Regressão Linear

Prática 04: Validação Cruzada

Profª Deborah Magalhães
Monitor: Davi Luis de Oliveira

Universidade Federal do Piauí

Olá!

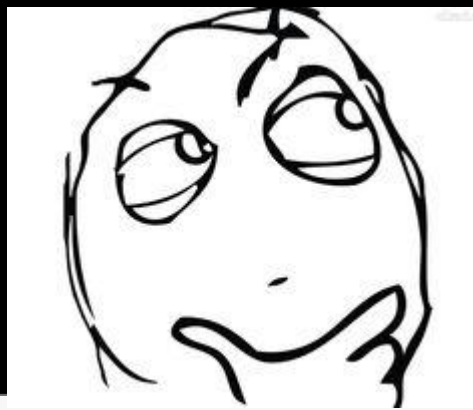
Curso: Bacharelado em Sistema de Informação

Disciplina: Sistemas Inteligentes

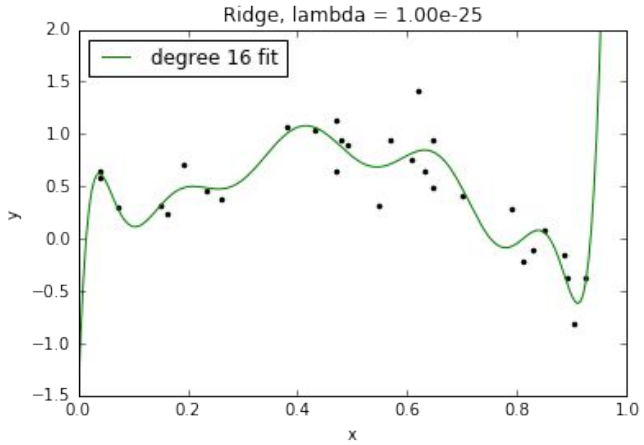
- **Visão Geral: Validação Cruzada**

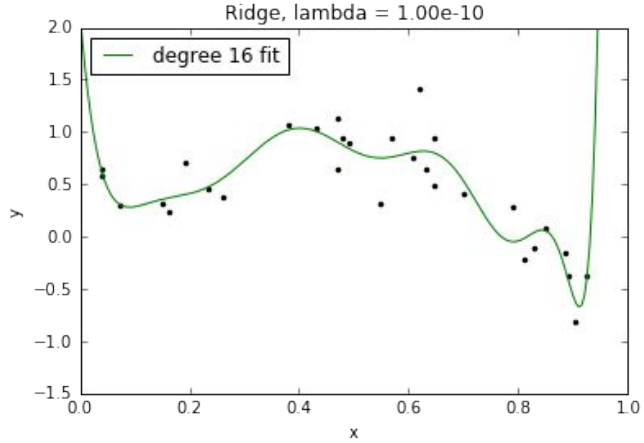
Você pode me encontrar em **deborah.vm@gmail.com**
(Dúvidas e sugestões serão bem-vindas =D)

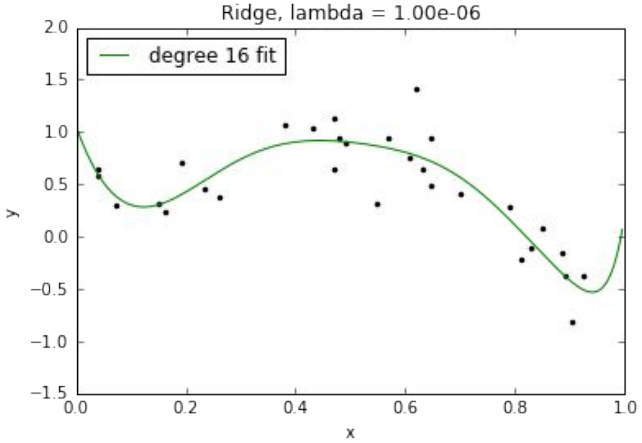
Na prática passada, qual valor de λ ofereceu o melhor fit aos dados?



