Relatório da atividade 1

disciplina SIN5016 - Aprendizado de máquina

Bruno Kemmer - NUSP 5910474

April 2020

1 Introdução

Como solicitado foram utilizados três datasets: *Iris Data Set*, *Pima Indians Diabetes Database* e *Hepatitis Data Set*.

1.1 *Iris Data Set*

Data set com 3 diferentes classes: Iris Setosa, Iris Versicolour, Iris Virginica. E com os seguintes atributos:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm

Como consta na descrição da atividade, o dataset foi dividido em 2/3 para treino e 1/3 para teste. Também foi codificada as três classes em:

- "Iris-setosa": [1,0,0]
- "Iris-versicolor": [0,1,0]
- "Iris-virginica": [0,0,1]

1.2 *Pima Indians Diabetes Database*

O dataset consiste em diferentes variáveis preditoras médicas (independentes) e uma variável classe (dependente). As variávies independentes incluem: número de gravidezes que a paciente teve, seu BMI, nível de insulina, idade entre outros.

1. Pregnancies - Gravidezes - número de vezes em que a pessoa já engravidou
2. Glucose - Nível de glicose - concentração de de glicose no plasma em 2 horas em um teste de tolerância oral.
3. BloodPressure - Diastolic blood pressure (mm Hg) - Pressão sanguínea diastólica

4. SkinThickness - Triceps skin fold thickness (mm) - Grossura da pele via o quanto é possível dobrar do triceps
5. Insulin - Nível de insulina 2-Hour serum insulin (mu U/ml) -
6. BMI - Body mass index - Índice de massa corporal
7. Age - Idade
8. Outcome - Variável classe (0 ou 1) - 268 de 768 exemplos são da classe 1, os outros são 0.

Como o dataset só contém uma classe (binário) a codificação da variável dependente foi: 0: -1 e 1: +1. Inicialmente foi utilizado o critério de divisão aleatório, porém como foi disponibilizado o dataset dividido, esta foi utilizada para treino e teste.

1.3 *Hepatitis Data Set*

- Número de instâncias: 155
- Tem valores faltantes: Sim
- Número de atributos: 19

Contém os seguintes atributos:

1. Class: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STEROID: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY: no, yes

Dataset binário e sua codificação foi feita como: 1: -1 e 2: +1.

Como nesse dataset tem dados faltantes, estes foram completados com a média de suas variáveis (colunas).

2 Normalizações

Para todos os datasets foi primeiramente treinado sem nenhuma normalização, normalizando todas as colunas para *z_score* - uma distribuição normal padrão (média 0 e desvio padrão 1) e também todas as colunas para *minmax - mínimo e máximo*. Foi tomado o cuidado de utilizar somente os dados de treinamento, e aplicar essas mesmas agregações (do dataset de treinamento) nos dados de teste no momento de inferência.

3 Regularização

Para o caso da regressão linear também foi implementada a versão com o decaimento dos pesos, em o termo de regularização lambda, e tendo a seguinte função para o cálculo dos pesos w:

$$w = (X^T X + \lambda I)^{-1} X^T y \tag{1}$$

Não tendo a quantidade de exemplos de treinamento N multiplicando o termo de regularização *lambda*.

4 Resultados

4.1 *Iris Data Set*

Utilizando o método de regressão linear foi obtida uma acurácia de 0,8444 nos dados de teste. Esse mesmo valor não mudou com os tipos de normalização (*z_score* e *minmax*).

Como podemos ver na Figura 1 existe um valor de *lambda* entre 0 e 1 em que o valor da acurácia de teste aumenta porém para todos os outros ela é menor. Seria interessante fazer uma busca para otimizar o parâmetro utilizando cross-validation no dataset.

Outro ponto interessante que podemos observar na Figura 2 é que para valores de *lambda* grandes o modelo começa a ter um decaimento de performance considerável, o que mostra que os valores de w estão ficando tão pequenos que o modelo não está conseguindo capturar os padrões contidos nos dados de treino.

