_i)} \\ &= \frac{exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}\end{aligned}$$

Exemplo 5.1

Suponha que ajustamos um modelo de regressão logística para classificar um cliente de um banco como *mau pagador* ($y = 1$) ou *bom pagador* (y=0) com base no seu saldo devedor (em $). O modelo estimado foi:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -10{,}6513 + 0{,}0055\, saldo\_devedor$$

a. Qual a probabilidade de uma pessoa com saldo devedor de $1500 não pagar o banco?

b. E saldo de $2500?

Odds Ratio - Razão de Chances

- No modelo de regressão logística, os parâmetros têm uma interpretação particular, chamada de *odds ratio*.
- A *odds ratio* (OR) é a medida da chance de um resultado ser classificado como positivo em comparação com a chance de ser classificado como negativo.
- A partir da Equação 1 podemos escrever

$$\frac{\pi_i}{1-\pi_i} = exp(\eta_i) = exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)$$

- Assim, podemos mensurar o acréscimo (ou decréscimo) da probabilidade π_i baseado na oscilação das variáveis explicativas.

Odds Ratio - Razão de Chances

- A interpretação dos coeficientes é a seguinte:
  - $exp(\beta_i) > 1$: a probabilidade de um resultado ser classificado como positivo é maior do que a de ser classificado como negativo.
    - **Exemplo:** se $exp(\beta_i) = 1{,}5$, para um aumento de uma unidade na variável preditora, a chance de um resultado positivo é $1{,}5$ vezes (ou 50% maior do que) a chance de um resultado negativo.
  - $exp(\beta_i) < 1$: a chance de classificação positiva é menor que a de classificação negativa.
    - **Exemplo:** se $exp(\beta) = 0{,}7$, para um aumento de uma unidade na variável preditora, a chance de um resultado positivo é 70 da chance de um resultado negativo.
      - Outra forma: a chance de um resultado negativo é aproximadamente 42,88% maior $(1/0{,}7 \approx 1{,}4286)$ que a chance de um resultado positivo.

Exemplo 5.2

Agora suponha que ajustamos o seguinte modelo para a variável resposta com as classes *mau pagador* ($y = 1$) ou *bom pagador* (y=0):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -10{,}8690 + 0{,}0057\, saldo\_devedor + 0{,}000\, salario - 0{,}6468\, estudante$$

- $saldo\_devedor$: dado em $
- $salario$: dado em $
- $estudante$: sim (1) ou não (0)

a. Interprete cada coeficiente.

b. Qual a probabilidade de uma pessoa com saldo devedor de $1000, salário $5000 e estudante não pagar o banco?

c. E se o saldo devedor for $2000?

Como fazer a classificação

- A previsão de um modelo de regressão logística é a probabilidade estimada de ocorrer um resultado positivo.
- Para usarmos essa previsão como um classificador, usamos um ponto de corte c tal que

$$\begin{cases} \pi_i \geq c & \Rightarrow y_i = 1 \text{ (positivo)} \\ \pi_i < c & \Rightarrow y_i = 0 \text{ (negativo)} \end{cases}$$

- Vejamos como determinar o ponto de corte.

Como fazer a classificação

- Uma forma de definirmos o ponto de corte c é escolher o valor que fornece maior acuária preditiva para o modelo.
- A acurácia é a proporção total de predições classificadas corretamente em relação ao número total de observações. Ela pode ser obtida a partir da matriz de confusão.

Matriz de confusão

	Negativo (real)	**Positivo (real)**
Negativo (predito)	Verdadeiro Negativo (VN)	Falso Negativo (FN)
Positivo (predito)	Falso positivo (FP)	Verdadeiro Positivo (VP)

$$\text{acurácia} = \frac{acertos}{total} = \frac{\text{VN} + \text{VP}}{\text{VN} + \text{FN} + \text{FP} + \text{VP}}$$

Exemplo 5.3

Considere o ponto de corte $c = 0.5$, faça a classificação e calcule a acurácia:

Caso 1:

Real	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**
π_i	0.18	0.44	0.15	0.39	0.62	0.47	0.37	0.59	0.42	0.15

Real	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**
π_i	0.87	0.66	0.07	0.77	0.91	0.74	0.66	0.27	0.78	0.71

Caso 2:

Real	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**
π_i	0.31	0.32	0.69	0.22	0.11	0.45	0.06	0.15	0.11	0.35

Real	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**P**	**P**	**P**
π_i	0.17	0.62	0.10	0.18	0.36	0.10	0.29	0.75	0.45	0.11

Como fazer a classificação

- No exemplo anterior, a acurácia é semelhante para ambos os modelos, mas no caso 2 o modelo comete mais erros do que acertos nas previsões positivas. Isso é comum quando há poucas instâncias pertencentes a uma classe, resultando em dados desbalanceados. Em tais casos, a acurácia não é um indicador adequado para avaliar o desempenho do modelo.
- À medida que aumentamos o valor de corte, classificamos mais observações como negativas.
  - Portanto, aumentamos o número de falsos negativos e reduzimos o número de falsos positivos.
- Uma forma de escolher o melhor ponto de corte é usando a curva ROC. Vejamos.