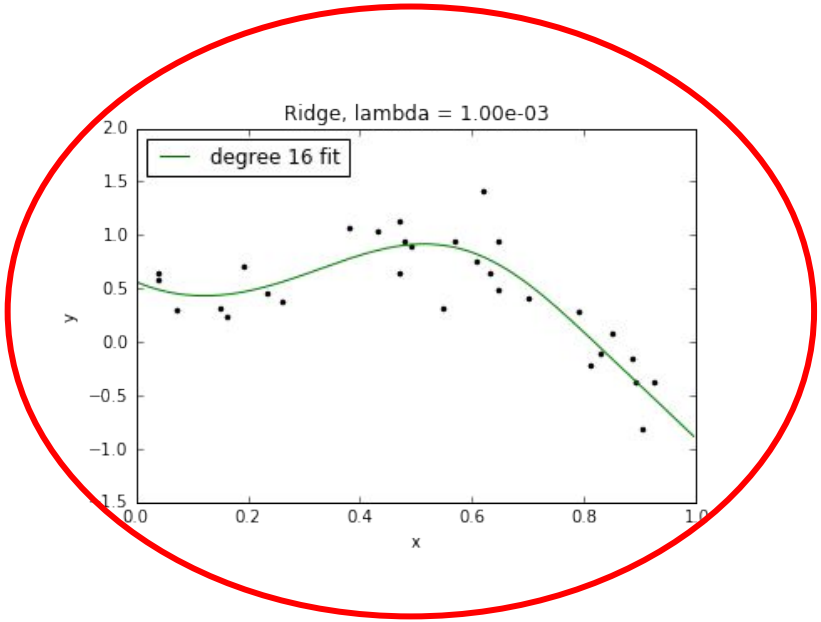
Regressão de Cume

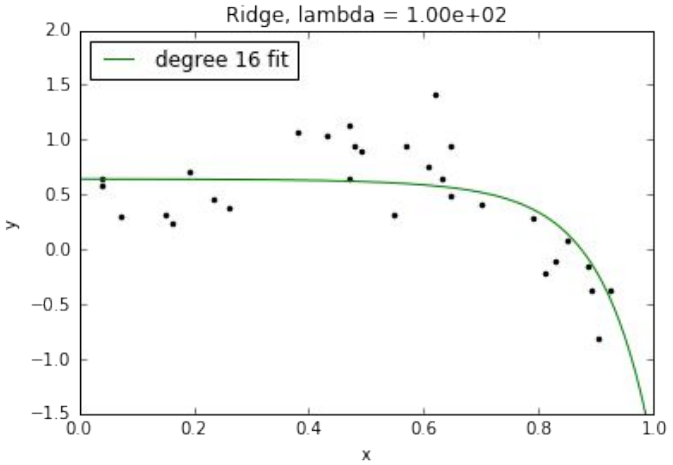




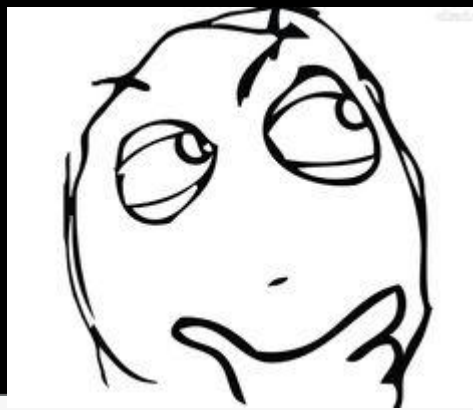


Regressão de Cume





Tem algum jeito de escolher os valores de lambda automaticamente?



