Introdução à Ciência de Dados

Prof. Francisco Rodrigues

Seleção de modelos e regularização

1 - Considere a base BostonHousing. Compare o coeficiente R2 obtido através de regressão linear múltipla, Lasso e ridge regression. Para os métodos Lasso e ridge regression, faça um gráfico de $\alpha \times R^2$ conforme feito no notebook da aula.

2 - Determine as variáveis que mais influenciam o preço de imóveis em Boston usando Lasso.

3 - Considere os dados gerados com o código a seguir. Usando regularização, ajuste o grau do polinômio que define o modelo mais adequado.

```
import numpy as np
from matplotlib import pyplot as plt
np.random.seed(10)

#função para gerar os dados
def function(x):
        y = x**4 + x**9
return y

# training set
N_train = 20
sigma = 0.2
x_train= np.linspace(0, 1,N_train)
y_train = function(x_train) + np.random.normal(0,sigma, N_train)
x_train = x_train.reshape(len(x_train), 1)

fig = plt.figure(figsize=(8, 4))
plt.scatter(x_train, y_train, facecolor="blue", edgecolor="b", s=100, label="training data")

# test set
N_test = 20
x_test=np.linspace(0, 1,N_test)
y_test = function(x_test) +  np.random.normal(0,sigma, N_test)
x_test = x_test.reshape(len(x_test), 1)

# Curva teorica
xt = np.linspace(0,1,100)
yt = function(xt)
plt.plot(xt,yt, '-r',  label="Theoretical curve")
plt.legend(fontsize=15)
plt.show(True)
```

4 - Realize a classificação da base Vehicles usando validação cruzada e o método grid_search para escolher os melhores hiperparâmetros do modelo regressão logística e knn.

5 - Verifique se o número de folds, usado na validação cruzada, influencia na classificação da base winequality-red. Use o modelo de regressão logística.