Curva ROC

- ROC é uma abreviação para *Receiver Operating Characteristic* (Característica de Operação do Receptor).
- Ela utiliza duas métricas: sensibilidade e $1-$ especificidade.
  - Sensibilidade: proporção de positivos classificados corretamente:

$$sensibilidade = \frac{veradadeiros\ positivos}{total\ de\ positivos} = \frac{VP}{VP + FP}$$

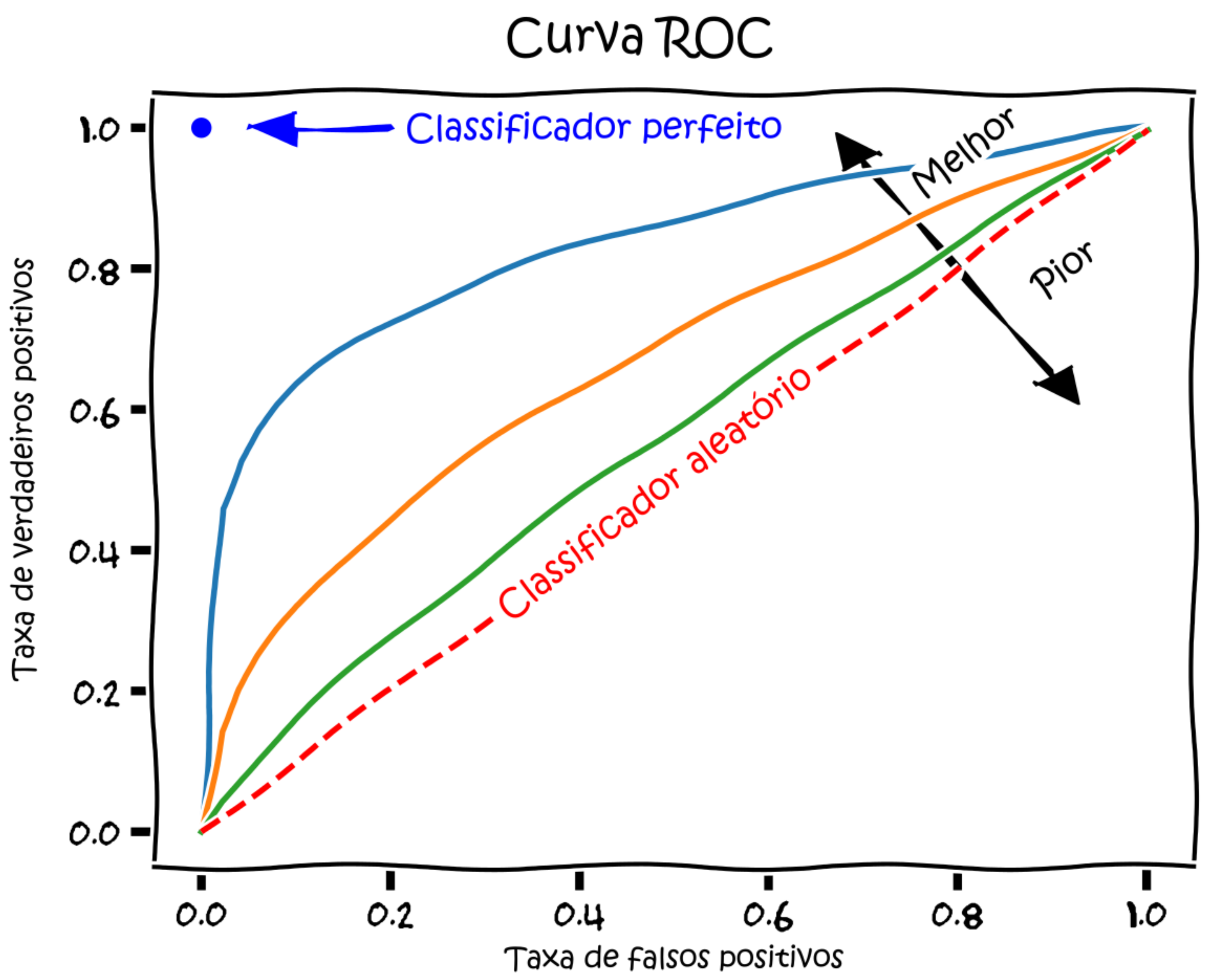
  - Especificidade: proporção de negativos classificados corretamente:

$$especificidade = \frac{veradadeiros\ negativos}{total\ de\ negativos} = \frac{VN}{VN + FN}$$

- Note que $1 - \ especificidade$ é a proporção de positivos classificados incorretamente.

Curva ROC

- A curva ROC nos fornece um gráfico entre sensibilidade e $1-$ especificidade quando aumentamos o valor de corte.



FONTE: Wikipedia.

- A curva sob a curva, conhecida com AUC (*area under the curve*), é usada como uma medida de qualidade do ajuste da regressão logística. Quanto maior, melhor (máximo: 1).

Curva ROC

- Baseado na curva ROC, o ponto de corte é aquele com maior equilíbrio entre a sensibilidade a especificidade.
- Geralmente aquele mais próximo do canto superior esquerdo da curva, representando alta sensibilidade e alta especificidade.

Validação do modelo

- Dividimos a amostra aleatoriamente em duas partes
  - Treinamento
  - Teste
- O modelo é ajustado com os dados da amostra treinamento e usado para prever as respostas da amostra teste.
- Objetivo: evitar superestimação do modelo.

Validação cruzada

- Dividimos a amostra em k partes iguais.
- Separamos uma parte k e ajustamos o modelo nas outras $k-1$ partes conjuntamente. Fazemos esse procedimento para todas as partes e combinamos os resultados.

Agora vamos ajustar um modelo no R