Divisão do conjunto de dados

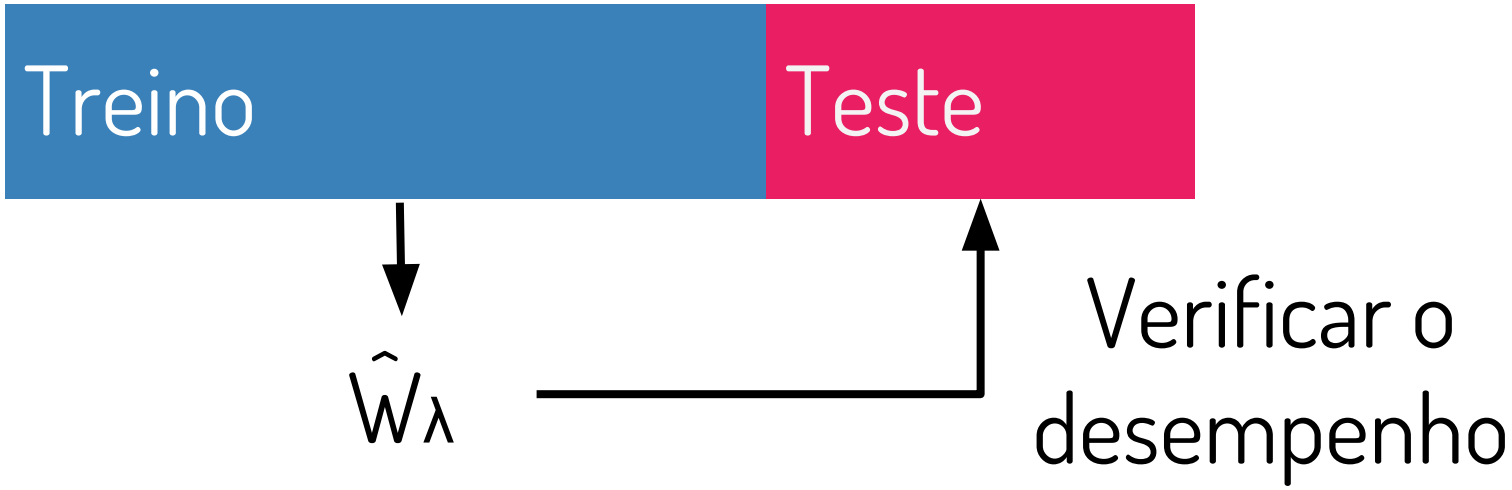
- A divisão do conjunto de dados pode ser utilizada:
  - Seleção do modelo
  - Avaliação do modelo

Divisão do conjunto de dados

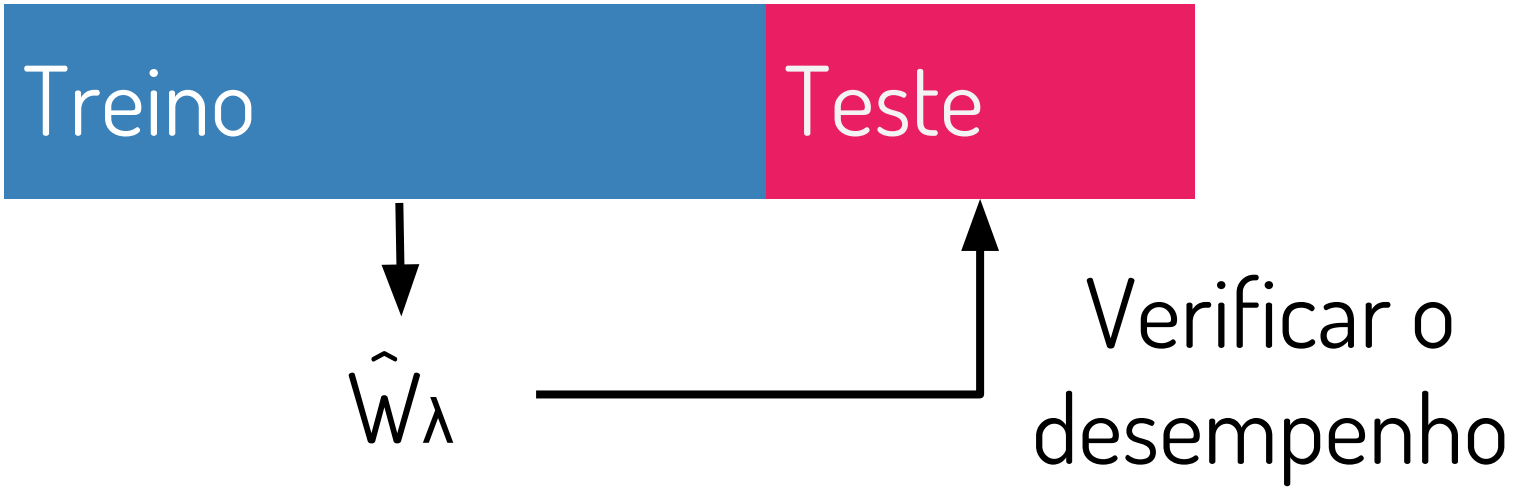
- O parâmetro λ serve para controlar o modelo:
  - Diferentes complexidades
  - Diferentes magnitudes dos coeficientes

