Noções de Inferência
Regressão Logística

Introdução

Preditiva.ai

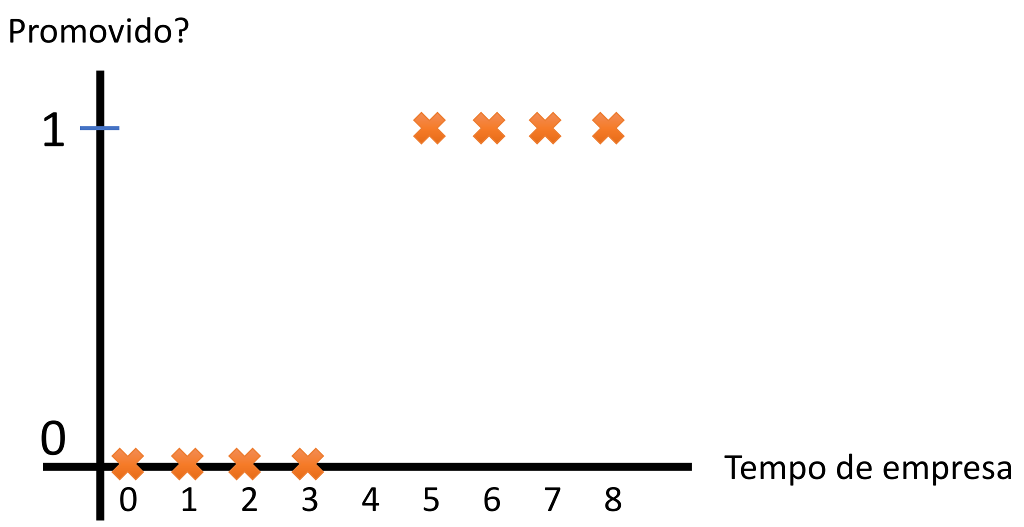
Regressão Logística
Motivação

Como vimos na últimas aulas, a regressão linear múltipla é uma boa técnica para ajustar *features* a uma variável resposta **quantitativa**. Mas nem sempre queremos modelar variáveis deste tipo. Em vez disso, e se quiséssemos ajustar um modelo de variável resposta **qualitativa**, como por exemplo: Funcionário Promovido ou Não Promovido? Podemos tentar usar a mesma técnica de regressão. Veja:

Y: (Func Promovido?)	Tempo de Empresa
Não	0
Não	1
Não	2
Não	3
Sim	5
Sim	6
Sim	7
Sim	8

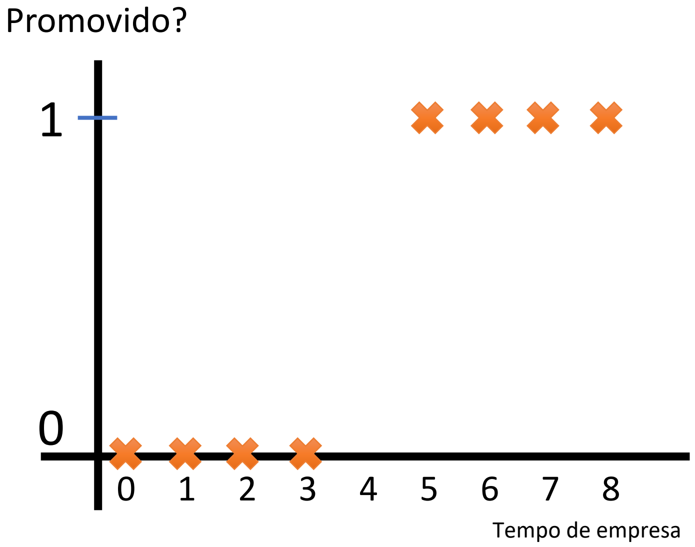
Supondo que funcionários promovidos sejam 1 e funcionários não promovidos sejam 0.

