Regressão Logística

1

O que é a Regressão Logística?

É utilizada para prever a probabilidade de um evento binário ocorrer;

Segue a mesma lógica do modelo de regressão linear com a particularidade da variável alvo ser binária;

É uma técnica muito utilizada quando não se tem bases com dados não normais;

Não tem muitas exigências de pressupostos, ampliando sua aplicabilidade;

Há uma infinidade de eventos de interesse que podem ser modelados pela regressão logística.

2

Regressão Logística?

Deriva seu nome da transformação logit usada como variável dependente;

Um modelo é definido como logístico se a função segue a seguinte equação:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Razão de Chance = [p(Sucesso)]/[1 – P(Sucesso)]

Em que pi indica a probabilidade de ocorrência, $x1$, …, xn representam os vetores de variáveis explicativas (ou independentes) e β0 e βx indicam os coeficientes do modelo.

3

Regressão Logística?

Os coeficientes logísticos são difíceis de interpretar em sua forma original, pois são expressos em termos de logaritmos quando usamos a função logit;

É possível aplicar a transformação de anti-logaritmo por meio da exponenciação dos coeficientes originais, gerando a razão de desigualdades:

$$Razão\ de\ Desigualdades_i\ (odds) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}$$

Simplificando, tem-se:

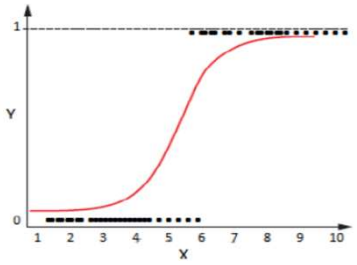
$$P(evento) = \frac{1}{1+e^{-(b_0+b_1 x_1+b_2 x_2+\cdots b_n x_n)}}$$

4

Regressão Logística?

Para cada observação, é previsto um valor de probabilidade entre 0 (0%) e 1 (100%);

•Os valores previstos para todos os valores da variável independente gera a curva logística:



5

Regressão Logística

Se a probabilidade prevista é maior do que 0,50 (ponto de corte de 50%), então a previsão é de que o resultado seja 1 (evento ocorreu);

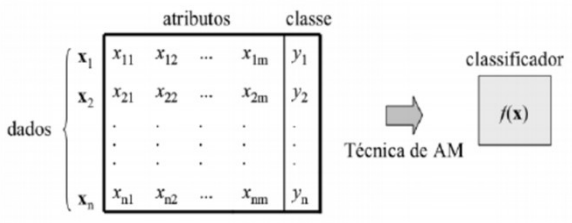
Caso a probabilidade prevista seja menor do que 0,50 (ponto de corte), então a previsão é de que o resultado seja 0 (evento não ocorreu);

Esse ponto de corte pode ser ajustado e faz parte das configurações de parâmetros para melhorar o modelo.

6

Modelo de Regressão Logística

A construção de um classificador por regressão logística pode ser representado de forma simplificada por:



Conjunto com n dados, onde cada observação xi possui m atributos e as variáveis yi representam as classes ou rótulos.

7

Avaliação do Modelo Logístico

Para entender os erros gerados por um classificador é possível visualizar por meio da construção da uma matriz de erros denominada matriz de confusão;

A partir da matriz é possível obter métricas de qualidade para a avaliação do desempenho de um classificador;

Resume o número de instâncias previstas corretas ou incorretas por um modelo de classificação.

8

Avaliação do Modelo Logístico

Representação da matriz de confusão:

Matriz de Confusão		Classe Atual	
		Negativa (-)	Positiva (+)
Classe Prevista	Negativa (-)	$f--(TN)$	$f+-(FN)$
	Positiva (+)	$f-+(FP)$	$f++(TP)$

9

Avaliação do Modelo Logístico

As seguintes terminologias são usadas para o entendimento da matriz de confusão:

Positivo verdadeiro (TP):é relacionado ao número de instâncias positivas previstas corretamente pelo classificador;

Negativo falso (FN): é o número de instâncias previstas erroneamente como negativos pelo classificador;

Positivo falso (FP):é o número de exemplos negativos previstos erroneamente como positivos pelo classificador;

Falso verdadeiro (TN):é o número de exemplos negativos previstos corretamente pelo classificador

10

Avaliação do Modelo Logístico

Uma das maneiras mais comuns de avaliar modelos é por meio da derivação de medidas que, tentam medir a qualidade do modelo;

Essas medidas geralmente podem ser obtidas a partir da matriz de confusão:

A **acurácia** é definida como sendo o número de instâncias corretas divididas pelo número total de instâncias, a saber:

$$Acurácia = \frac{TP + TN}{(TP + TN + FN + FP)}$$

11

Avaliação do Modelo Logístico

A medida da **precisão** determina o percentual de registros que são positivos no grupo que o classificador previu como classe positiva. Assim, tem-se:

$$Precisão = \frac{TP}{TP + FP}$$

Outra medida é denominada de **lembrança** (em inglês Recall) mede o percentual de instâncias positivas previstas corretamente pelo classificador.

$$Recall = \frac{TP}{TP + FN}$$

12

Avaliação do Modelo Logístico

Área da curva ROC (em inglês **ROC curve area)**:

Representação gráfica para descrever o desempenho de um sistema classificador binário;

É baseada na taxa de verdadeiros positivos TPR,e na taxa de falsos positivos FPR;

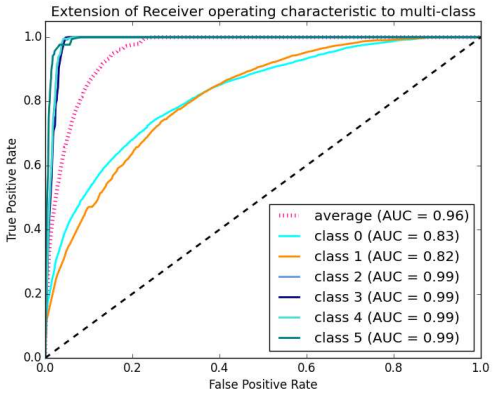
É muito utilizada quando queremos comparar o desempenho entre diversos classificadores;

Seus valores variam entre zero e um e é muito utilizada quando se tem classe desbalanceada.

