Classificador bayesiano no dataset Parkinson.

Arthur Felipe Reis Souza
Electrical Engineering Department,
Federal University of Minas Gerais,
Belo Horizonte, Brazil
arthurfreisouza@gmail.com

Antônio de Pádua Braga and Frederico Gualberto Ferreira Coelho
Electrical Engineering Department,
Federal University of Minas Gerais,
Belo Horizonte, Brazil
apbraga@cpdee.ufmg.br, fredgfc@ufmg.br

November 17, 2024

Abstract

1 Introdução

Este relatório tem por objetivo mostrar o processo de classificação bayesiana. As bases de dados que serão a XOR e duas distribuições normais.

2 Dados

Os dados foram gerados através de distribuições normais, e estão retratados nas imagens abaixo :

3 Aplicação do algoritmo

O classificador bayesiano leva em consideração a independência entre os atributos, bem como a probabilidade a priori de ocorrência da classe ao realizar a classificação. Ele se baseia no teorema de Bayes, que é descrito pela seguinte equação:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

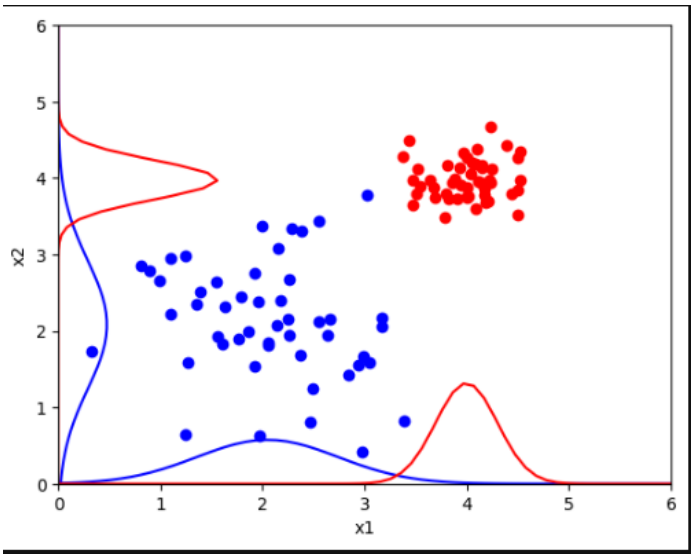


Figure 1: Dados gerados por 4 distribuições normais.

Apesar de considerar a independência entre as características, o mesmo pode ser extendido para casos onde há uma certa correlação entre os atributos. Nesses casos, o classificador irá considerar a matriz de covariâncias na estimativa da função densidade.

4 Resultados

Após aplicar o classificador bayesiano em ambas as bases de dados e, utilizando o K-Fold Cross Validation, com K = 10, obtemos que a acurácia de 72%. A superfície de separação para essa base de dados é mostrada abaixo :

Apesar de uma acurácia aceitável, o resultado não foi bom.

Para a segunda base de dados, a acurácia obtida foi de 81%.

5 Conclusão

Com este relatório, foi possível observar o classificador bayesiano em duas base de dados distintas. Os resultados obtidos mostram que o classificador teve um bom desempenho na base de dados 1, mas melhorou consideravelmente na base de dados 2. A matriz de covariâncias, nesse caso, contribuiu consideravelmente na predição de uma nova amostra.

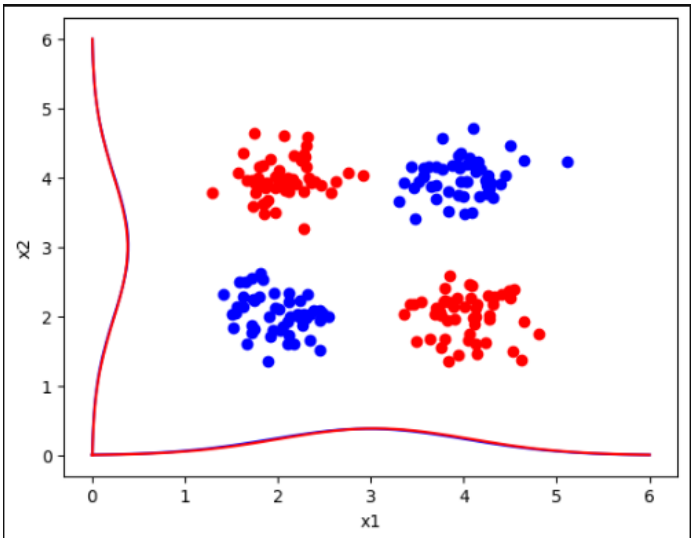
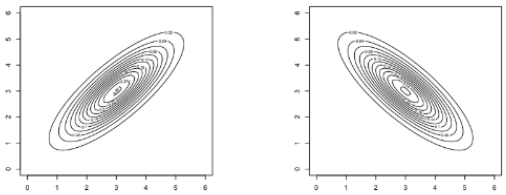


Figure 2: 4 distribuições geradas, seguindo a lógica da XOR.

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((\frac{x_1-\mu_1}{\sigma_1})^2 - (\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}) + (\frac{x_2-\mu_2}{\sigma_2})^2\right)} \tag{8.6}$$

em que ρ é o coeficiente de correlação linear entre as variáveis x_1 e x_2.



(a) $\rho = 0.8$ (b) $\rho = -0.8$

Figure 3: Equação geral da estimativa do classificador de bayes Multivariado.

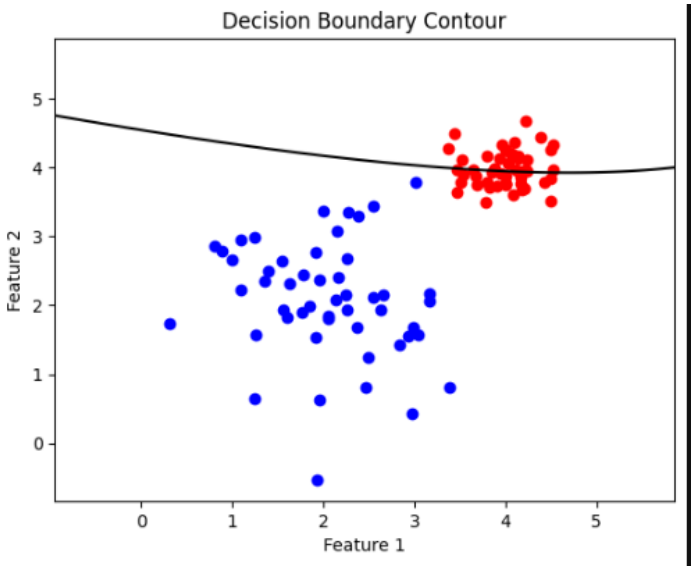


Figure 4: Superfície de separção base de dados 1.