Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Ciência de Dados:
Projeto: Classificação

Exercício 1 Considere a base de dados sobre doenças cardíacas:
`https://www.kaggle.com/ronitf/heart-disease-uci`
Faça o pré-processamento dos dados e classifique os pacientes de acordo com a variável "target". Considere os classificadores: Bayesiano paramétrico, Bayesiano não-paramétrico e Naive Bayes.

Exercício 2 No classificar não-paramétrico, verifique o efeito do hiperparâmetro h na classificação dos dados de diabetes, encontrando seu melhor valor:
`https://www.kaggle.com/uciml/pima-indians-diabetes-database`

Exercício 3 Considere o código abaixo para gerar dados artificialmente.

```
from sklearn import datasets
import matplotlib.pyplot as plt
plt.figure(figsize=(6,4))
n_samples = 1000
data  = datasets.make_moons(n_samples=n_samples, noise=.05)
X = data[0]
y = data[1]
plt.scatter(X[:,0], X[:,1], c=y, cmap='viridis', s=50, alpha=0.7)
plt.show(True)
```

Compare os resultados para os métodos Naive Bayes, Classificador Bayesiano paramétrico e o classificador Bayesiano não-paramétrico.

Exercício 4 Considerando os dados artificiais do exercício anterior, mostre as regiões de separação para os métodos Naive Bayes e Bayesiano paramétrico. Verifique como a região muda de acordo com a variável h no método não-paramétrico.

Exercício 5 Gere dois conjuntos de pontos em duas dimensões usando o código a seguir:

```
from sklearn.datasets import make_blobs
import numpy as np
import matplotlib.pyplot as plt
n = 500
c = [(1,1), (10,10)] #center of the points
std = [5.0, 2] # standard deviation
nc = [400,50] #number of points in each class
X, y = make_blobs(n_samples=n, n_features=2, cluster_std=std,centers= c)
plt.scatter(X[:,0],X[:,1], c=y)
plt.show(True)
```

Compare os classificadores Naive Bayes e Bayesiano Paramétrico variando a separação entre as nuvens de pontos – mantenha a posição de uma classe fixa e mude a posição do centro da outra classe, calculando a distância entre os centros.

Continua...