IA e ML Aplicados a Finanças

Prof. Leandro Maciel

AULA 7: Classificação e Regressão Logística

Agenda

1. Classificação

1. Classificação

- Regressão/previsão: variável dependente é uma variável contínua;
- Não estamos mais interessados em explicar o salário, preço, quantidade...
- Analista está interessado em **classificar** um item em uma **categoria**;
- Estimar a probabilidade de um item pertencer a uma categoria;
- **Modelos de classificação** ou **classificadores**:

  Reconhecimento de imagens, facial, spams, rating etc...

1. Classificação

- Classificação como um mapeamento entrada-saída:

$$y = f(x_1, x_2, \ldots)$$

- $y \rightarrow$ variável resposta ou **rótulo**, assumindo k valores distintos;
- k representa o número de classes, rótulos ou categorias;
- $x_1, x_2, \ldots \rightarrow$ variáveis explicativas ou **atributos** (*features*);
- Classificador $f(\cdot)$ faz o mapeamento associado;
- Classifica objeto com atributos $(x_1, x_2, \ldots)$ a uma classe em y.

1. Classificação

- Aprendizagem em classificação:
  - supervisionada $\rightarrow$ fornecemos as classes no conjunto treinamento;
  - pressuposto $\rightarrow$ objetos de mesma classe têm atributos similares;
  - classificador define a regra de associação.
- Aplicações em administração e finanças:
  - maketing direcionado, padrão de consumo/sentimento de clientes;
  - categorização de produto, modelagem de crédito, regras de *trading*.

2. Regressão Logística

2. Regressão Logística

- Modelo clássico (estado-da-arte): **Regressão Logística**;
- Variável dependente y categórica $\rightarrow$ 0 ou 1 (fracasso ou sucesso);
- Estamos interessados em prever a probabilidade de sucesso de y:

$$\Pr(y = 1) = f\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\right)$$

- Se y binária, modelo binomial, c.c. regressão multinomial;
- Objetiva-se determinar uma prob. condicional: $\Pr(y = 1|x_1, x_2, \ldots, x_k)$.

- Precisamos de uma função que respeite restrição $0 \leq \Pr(y = 1|x) \leq 1$;
- **Modelo geral de regressão logística:**

$$\Pr(y = 1) = g(x) = \frac{1}{1 + e^{-f(x)}}, \; f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

