Regressão Logística Multinomial no R

Letícia Thomaz

Formada em Estatística pelo IME-USP

Pós Graduação em Análise de Big Data pela FIA

Atuária Sr. na SulAmérica

R-Ladies São Paulo

https://www.linkedin.com/in/leticiathomaz/

https://github.com/lodthomaz/

Agenda

- Modelos de Aprendizado
- Modelos de Classificação
- Regressão Logística
- Regressão Logística Multinomial
- Dataset Iris **(R)**
- Dataset Titanic **(R)**
- Avaliação do Modelo

1.

Modelos de Aprendizado

Contextualizando ...

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

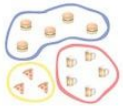
Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»



DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Find hidden dependencies

Identify sequences

ASSOCIATION

«Find what clothes I often wear together»

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

Regressão Logística

UNSUPERVISED

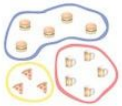
Divide by similarity

Identify sequences

Find hidden dependencies

CLUSTERING

«Split up similar clothing into stacks»



ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

2.
Modelos de Classificação

O que são

- Modelos utilizados quando a variável de interesse é **categórica**
- Objetivo é estimar um “classificador” com base nos dados

Exemplos

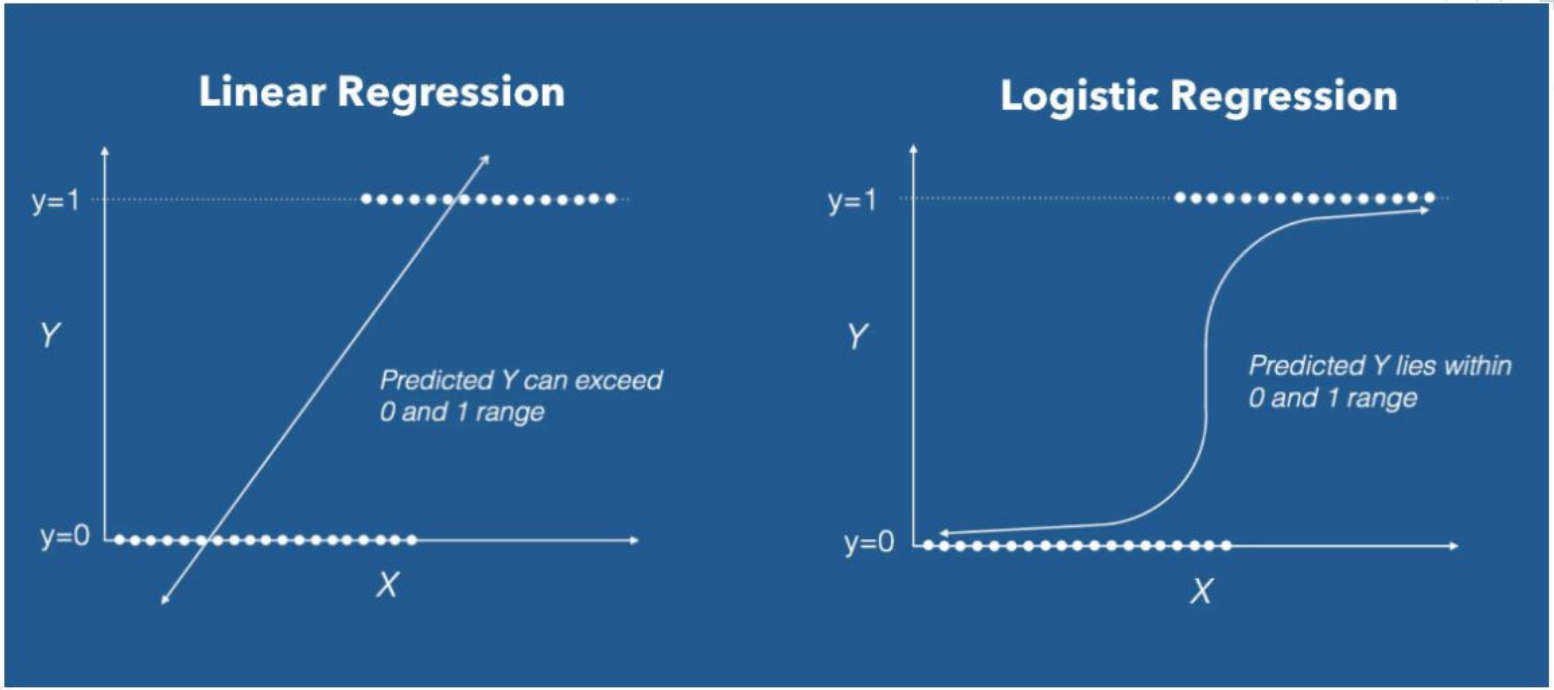
- Morte ou Sobrevivência no Titanic (clássico!!!)
- Fraude em transação de cartão de crédito
- Classificar o tipo de acidente nas rodovias brasileiras (sem vítimas, com vítimas feridas, com vítimas fatais)
- Filtrar e-mail: spam ou não spam
- Condição médica de um paciente no PS (infarto, overdose, ataque eplético)