Podemos plotar o seguinte gráfico:

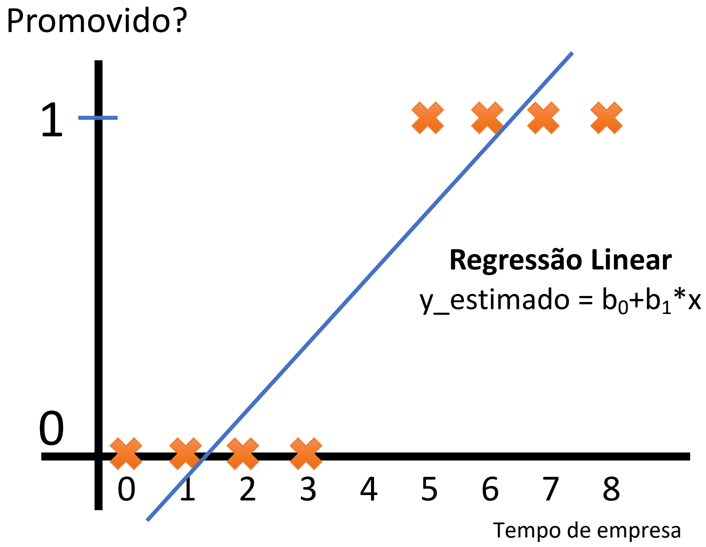


Regressão Logística

Motivação

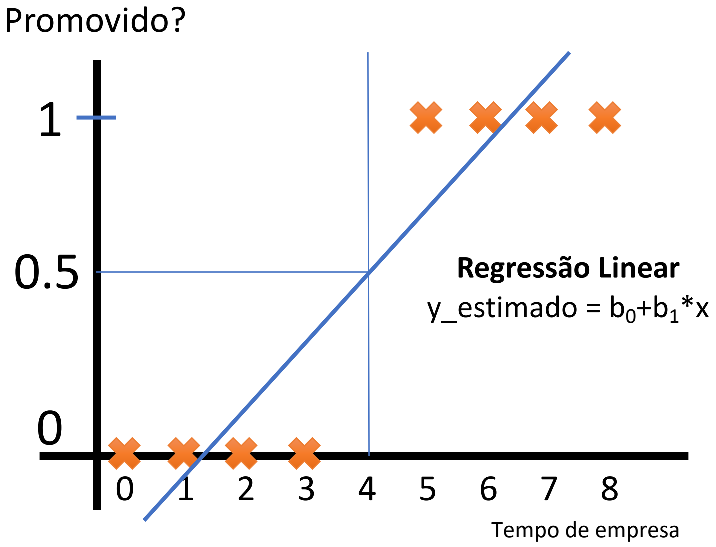


Passo 1



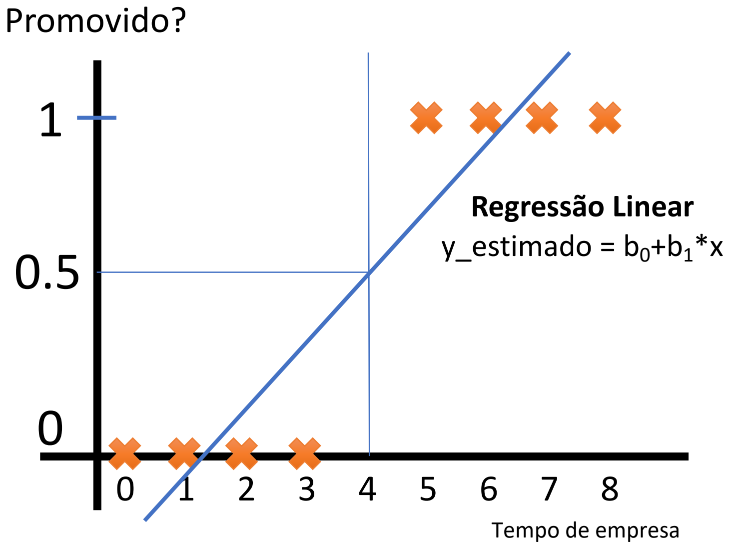
Podemos ajustar uma reta de regressão conforme acima.