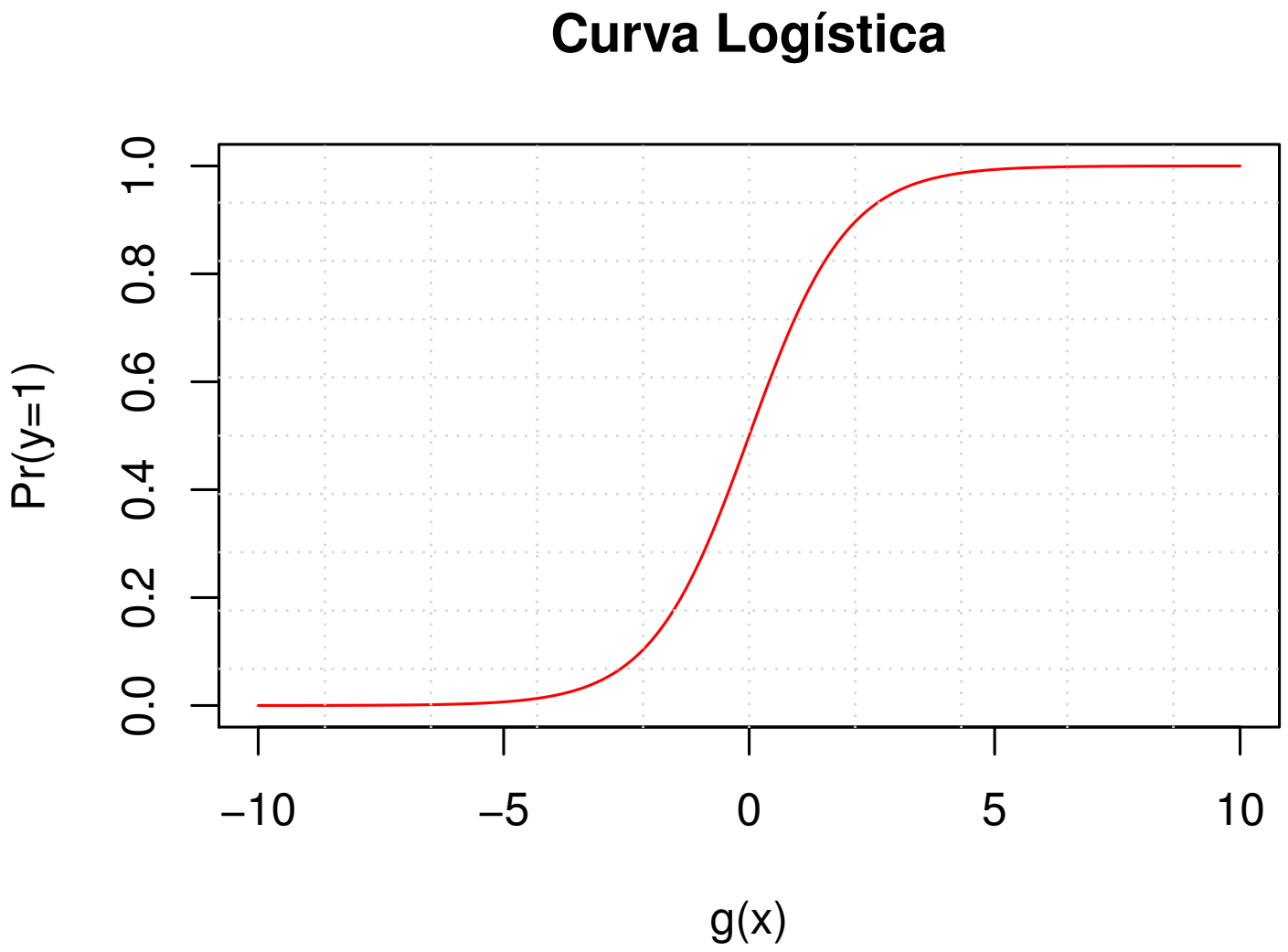
- $g(x)$ é a chamada função logística;
- $f(x)$ é uma função linear dos atributos

- Função logística $g(x)$ apresenta as seguintes características:

$$g(x) \to +\infty \quad \Rightarrow \quad \Pr(y = 1) \to 1$$

$$g(x) \to -\infty \quad \Rightarrow \quad \Pr(y = 1) \to 0$$

- Coeficientes estimados por Máxima Verossimilhança: $\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\}$;
- Estimativas maximizam a probabilidade de ocorrência da amostra.


Curva Logística
Pr(y=1)
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
g(x)

- Qual interpretação dos parâmetros estimados?
- Podemos escrever a **razão de chances**:

$$log\left[\frac{\Pr(y=1)}{1-\Pr(y=1)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- $\frac{\Pr(y=1)}{1-\Pr(y=1)} \rightarrow$ prob. sucesso em relação prob. fracasso;
- $\Pr(y=1) = 0,8 \rightarrow 0,8/0,2 = 4$: prob. sucesso 4x maior que de fracasso.

- Interpretação dos parâmetros estimados:

  "**Impacto da variável sobre a razão de chances de ocorrência**"

- Positivo $\rightarrow$ aumenta a probabilidade relativa de ocorrência do evento;

- Negativo $\rightarrow$ reduz a probabilidade relativa de ocorrência do evento;

- Regra de classificação em modelos logísticos de acordo com a saída:

  Se $\Pr(y = 1) > 0,5 \rightarrow y = 1$;

  Se $\Pr(y = 1) < 0,5 \rightarrow y = 0$.