Divisão do conjunto de dados

- Anteriormente, falamos da divisão treino/teste para solucionar a questão do overfitting



Divisão do conjunto de dados



* Repetir o processo para diferentes valores de λ e escolher os coeficientes que minimizam o erro de teste

Divisão do conjunto de dados

Se os dados de teste não representam todas as **variações** do mundo, considerar que o **erro de teste** é uma **boa medida** para selecionar o modelo é ser bastante otimista!

Ingênuo

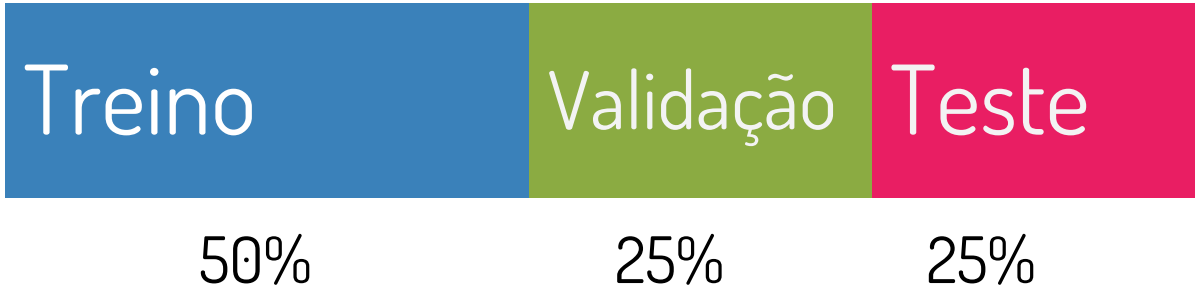
Divisão do conjunto de dados

1. Encontrar os coeficientes ($\hat{w}$) que se ajustam ao conjunto de teste
2. Selecionar o λ^* que minimiza o erro de validação
3. Em seguida, avaliar o desempenho do modelo para o conjunto de teste, isso vai nos oferecer o erro de generalização