Já no caso utilizando a regressão logística, com a normalização *z_score* foi obtida a acurácia de 0,9778 e com *minmax* 0,9333. Utilizando uma taxa de aprendizagem de 0,1 e um limite de iterações de 1.000.

4.2 *Pima Indians Diabetes Database*

Utilizando o método de regressão linear foi obtida uma acurácia de 0,7565 nos dados de teste. A mesma não se alterou ao utilizar a normalização via *z_score* porém caiu para 0,6522 ao utilizar *minmax*.

Ao utilizar a regressão logística, sem normalização a função de erro não convergiu, aplicando a normalização *z_score* foi obtida uma acurácia de 0,7478 e ao aplicar *minmax* a acurácia obtida foi de 0,7522. Em todos esses testes foi utilizada uma taxa de aprendizagem de 0,5 e um limite de iterações de 1.000.

4.3 *Hepatitis Data Set*

Utilizando o método de regressão linear foi obtida uma acurácia de 0,8298 nos dados de teste. A mesma não se alterou ao utilizar a normalização via *z_score* porém caiu para 0,1915 ao utilizar *minmax*.

Ao utilizar a regressão logística novamente não convergiu nos dados não normalizados, utilizando a normalização *z_score* foi obtida uma acurácia de 0,7872 e usando *minmax* 0,8298. Em todos esses testes foi utilizada uma taxa de aprendizagem de 0,1 e um limite de iterações de 2.000.

4.4 Resumo

Dataset	Regressor linear			Regressor logístico	
	Sem normalização	*z_score*	*minmax*	*z_score*	*minmax*
Iris	0,8444	0,8444	0,8444	0,9778	0,9333
Pima Indians Diabetes	0,7565	0,7565	0,6522	0,7478	0,7522
Hepatitis	0,8298	0,8298	0,1915	0,7872	0,8298