Passo 2

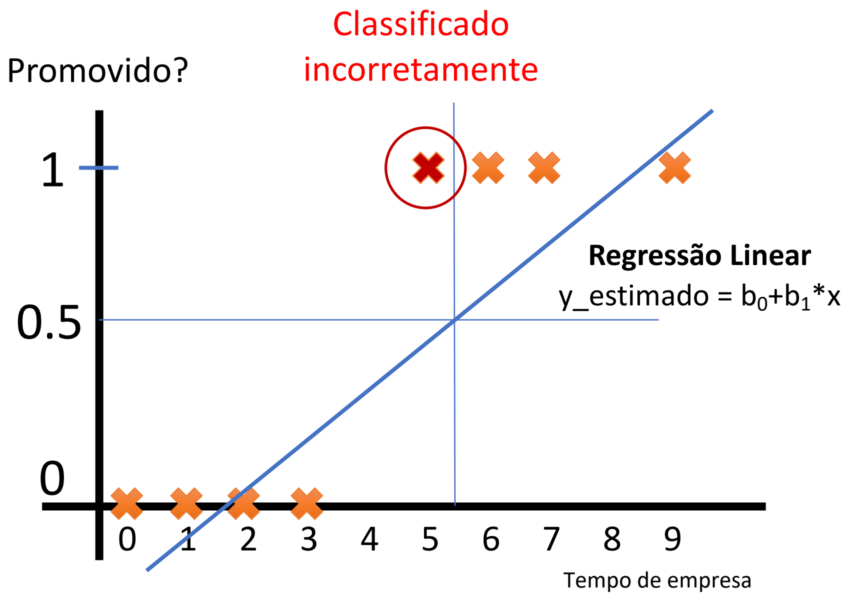


Com a reta ajustada, podemos dizer que valores **de y_estimado > 0,5 são de funcionários Promovidos**. Da mesma forma, Não Promovidos quando y_estimado < que 0,5.

Regressão Logística
Motivação



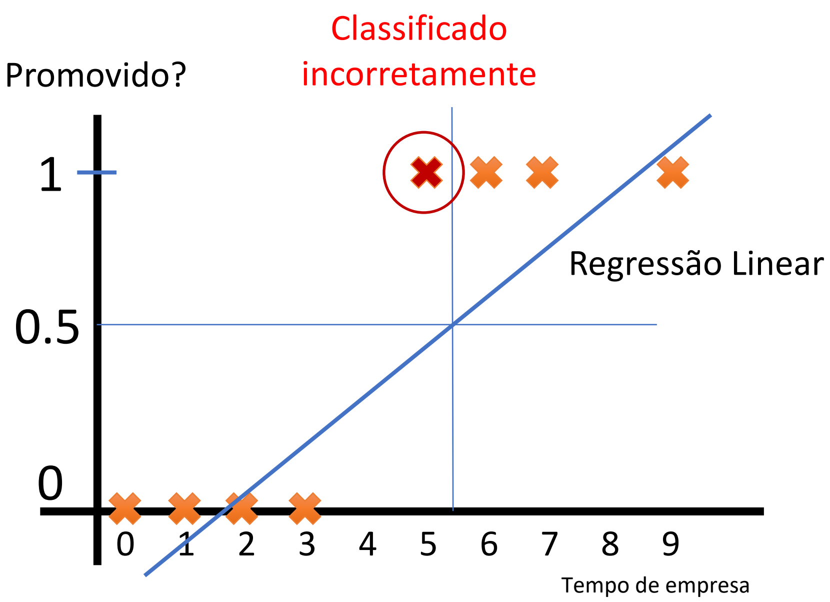
Mas o que acontece quando adicionamos mais uma observação?



Usar o mesmo critério de y > 0,5 **não classifica corretamente** todos os funcionários promovidos.

Regressão Logística

Motivação



Usar o mesmo critério de y > 0,5 agora **não classifica corretamente** todos os funcionários promovidos.

Outro **problema** dessa abordagem é que os valores estimados de y podem assumir valores menores que 0 ou maiores que 1, o que não faz sentido devido à própria característica binária do problema de classificação (no exemplo, 1: Promovido e 0: Não Promovido).

Para resolver esse problema que surge a **regressão logística**, um tipo de modelo que ajusta variáveis respostas qualitativas melhor que a reta de regressão linear.

Além disso, ela tem a propriedade de 0 < y_estimado < 1.