Divisão do conjunto de dados

Treino	Validação	Teste

Divisão do conjunto de dados

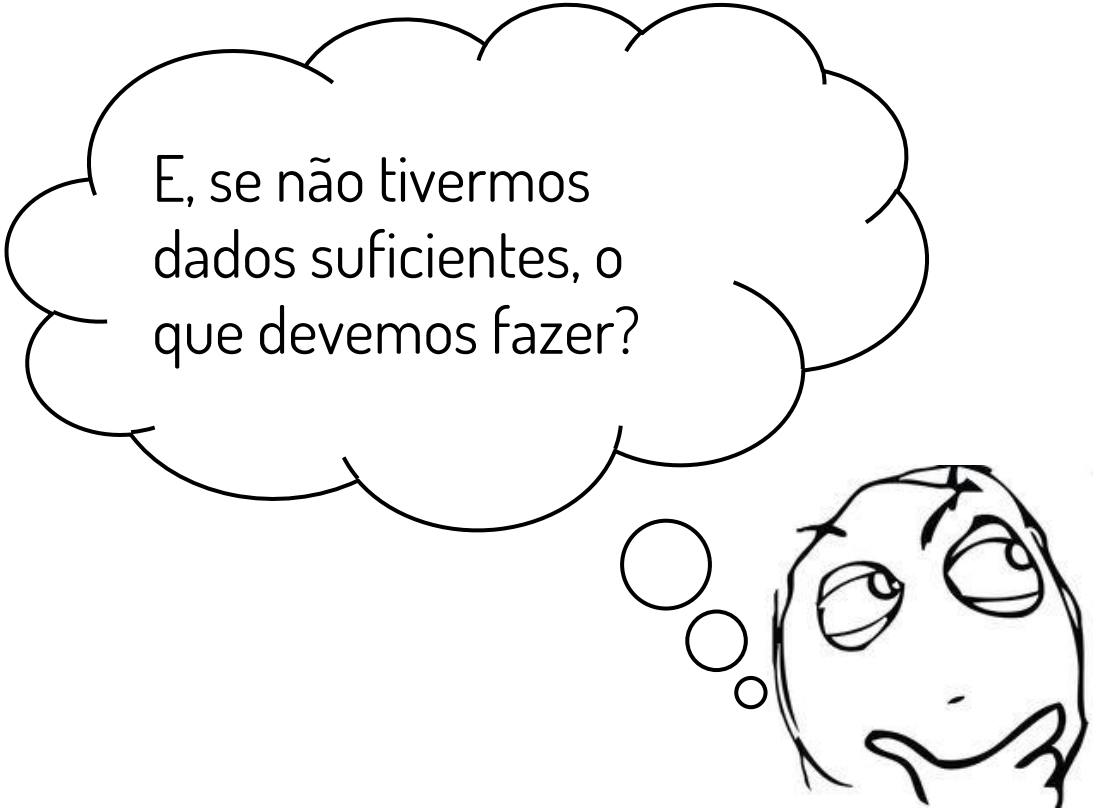


* Não existem regras!

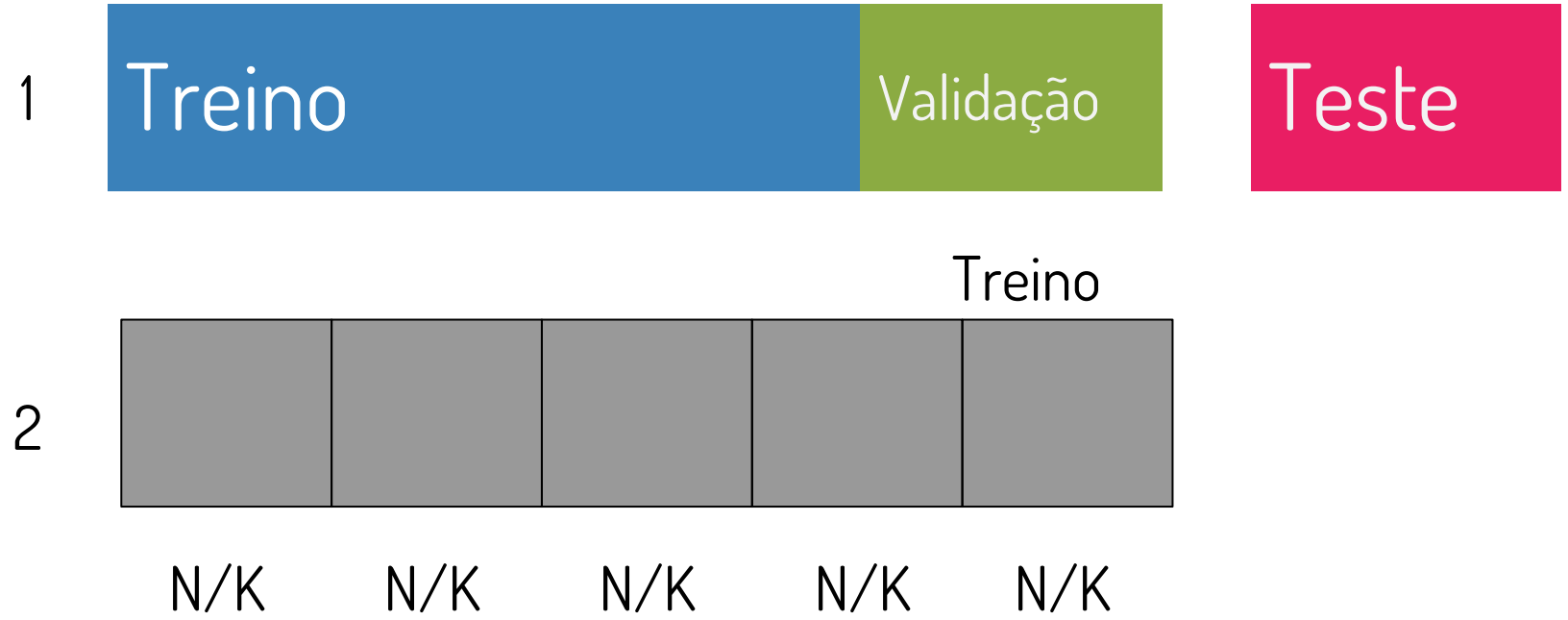
Divisão do conjunto de dados

- Problema: estamos assumindo que existem dados suficientes para realizar essa divisão. **Isso é uma suposição muito forte!!**