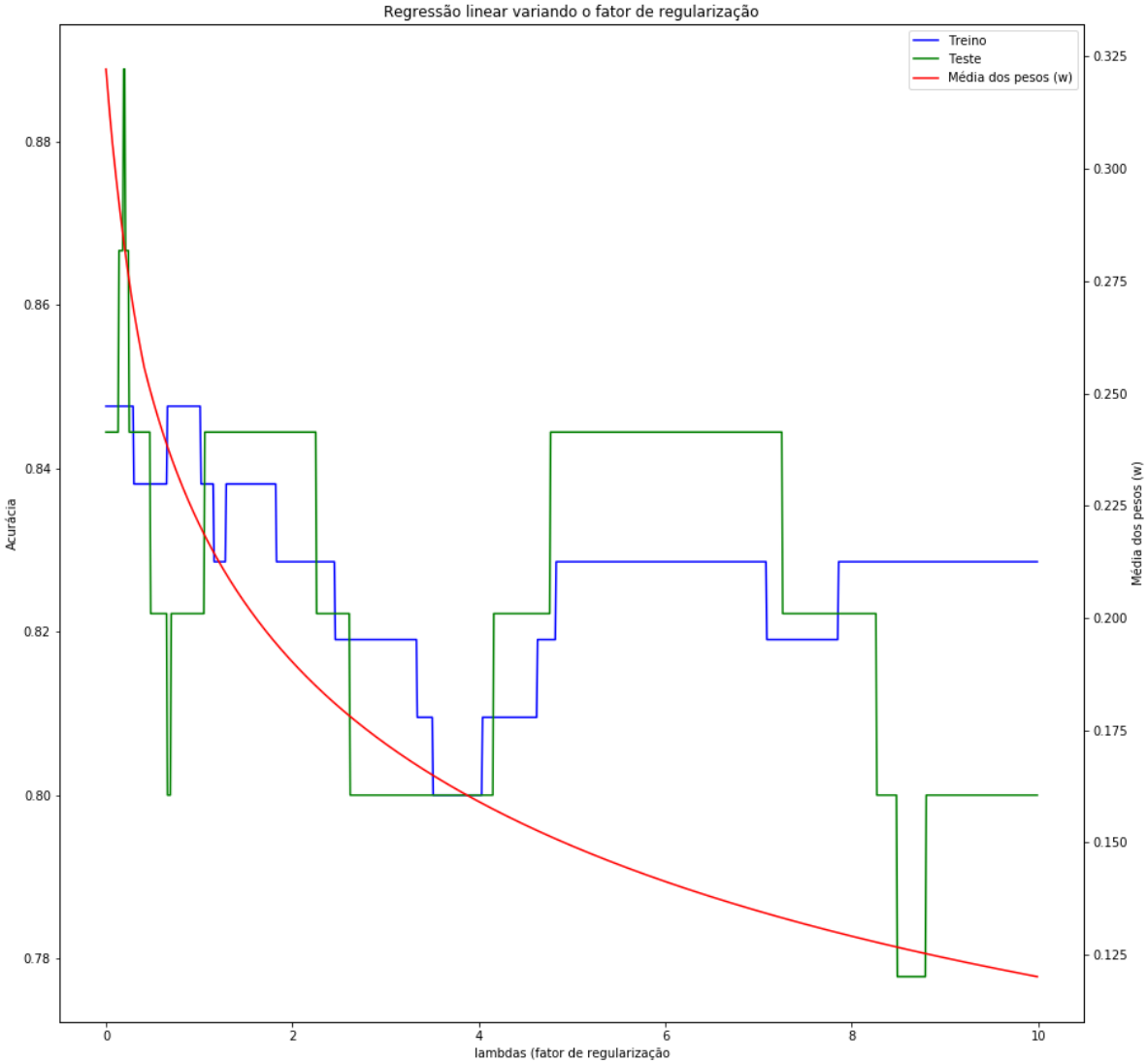

Figura 1: Acurácia de treino e teste (eixo-y esquerda) e média dos valores absolutos dos pesos w (eixo-y direita) ao variar os valores de lambda.

