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 utilizando a regra de Bayes como base, onde o conjunto são características que descrevem a doença de Parkinson.

2 Dados

Os dados são obtidos no site https://archive.ics.uci.edu/dataset/174/parkinsons, e contém 197 instâncias de 23 variáveis. O objetivo do estudo é, com base nas características, concluir se um paciente tem ou não a doença de Parkinson. Para isso, o classificador Multivariado de Bayes foi utilizado.

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

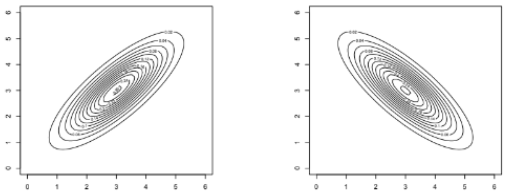
classificação. Ele se baseia no teorema de Bayes, que é descrito pela seguinte equação:

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

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.

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

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 1: Equação geral da estimativa do classificador de bayes Multivariado.

4 Resultados

Após aplicar o classificador bayesiano na base de dados Parkinson, utilizando o K-Fold Cross Validation, com K = 10, obtemos que a acurácia de 70%. Ao desconsiderar o K-Fold e separar os dados em treino e teste, a seguinte matriz de confusão foi obtida :

O classificador não obteve um bom desempenho. Ele está classificando pessoas que tem parkinson como pessoas que não tem.

5 Conclusão

Com este relatório, foi possível aplicar o classificador Bayesiano na base de dados de Parkinson. Os resultados obtidos mostram que o classificador teve um bom desempenho ao identificar pessoas com a doença. No entanto, ele classificou erroneamente 13 pessoas como não tendo a doença, quando na verdade elas a possuem. Problemas como esses nos levam a refletir sobre a importância de uma análise mais profunda das métricas a serem consideradas. Afinal, seria menos prejudicial classificar 13 pessoas como tendo Parkinson, quando elas não

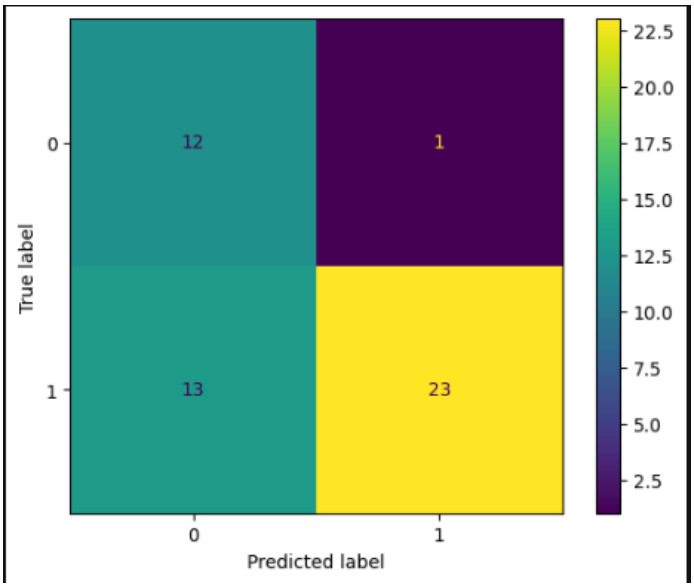


Figure 2: Matriz de confusão classificador de bayes na base de dados Parkinson.

têm, do que cometer o erro oposto e classificar pessoas que realmente possuem a doença como se não tivessem.