Divisão do conjunto de dados

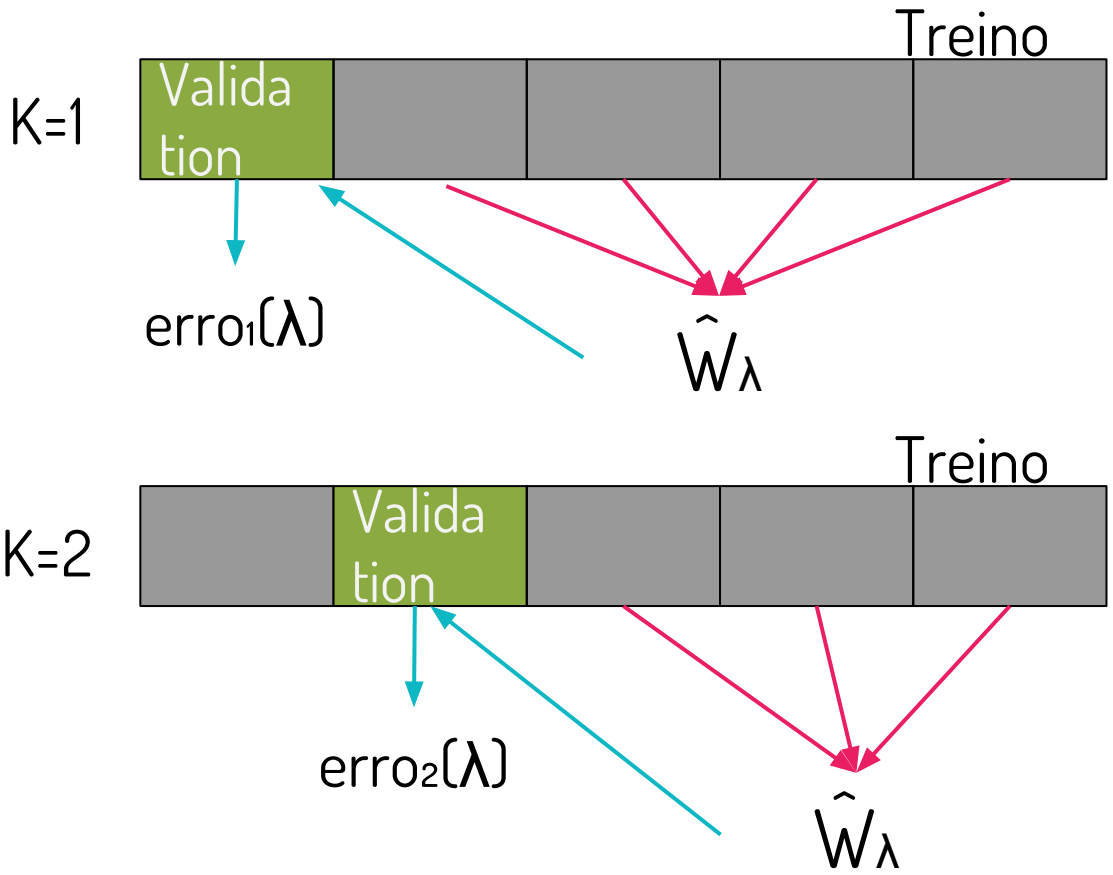


Validação Cruzada



Validação Cruzada

Validação Cruzada



Validação Cruzada

1. Repito o processo até K=5
2. Associado a cada K, eu tenho um erro de validação que será utilizado para calcular o erro de validação cruzada:

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} error_k(\lambda)$$

3. Esse procedimento é repetido para cada λ e, eu seleciono aquele que minimiza CV(λ)

Validação cruzada

Resposta: A melhor configuração é K=N, ou seja, os blocos possuem tamanho igual a 1. Essa configuração específica recebe o nome de "leave one out"

