

1 Problema

Dataset: <https://archive.ics.uci.edu/ml/datasets/banknote+authentication>

Dataset Information:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Attribute Information:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

2 Classificadores

(a) Treine o classificador K-Vizinhos, utilizando para isso as seguintes medidas de distância:

1. Distância Euclidiana

$$d_{\text{Euclidiana}}(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

2. Distância Manhattan

$$d_{\text{Manhattan}}(X, Y) = \sum_{i=1}^n |x_i - y_i|$$

- (b) Treine um classificador bayesiano - Caso Univariado. A função de densidade da distribuição normal univariada é dada por:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]$$

- (c) Treine um classificador bayesiano - Caso Multivariado. A função de densidade normal multi-variada em d dimensões é dada por:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

Matriz de covariância:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

- \mathbf{x} é um vetor coluna com d componentes
- $\boldsymbol{\mu}$ é o vetor de médias de d componentes
- Σ é a matriz de covariâncias $d \times d$
- $|\Sigma|$ é o determinante de Σ
- Σ^{-1} é a inversa de Σ
- $(\mathbf{x} - \boldsymbol{\mu})^t$ é a transposta de $(\mathbf{x} - \boldsymbol{\mu})$

OBS: No Python esta matriz pode ser encontrada ao utilizar a biblioteca numpy e aplicar (np.cov).

- (d) Use validação cruzada “10-folds” (com random_state=42) para avaliar e comparar os classificadores.

- (e) Calcule a média e o desvio padrão das métricas de avaliação: Acurácia, Precisão e F1-Score, considerando todos os folds da validação cruzada. Em seguida, apresente os resultados em uma tabela comparativa, seguindo o modelo abaixo, de forma a facilitar a visualização das diferenças de desempenho entre os classificadores.

Exemplo de Tabela com Médias, Desvio Padrão e Tempos

Classificador	Acurácia	Precisão	F1-Score	Tempo Treino (s)	Tempo Teste (s)
K-vizinhos (Dist. Euclidiana)	0.85 ± 0.03	0.84 ± 0.04	0.84 ± 0.04	0.12 ± 0.01	0.03 ± 0.00
K-vizinhos (Dist. Manhattan)	0.88 ± 0.02	0.87 ± 0.03	0.87 ± 0.03	0.11 ± 0.01	0.02 ± 0.00
K-vizinhos (Dist. Chebyshev)	0.86 ± 0.03	0.85 ± 0.03	0.85 ± 0.03	0.13 ± 0.01	0.03 ± 0.00
Bayesiano (Multivariado)	0.90 ± 0.02	0.91 ± 0.02	0.90 ± 0.02	0.08 ± 0.01	0.01 ± 0.00
Bayesiano (Univariado)	0.87 ± 0.03	0.86 ± 0.03	0.86 ± 0.03	0.07 ± 0.01	0.01 ± 0.00

Tabela 1: Análise comparativa do desempenho dos classificadores, com médias, desvios padrão (\pm) e tempos de treino e teste.

Análise Comparativa dos Classificadores

Com base na Tabela 1, realize uma análise geral do desempenho dos classificadores, considerando tanto as métricas de avaliação quanto os tempos de execução. Discuta:

- Comparação das métricas (acurácia, precisão e F1-Score) entre os classificadores;
- Comparação dos tempos de treino e de teste entre os métodos;
- Relação entre desempenho (métricas) e eficiência computacional (tempos);
- Pontos fortes e limitações de cada classificador observados nos resultados.

Apresente suas conclusões destacando qual classificador oferece melhor equilíbrio entre desempenho e tempo de execução para este problema em específico.

Observações importantes

- A entrega das implementações é obrigatória. Equipes que não entregarem receberão nota zero.
- A pontualidade é um critério avaliativo.
- Os trabalhos serão submetidos a softwares de detecção de plágio, incluindo IA (ChatGPT, Deep Seek, etc.). Qualquer caso de plágio resultará em nota zero para as equipes envolvidas.
- Não é permitido o uso de bibliotecas com implementações prontas dos modelos exigidos, como o sklearn. Também não é permitido o uso do pandas.
- Os slides da apresentação deverão ser enviados no AVA dentro do prazo estabelecido. Os slides devem conter: Introdução, Explicação do Código, Funcionamento dos Algoritmos, Resultados e Conclusões. **Não é necessário entregar um documento de relatório separado. É necessário apenas os slides de apresentação.**
- Caso algum membro não consiga apresentar, deverá marcar em um outro momento a apresentação individual para o professor. Caso contrário, será atribuída nota zero.