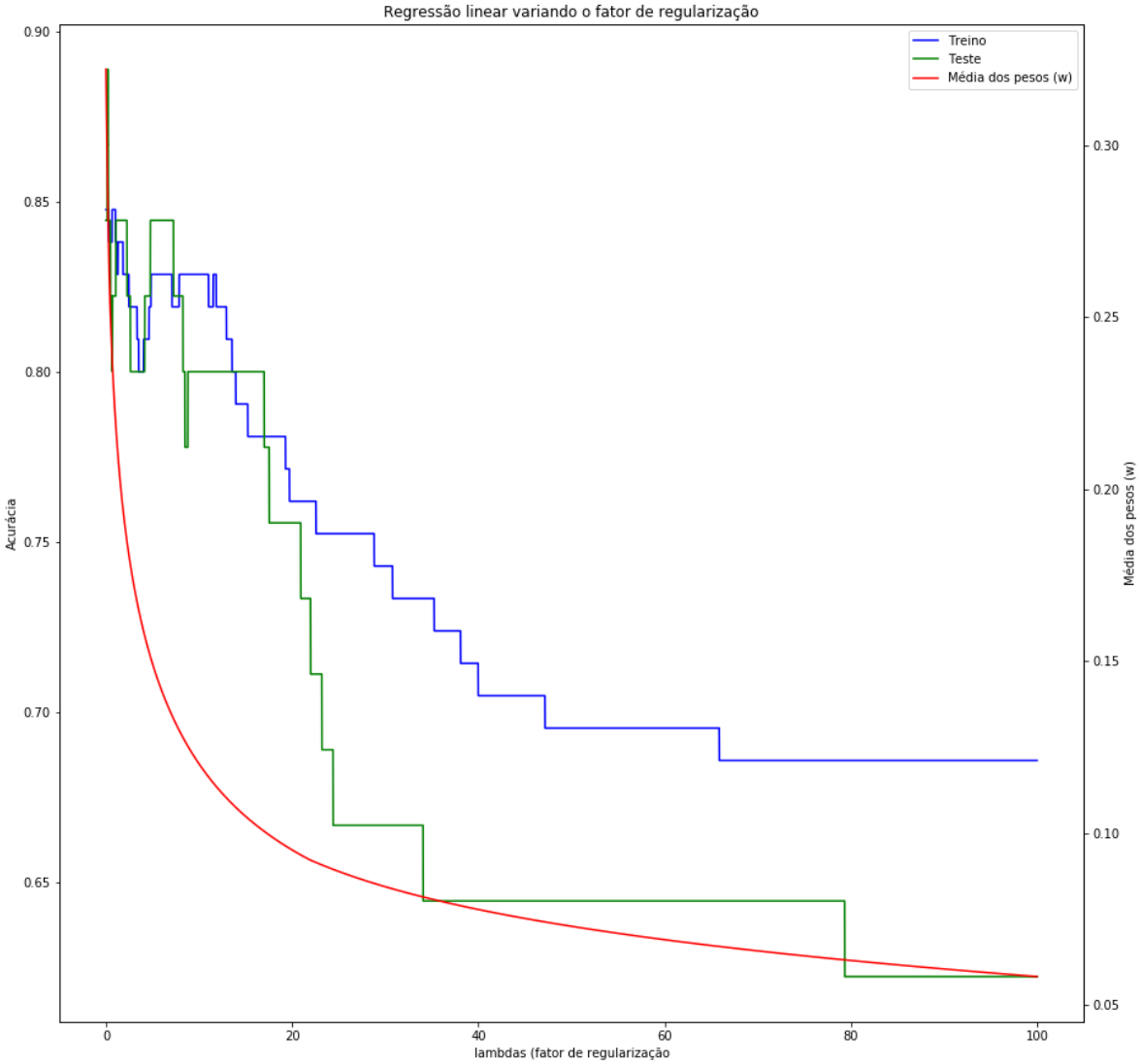

Figura 2: Mesmo gráfico mas ao variar os valores de lambda de 0 até 100.

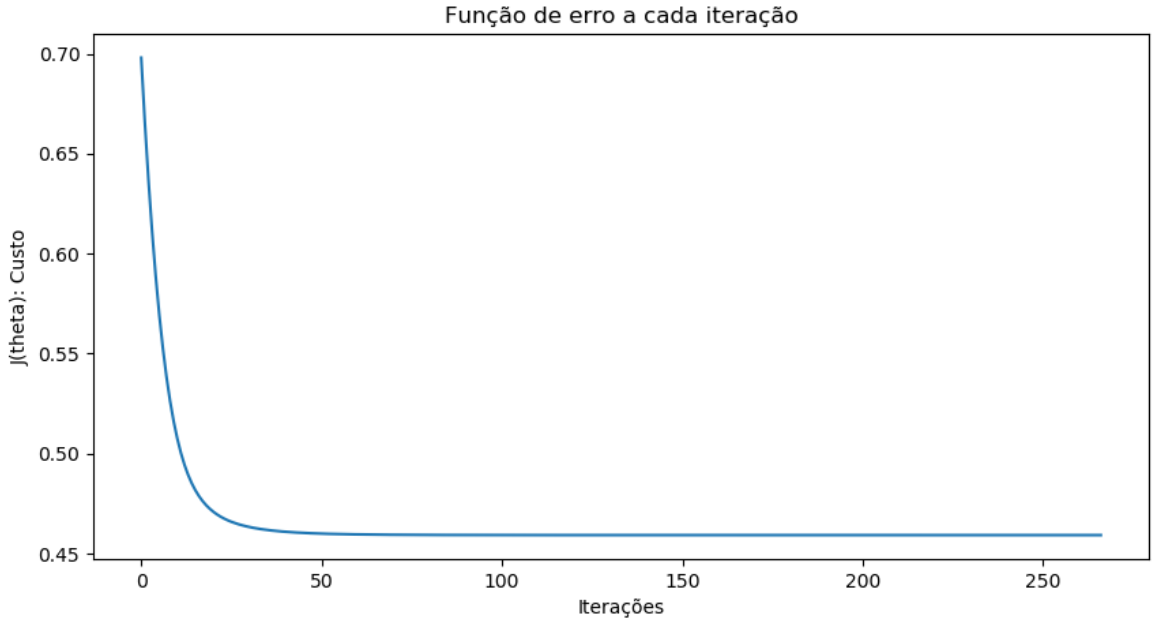

Figura 3: Função de custo $J(\theta)$ da regressão logística com valores normalizados utilizando *z_score* e $\lambda = 0.5$.