Problema: computacionalmente custosa porque N fits são calculados para cada λ

Passo 11: Definir a função que realiza a validação cruzada "leave one out"

```
def loo(data, deg, l2_penalty_values):

    data = polynomial_features(data, deg)

    num_folds = len(data)
    folds = graphlab.cross_validation.KFold(data,num_folds)

    l2_penalty_mse = []
    min_mse = None
    best_l2_penalty = None
```

Passo 11: Definir a função que realiza a validação cruzada "leave one out"

```
for l2_penalty in l2_penalty_values:
        next_mse = 0.0
        for train_set, validation_set in folds:
        model = graphlab.linear_regression.create(train_set,target='Y',
                                l2_penalty=l2_penalty,
                                validation_set=None,verbose=False)

        y_test_predicted = model.predict(validation_set)
        next_mse += ((y_test_predicted-validation_set['Y'])**2).sum()

        next_mse = next_mse/num_folds
        l2_penalty_mse.append(next_mse)
        if min_mse is None or next_mse < min_mse:
        min_mse = next_mse
        best_l2_penalty = l2_penalty
```

Passo 11: Definir a função que realiza a validação cruzada "leave one out"

```
    return l2_penalty_mse,best_l2_penalty
```

Passo 12: Gere diferentes valores de lambda

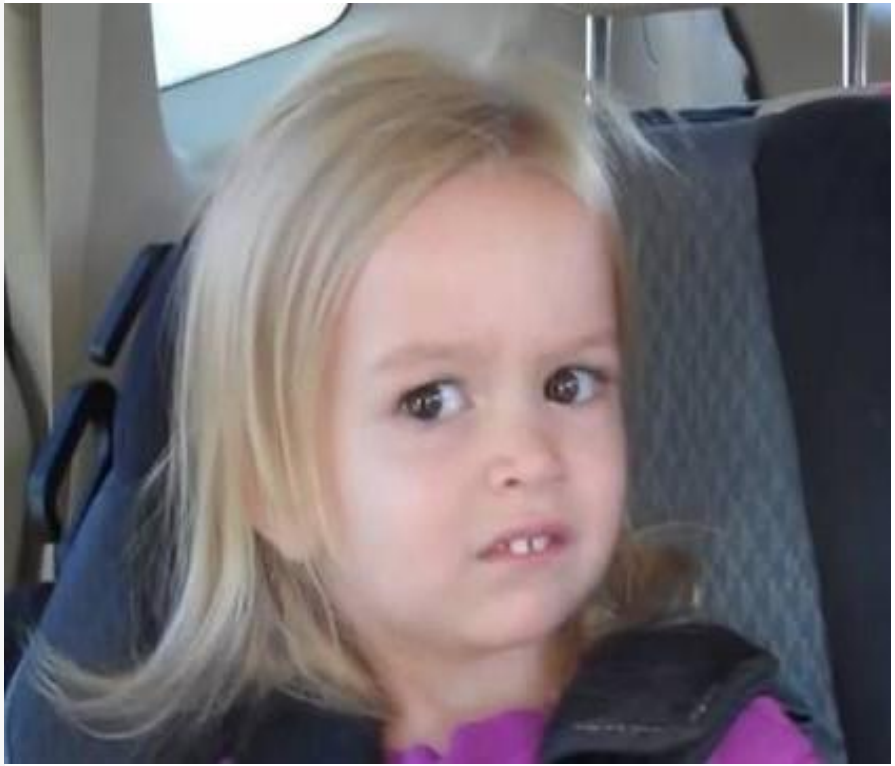
```
l2_penalty_values = numpy.logspace(-4, 10, num=10)
```

Passo 13: gere o modelo de regressão de cume (grau 16) considerando o melhor lambda

```
l2_penalty_mse,best_l2_penalty = loo(data, 16, l2_penalty_values)
```

Atividade Computacional (1.5 pontos)

1. Extrapole a função loo para K=5 e k=10
2. Imprima os coeficientes e plot a predição do modelo de regressão de cume de grau 16 com o parâmetro lambda definido pela função loo, K=5 e K=10



Dúvidas?Sugestões? Inquietações? Aconselhamentos?

- Desabafe em: **deborah.vm@gmail.com**