13

Avaliação do Modelo Logístico

Gráfico da área da curva ROC (ROC curve area):



14

Avaliação do Modelo Logístico

Utilização de métricas conjuntamente:

É comum encontrar situações em que um modelo aparenta ser melhor que outro para algumas das métricas, mas pior com relação a outras;

Nesses casos, utilizar uma única medida pode dar a falsa impressão de que o desempenho pode ser avaliado utilizando-se apenas essa medida;

Para uma avaliação mais precisa é ideal que seja utilizado um conjunto de métricas levando em consideração o objetivo da pesquisa.

15

Exemplo Prático do uso de Regressão Logística

Suponha que uma concessionária esteja interessada em aprimorar sua política de vendas para minimizar perdas com clientes. Uma das medidas que se encontram em cogitação é exigir garantias adicionais que não possuem renda fixa, especialmente quando responsáveis pelas despesas da família. Por considerar que as exigências devem variar em função do risco de inadimplência associado à operação, o controller solicitou um estudo baseado no histórico dos últimos 12 meses. Para tanto, tomou-se uma amostra aleatória de 92 clientes, em relação aos quais foram consideradas as seguintes variáveis: Renda Mensal, Número de Dependentes e de o sujeito possui Vínculo Empregatício.

16

Exemplo Prático do uso de Regressão Logística

De acordo com o comportamento apresentado no período, cada um foi classificado como adimplente ou inadimplente.

Seguem os códigos das variáveis, a saber:

ST (Status) → Se Inadimplente, rotula-se 1 (um); se Adimplente, rotula-se 0 (zero);

R → Renda mensal média dos últimos 12 meses, em milhares de reais

ND → Número de Dependentes

VE → Atividade com Vínculo Empregatício, rotula-se com 1 (um); e sem Vínculo Empregatício, rotula-se 0 (zero)

17

Exemplo Prático do uso de Regressão Logística

Use a base de dados GarantiasdeVendas.csv

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	92	100,0
	Missing Cases	0	,0
	Total	92	100,0
Unselected Cases		0	,0
Total		92	100,0

a. If weight is in effect, see classification table for the total number of cases.

18

Exemplo Prático do uso de Regressão Logística

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
VE	0	42	1,000
	1	50	,000

19

Exemplo Prático do uso de Regressão Logística

Classification Table[a,b]

			Predicted		
			ST		
	Observed		0	1	Percentage Correct
Step 0	ST	0	51	0	100,0
		1	41	0	,0
	Overall Percentage				55,4

a. Constant is included in the model.

b. The cut value is ,500

20

Exemplo Prático do uso de Regressão Logística

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	50,307[a]	,563	,754

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,169	8	,417

21

Exemplo Prático do uso de Regressão Logística

Classification Table[a]

Observed			Predicted ST 0	Predicted ST 1	Percentage Correct
Step 1	ST	0	45	6	88,2
		1	4	37	90,2
	Overall Percentage				89,1

a. The cut value is ,500

22

Exemplo Prático do uso de Regressão Logística

Acurácia

$$Acurácia = \frac{TP + TN}{(TP + TN + FN + FP)}$$

$$Acurácia = \frac{45+37}{(45+37+4+6)} = 0{,}8913 = 89{,}13\%$$

23

Exemplo Prático do uso de Regressão Logística

Precisão

$$Precisão = \frac{TP}{TP + FP}$$

$$Precisão = \frac{45}{45+6} = 0{,}8824 = 88{,}24\%$$

24

Exemplo Prático do uso de Regressão Logística

Lembrança ou Recall

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{45}{45 + 4} = 0{,}9184 = 91{,}84\%$$

25

Exemplo Prático do uso de Regressão Logística

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	R	-1,882	,489	14,845	1	,000	,152	,058	,397
	ND	,860	,386	4,965	1	,026	2,362	1,109	5,031
	VE(1)	-2,822	,852	10,969	1	,001	,059	,011	,316
	Constant	4,300	1,489	8,341	1	,004	73,680		

a. Variable(s) entered on step 1: R, ND, VE.

$$P(evento) = \frac{1}{1 + 2{,}7182^{-(4{,}3 + (-1{,}882 Renda) + (0{,}86 Dependentes) + (-2{,}822 Vínculo))}}$$

26

Exemplo Prático do uso de Regressão Logística

2,7182	**4,3**	**-1,882**	**0,86**	**-2,822**		
e		**x1=renda**	**x2=numdep**	**x3=vincempre**		
		4	3	0		
	4,3	-7,528	2,580	0,000	0,648	1,912
	0,648					
	1,91167628					%
	34,34%				0,343445	34,34

2,7182	**4,3**	**-1,882**	**0,86**	**-2,822**		
e		**x1=renda**	**x2=numdep**	**x3=vincempre**		
		2	4	1		
	4,3	-3,764	3,440	-2,822	-1,154	0,315
	-1,154					
	0,31538371					%
	76,02%				0,760234	76,02

27