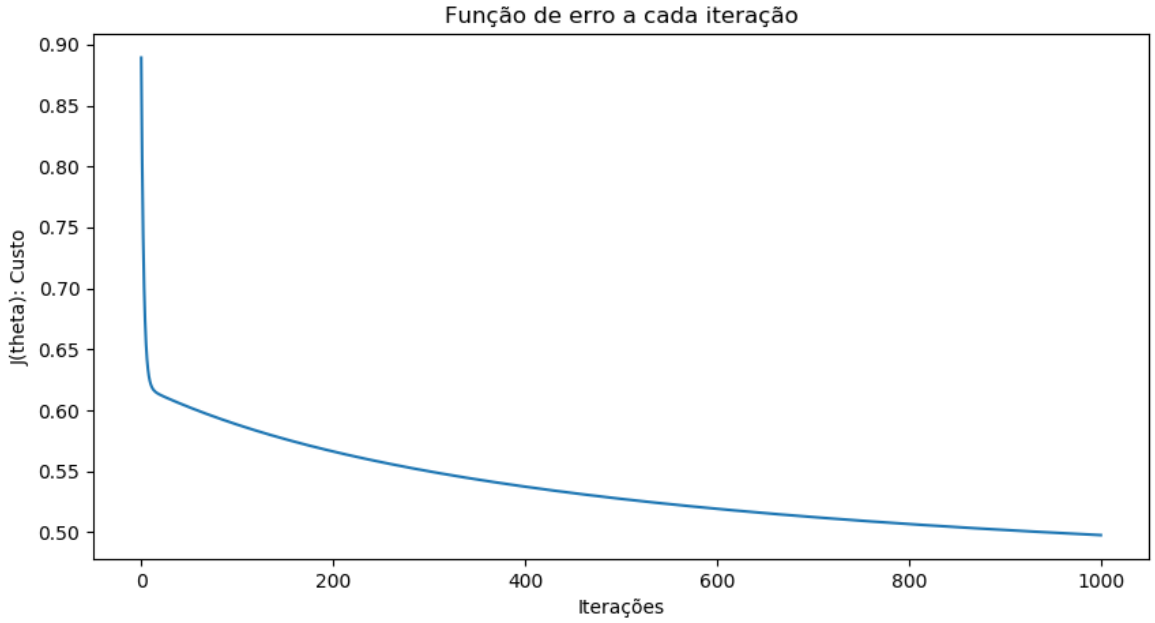

Figura 4: Função de custo $J(\theta)$ da regressão logística com valores normalizados utilizando *minmax* e $\lambda = 0.1$.