3.
Regressão Logística

Regressão Logística

- Quando queremos predizer uma **categoria**
- Ao invés de modelar a variável resposta (Y) diretamente, a regressão logística vai modelar a **probabilidade** de Y pertencer a uma determinada classe
- Valores estão contidos entre 0 e 1
- Utiliza a função de ligação logit

Regressão Logística



https://www.datacamp.com/community/tutorials/logistic-regression-R

Função Logística

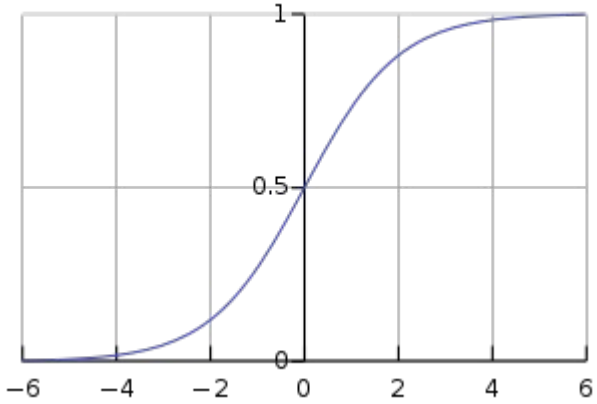
Quando vamos estimar uma probabilidade precisamos limitar o valor entre 0 e 1

$$p(x) = Pr(Y = 1 \mid X)$$

Difícil interpretar os coeficientes

$$p(x) = e^{g(x)} / (1+e^{g(x)})$$

$$g(x) = \beta 0 + \beta 1 X1 + \dots \beta n Xn$$



Logit

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

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$$

Logit

4.

Regressão Logística Multinomial

Regressão Logística Multinomial

- Extensão do modelo de regressão logística binária
- Variável resposta tem mais de duas categorias
- Ao invés de estimar um “classificador”, irá estimar k-1, sendo k o número de categorias da variável
- Precisa passar uma categoria de referência da variável resposta

5.

Avaliação do Modelo

Métricas

Acurácia: indica a performance geral do modelo. De todas as classificações, quantas o modelo classificou corretamente

Precisão (Pos Pred Value): dentre todas as classificações da classe positivo que o modelo fez, quantas estão corretas

Recall (Sensibilidade): dentre todas as situações de classe positivo como valor esperado, quantas estão corretas

Métricas

Acurácia é uma boa indicação geral de como o modelo performou. Para bases muito desbalanceadas, não é indicado olhar só para essa medida.

Precisão pode ser usada em uma situação em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos.

Recall pode ser usada em uma situação em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos.

6.

Exemplos R

https://github.com/lodthomaz/MultinomialLogReg_EstatiDados

Iris

- Conjunto de dados de 3 espécies da flor *Iris*
- Base de dados contém 50 amostras de cada uma delas: setosa, virginica e versicolor
- 4 variáveis: comprimento e largura das sépalas, comprimento e largura das pétalas

É possível estimar um classificador que classifique a espécie corretamente com base nas outras variáveis?