Regressão Logística

Introdução

Para modelar o problema de classificação, a regressão logística trabalha com uma **transformação** de variável conforme a seguir:

1) Criamos uma variável Z que é a regressão linear múltipla das features do problema:

$$z = \beta_0 + \beta_1 * x$$

2) Fazemos uma transformação da variável Z conforme abaixo:

$$g(z) = \frac{1}{1 + e^{-z}}$$

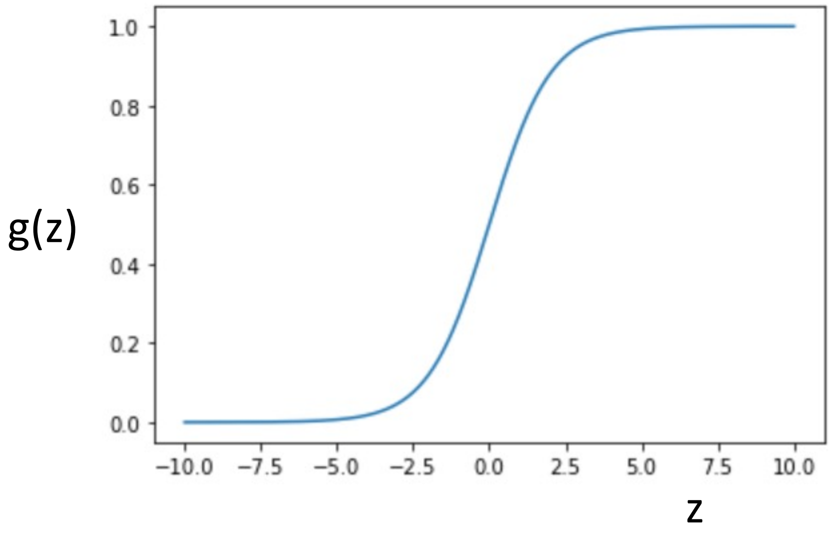
(Função Sigmoid ou Função Logística)

Essa transformação usando a **função de Sigmoid** garante que o y_estimado esteja entre 0 e 1.

3) Por fim, estimamos a probabilidade de Y = 1 usando a seguinte relação:

$$P(y = 1) = g(z)$$

Exemplo de Função de Sigmoid



Regressão Logística
Exemplo

Aplicando os coeficientes na tabela, temos o seguinte resultado:

Supondo que: $z = -4.5446 + 1.1361 * x$

Y: (Func Promovido?)	Tempo de Empresa	z	g(z)
Não	0	-4,5446	0,0105
Não	1	-3,4085	0,0320
Não	2	-2,2723	0,0934
Não	3	-1,1362	0,2430
Sim	5	1,1361	0,7570
Sim	6	2,2723	0,9066
Sim	7	3,4084	0,9680
Sim	8	4,5446	0,9895

$$P(y = 1) = g(z) = \frac{1}{1 + e^{-z}}$$

Probabilidade de Y = 1 é próxima de 0 quando o funcionário **não** foi promovido.

Probabilidade de Y = 1 é próxima de 1 quando o funcionário foi promovido.

Regressão Logística

Exemplo

Aplicando os coeficientes na tabela, temos o seguinte resultado:

Y: (Func Promovido?)	Tempo de Empresa	z	g(z)	Y Estimado
Não	0	-4,5446	0,0105	Não
Não	1	-3,4085	0,0320	Não
Não	2	-2,2723	0,0934	Não
Não	3	-1,1362	0,0243	Não
Sim	5	1,1361	0,7570	Sim
Sim	6	2,2723	0,9066	Sim
Sim	7	3,4084	0,9680	Sim
Sim	8	4,5446	0,9895	Sim

Finalmente, para chegarmos ao Y Estimado basta adotarmos um **corte (threshold)** de probabilidade. Um valor padrão é **0,5**.

$$y_{\text{estimado}} = \begin{Bmatrix} Sim\ quando\ g(z) \geq 0{,}5 \\ Não\ quando\ g(z) < 0{,}5 \end{Bmatrix}$$

Regressão Logística

Exemplo

Por fim, podemos medir a **acurácia** do modelo com uma medida muito simples. Veja:

	Y Estimado	
Y Real	Sim	Não
Sim	4	0
Não	0	4

$$Acurácia = \frac{soma\ dos\ acertos}{total\ de\ linhas\ da\ base} = \frac{4+4}{8} = 100\%$$

A matriz acima chamada **Matriz de Confusão (Confusion Matrix)** é muito importante para medir a performance de modelos de classificação como esse. Vamos falar mais dessa técnica nas próximas aulas.