3. Rating de Crédito

3. Rating de Crédito

- **Definir prob. default no pagamento fatura de cartão de crédito**...
- Categorias/classes $\rightarrow$ prob. default igual a 1 ou 0;
- Atributos $\rightarrow$ características dos clientes;
- Dados: "*Default of credit card clients Data Set*";
- Fonte *UC Irvine Machine Learning Repository*:

`https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

- Estimar o seguinte modelo de regressão logística:

$$\Pr(y = 1) = g(x) = \frac{1}{1 + e^{-f(x)}}, \text{ com}$$

$$f(x) = \beta_0 + \beta_1 limite + \beta_2 sexo + \beta_3 educacao + \beta_4 estadoCivil...$$

- $y = 1 \rightarrow$ default; $y = 0 \rightarrow$ não default;
- Seleção de variáveis: *expert knowledge* e técnicas lineares/não-lineares...

3. Rating de Crédito

- Divisão da amostra em:

  Treinamento ($\sim 75\%$) e **Teste** ($\sim 25\%$) (*holdout sample*);

- Vários métodos de divisão: **validação-cruzada** (*cross-validation*);

- Amostras devem ser **representativas** - evitar *overfitting*;

- Classes representadas proporcionalmente nas amostras;

- Seleção aleatória como mecanismo apropriado.

3. Rating de Crédito

- Coeficientes negativos e positivos:

  Positivo $\rightarrow$ aumenta a probabilidade relativa de ocorrência do evento;

  Negativo $\rightarrow$ reduz a probabilidade relativa de ocorrência do evento;

- Positivo aumenta a razão de changes de não pagamento (default);
- Negativo reduz a razão de changes de não pagamento (default);
- Resultados na amostra teste (desconhecida)?

- **Matriz de confusão** para avaliar a qualidade do classificador;
- Relaciona número de classificações corretas e falsas nas classes.

Classes	*Referência*	
Previsão	0	1
0	V_N (Verdadeiro Negativo)	F_N (Falso Negativo)
1	F_P (Falso Positivo)	V_P (Verdadeiro Positivo)

- Qual erro (falso) é pior? Depende da aplicação em questão;
- Prob. de default, câncer, acesso por digital conta bancária...

- **Taxa de acerto total** (medida de acurácia):

$$\text{Acurácia} = \frac{V_N + V_P}{V_N + V_P + F_N + F_P}$$

- **Taxa de erro total**:

$$\text{Taxa de Erro} = 1 - \text{Acurácia}$$

$$= 1 - \frac{V_N + V_P}{V_N + V_P + F_N + F_P} = \frac{F_N + F_P}{V_N + V_P + F_N + F_P}$$

- Outras medidas de qualidade do classificador...
- **Sensitivity** $\rightarrow$ V_P corretamente identificado (percentual):

$$\text{Sensitivity} = \frac{V_P}{V_P + F_N}$$

- **Specificity** $\rightarrow$ V_N corretamente identificado (percentual):

$$\text{Specificity} = \frac{V_N}{V_N + F_P}$$

- Aplicação define a medida de qualidade a ser maximizada/minimizada.

Problema

Obtenha um melhor modelo de regressão logística para os dados de cartão de crédito incluindo as demais variáveis que estão disponíveis. Em seguida, faça a divisão das amostras entre treinamento e teste de forma aleatória (mantendo as mesmas proporções) e verifique se o seu modelo tem o desempenho afetado de forma significativa.

- Próxima aula...
  - outros classificadores em ML;
  - algoritmo K-NN...

4. Bibliografia

HASTIE, Trevor, TIBSHIRANI, R., & FRIEDMAN, Jerome. **The Elements of Statistical Learning**. Data Mining, Inference, and Prediction. 2 Ed. Springer, 2008. Capítulo 4.

WOOLDRIDGE, Jeffrey M. **Introdução à Econometria** - Uma abordagem Moderna. Tradução da 4ª Ed. São Paulo: CENGAGE Learning, 2011. Capítulo 17.

ADLER, Joseph. **R in a Nutshell** - A Desktop Quick Reference. 2 Ed. Sebastopol, CA: O'Reilly, 2012. Capítulo 21.

Prof. Leandro Maciel

leandromaciel@usp.br