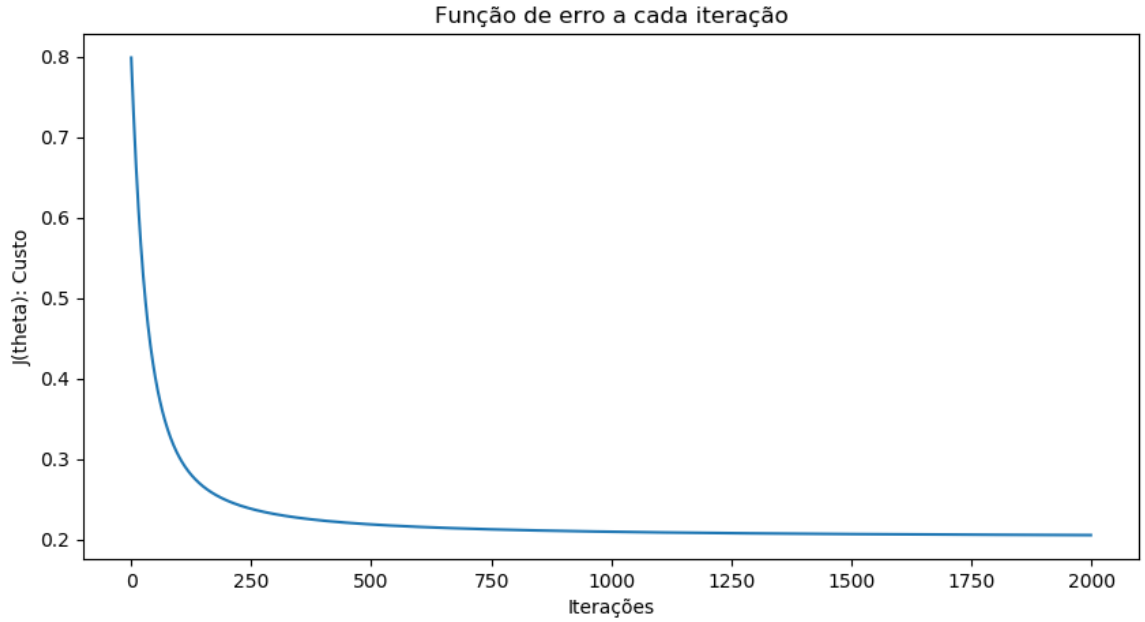

Figura 5: Função de custo $J(\theta)$ da regressão logística com valores normalizados utilizando *z_score* e $\lambda = 0.1$.

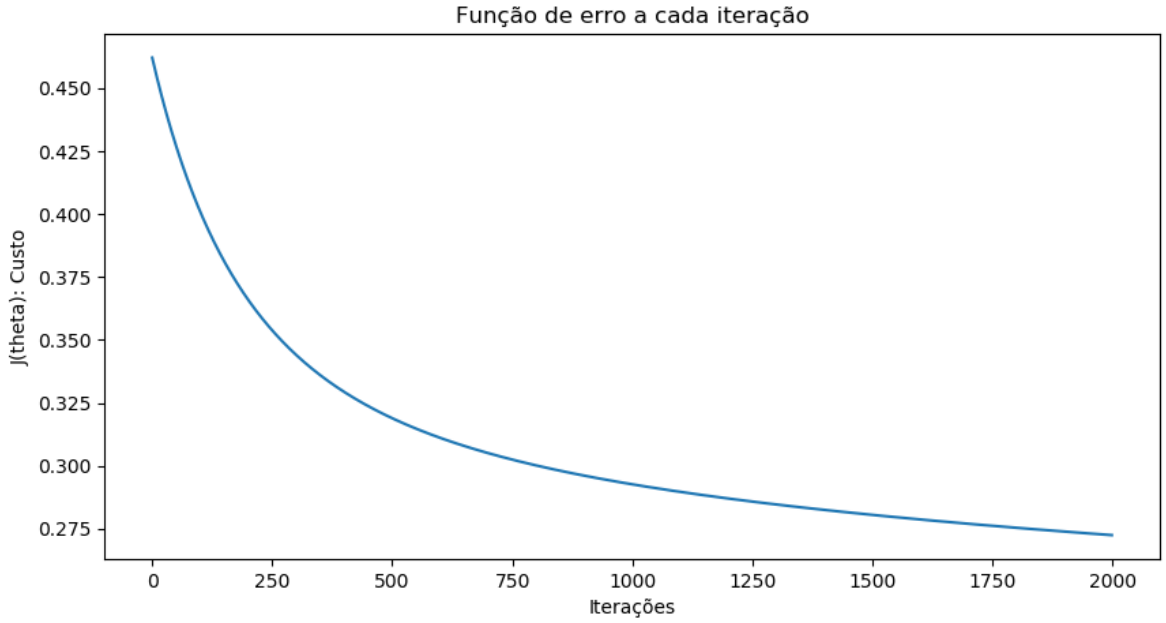

Figura 6: Função de custo $J(\theta)$ da regressão logística com valores normalizados utilizando *minmax* e $\lambda = 0.1$.