Preditiva.ai

Aprendizado Supervisionado
Regressão Logística

Demonstração

Demonstração

Regressão Logística no Knime

Regressão Logística

Base Diabetes

Preditiva.ai

Hands on

Ajuste um modelo de regressão logística e calcule sua acurácia

Roteiro:

1. Importe a base “diabetes.csv”;
2. Ajuste um modelo de regressão logística (Target = “Outcome”);
3. Exporte os coeficientes do modelo para o Excel;
4. No Excel, crie uma coluna chamada Y_Estimado usando o corte de 0,5 na variável g(z) utilizando os coeficientes do modelo;
5. Calcule a acurácia do modelo.

Preditiva.ai

Aprendizado Supervisionado
Regressão Logística

Interpretando os coeficientes do modelo

Regressão Logística

Interpretabilidade dos coeficientes

Preditiva.ai

Diferentemente da Regressão Linear Múltipla, em que o Target é linearmente correlacionado com a combinação linear de coeficientes e Features (ou seja, $y = \beta_0 + \beta_1$*x), na **Regressão Logística apenas o ln(Odds) é correlacionado com tal combinação**. Veja:

Regressão Linear Múltipla:

$$y = \beta_0 + \beta_1 * x$$

Regressão Logística:

$$\ln(Odds) = \beta_0 + \beta_1 * x$$

Sendo que:

$$Odds = \frac{P(y=1)}{1 - P(y=1)}$$

E de onde vem isso?

$$z = \beta_0 + \beta_1 * x$$

$P(y=1) = \frac{1}{1+e^{-z}}$ e $1 - P(y=1) = 1 - \frac{1}{1+e^{-z}}$

$$\frac{P(y=1)}{1-P(y=1)} = \frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} = \frac{\frac{1}{1+e^{-z}}}{\frac{1+e^{-z}-1}{1+e^{-z}}}$$

$$\frac{P(y=1)}{1-P(y=1)} = e^{z}$$

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = z = \beta_0 + \beta_1 * x$$

Conclusão: Na Regressão Logística, as variáveis explicativas **(features) são lineares com o Ln (Odds)**.

Regressão Logística
Interpretabilidade dos coeficientes

Preditiva.ai

E como isso ajuda na interpretação dos coeficientes da Regressão? Basta fazer mais algumas manipulações matemáticas e chegamos na interpretação:

Vamos chamar de $p = P(y=1)$:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x_1$$

$$e^{\ln\left(\frac{p}{1-p}\right)} = e^{\beta_0+\beta_1 * x_1}$$

$$\frac{p}{1-p} = e^{\beta_0} * e^{\beta_1 * x_1}$$

Com isso, o que acontece se aumentarmos a variável x1 em uma unidade?

$$\frac{p}{1-p} = e^{\beta_0} * e^{\beta_1 * (x_1+1)}$$

Aumentamos X1 em uma unidade

$$\frac{p}{1-p} = e^{\beta_0} * e^{\beta_1 x_1} * e^{\beta_1}$$

O coeficiente da variável que aumentou vira uma exponencial.

$$\frac{p}{1-p} * e^{\beta_1}$$

Conclusão: Diferentemente da Regressão Linear Múltipla, em que aumentos das varáveis explicativas produzem aumentos em Y (target) diretamente, na **Regressão Logística,**

- **Ao se aumentar em x unidades a variável X_n , a Odds (p/1-p) é multiplicada por $e^{x\beta_n}$**
- **Ao se diminuir em x unidades a variável X_n , a Odds (p/1-p) é multiplicada por $e^{-x\beta_n}$**

Regressão Logística

Interpretabilidade dos coeficientes: Exemplo 1

Preditiva.ai

Em uma operação de financiamento de veículos, um modelo logístico foi treinado. Sendo P(y=1) a probabilidade do cliente ser um BOM pagador, interprete o modelo:

$$Ln(Odds) = 0,5 + 0,3 * Valor\ de\ Entrada\ (em\ milhares) - 0,4 * Qte\ de\ Dívidas\ no\ mercado$$

Possíveis interpretações:

1. Não dar entrada e não ter dívidas no mercado ($e^{0,5}$=1,65) resulta em uma chance de 1,65 de ser bom cliente sobre não ser bom cliente.