Resultados Iris

```
> confusionMatrix(df_teste$Species, df_teste$fitted_values)
Confusion Matrix and Statistics

            Reference
Prediction   virginica setosa versicolor
  virginica          8      0          2
  setosa             0     10          0
  versicolor         1      0          9

Overall Statistics

               Accuracy : 0.9
                 95% CI : (0.7347, 0.9789)
    No Information Rate : 0.3667
    P-Value [Acc > NIR] : 1.888e-09

                  Kappa : 0.85

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: virginica Class: setosa Class: versicolor
Sensitivity                    0.8889       1.0000            0.8182
Specificity                    0.9048       1.0000            0.9474
Pos Pred Value                 0.8000       1.0000            0.9000
Neg Pred Value                 0.9500       1.0000            0.9000
Prevalence                     0.3000       0.3333            0.3667
Detection Rate                 0.2667       0.3333            0.3000
Detection Prevalence           0.3333       0.3333            0.3333
Balanced Accuracy              0.8968       1.0000            0.8828
```

Titanic

Base de dados com informações sobre os passageiros do titanic. Muito utilizada em competições de ML no ***kaggle*** com o objetivo de detectar os passageiros que sobreviveram.

Mas vamos usar essa base para tentar prever à qual classe aquele passageiro pertencia: 1ª, 2ª ou 3ª classe.

Dados: https://www.kaggle.com/c/titanic/data

Resultados Titanic

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 35  6  1
         2  3 14 19
         3  1  1 96

Overall Statistics

               Accuracy : 0.8239
                 95% CI : (0.7594, 0.8771)
    No Information Rate : 0.6591
    P-Value [Acc > NIR] : 9.051e-07

                  Kappa : 0.683

 Mcnemar's Test P-Value : 0.0006429

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8974   0.66667   0.8276
Specificity            0.9489   0.85806   0.9667
Pos Pred Value         0.8333   0.38889   0.9796
Neg Pred Value         0.9701   0.95000   0.7436
Prevalence             0.2216   0.11932   0.6591
Detection Rate         0.1989   0.07955   0.5455
Detection Prevalence   0.2386   0.20455   0.5568
Balanced Accuracy      0.9232   0.76237   0.8971
```

Resultados Titanic

```
> # Resumo do Modelo
> summary(modelo)
Call:
multinom(formula = Pclass ~ ., data = df_treino)

Coefficients:
  (Intercept)  Survived1    Sexmale         Age    SibSp    Parch  EmbarkedC EmbarkedQ
2    4.599933 -0.7223733 -0.8227987 -0.07499522 1.743236 0.729511 0.05045065  2.778236
3    7.389673 -1.6436351 -0.4309406 -0.09402902 2.620887 1.416826 0.77224662  4.996631
  EmbarkedS        Fare
2  1.771246 -0.09594606
3  1.620796 -0.22311060
```

Survived (3): -1,6436 é a estimativa do efeito no log do odds ratio para um aumento de uma unidade na variável survived.
Ou seja, caso o passageiro tenha sobrevivido ao Titanic, o log da chance dele pertencer à 3ª classe e não à 1ª é diminuído em 1,6436 unidades.

Resultados Titanic

```
> # Exponencial dos Coeficientes do Modelo
> exp(coef(modelo))
  (Intercept) Survived1   Sexmale       Age     SibSp     Parch EmbarkedC EmbarkedQ EmbarkedS
2    99.47767 0.4855984 0.4392008 0.9277479  5.715809 2.074066  1.051745  16.09062  5.878174
3  1619.17670 0.1932762 0.6498975 0.9102563 13.747918 4.124008  2.164624 147.91395  5.057113
       Fare
2 0.9085130
3 0.8000264
```

Survived (3): Se mantivermos todas as demais variáveis constantes, para um passageiro que sobreviveu, a chance dele estar na 3ª classe é 0,1933 vezes a chance dele estar na 1ª.

Obrigada!

Perguntas?

https://www.linkedin.com/in/leticiathomaz/

https://github.com/lodthomaz/

Slides Carnival

Free templates for all your presentation needs

For PowerPoint and Google Slides

100% free for personal or commercial use

Ready to use, professional and customizable

Blow your audience away with attractive visuals