2. A cada **aumento** de R$ 1 Mil Reais de Entrada ($e^{0,3}$=1,35) aumenta em 35% a chance de ser bom cliente.

3. A cada **diminuição** de R$ 1 Mil Reais de Entrada ($e^{-0,3}$=0,74) diminui em (1 - 0,74)% = 26% a chance de ser bom cliente.

4. A cada **aumento** de uma dívida no mercado ($e^{-0,4}$= 0,67) **diminui** em (1-0,67)% = 33% a chance de ser bom cliente.

Dicas gerais:

- Aumentos de variáveis com coeficientes positivos produzem aumentos na P (Y=1).
- Diminuições de variáveis com coeficientes positivos produzem diminuições na P (Y=1).
- Aumentos de variáveis com coeficientes negativos produzem diminuições na P (Y=1).
- Diminuições de variáveis com coeficientes negativos produzem aumentos na P (Y=1).

Regressão Logística
Interpretabilidade dos coeficientes: Exemplo 2

Preditiva.ai

Em uma operação de financiamento de veículos, um modelo logístico foi treinado. Sendo P(y=1) a chance do cliente ser um BOM pagador, interprete o modelo com o **acréscimo da varíavel categórica "escolaridade"** usando as seguintes Dummies:

	Dummies	
Escolaridade	**D_Medio**	**D_Graduacao**
Ensino Medio	1	0
Graduação	0	1
Pós Graduação	0	0

$$Ln(Odds) = 0{,}6 + 0{,}2 * Valor\ de\ Entrada\ (em\ milhares) - 0{,}1 * Qte\ de\ Dívidas - 0.05 * D_{Medio} - 0.01 * D_{Graduacao}$$

Possíveis interpretações:

1. Não dar entrada, não ter dívidas no mercado e ter escolaridade "Pós Graduação" ($e^{0,6}$= 1,82, pois D_Medio = 0 e D_Graduacao = 0) resulta em aumento de 82 % a chance de ser bom cliente.

2. Clientes com escolaridade "Ensino Médio", tudo o mais constante*, ($e^{-0.05}$= 0,951) tem (1-0,95)% = 5% menos chances de serem bons clientes em comparação com clientes de escolaridade "Pós Graduação".

3. Clientes com escolaridade "Graduação", tudo o mais constante*, ($e^{-0.01}$= 0,99) tem (1-0,9)% = 1% menos chances de serem bons clientes em comparação com clientes de escolaridade "Pós Graduação".

Preditiva.ai

Demonstração

Interpretação da base de diabetes com o uso da regressão

Regressão Logística

Risco de Crédito

Preditiva.ai

Hands on

Uma fintech de Crédito iniciou sua operação de concessão de empréstimo pessoal e acompanhou a performance de pagamento de 1.000 clientes após 1 ano. Com base dessa amostra, pediu para a área de Ciência de Dados desenvolver um modelo de concessão de crédito para conseguir aprovar mais contratos com a menor inadimplência possível.

Objetivo: Desenvolver um modelo para realizar a concessão de crédito de forma mais acurada.

Roteiro:

1. Faça uma análise exploratória da base “emprestimos.csv”;
2. Faça um ranking de IV’s. Quais variáveis são mais promissoras?
3. Desenvolva um modelo de Regressão Logística;
4. Calcule a Acurácia;
5. Interprete o modelo;
6. As variáveis mais importantes que o modelo trouxe batem com as identificadas pelo ranking de IV ?
7. **Bonus**: Discuta uma estratégia de uso desse modelo para melhorar a concessão de crédito na fintech.

Preditiva.ai