Benchmark de Algoritmos de Classificação Supervisionada

Projeto Final - Planejamento e Análise de Experimentos

Ana Júlia de Lima Martins
Antônio Anunciação
Melchior Melo

Universidade Federal de Minas Gerais
Departamento de Engenharia Elétrica – DEE
Programa de Pós-Graduação em Engenharia Elétrica – PPGEE

14 de janeiro de 2025.

Conteúdo

Descrição do problema

- O câncer de mama é o tipo de câncer mais comum entre as mulheres no Brasil e no mundo. A maioria dos casos, quando tratados adequadamente e em tempo oportuno, apresentam bom prognóstico e possibilitam melhores resultados estéticos.
- Nos últimos, anos a aplicação de estratégias de aprendizado de máquina tem se mostrado eficaz no auxílio do diagnóstico precoce desse tipo de câncer.
- Dentre os diversos algoritmos disponíveis para classificação supervisionada, haveria algum com melhor desempenho do que os demais para disgnóstico de câncer de mama, considerando diferentes cenários de ruído no treinamento dos modelos?

Descrição do Problema

Objetivo:

- Comparar o desempenho de diferentes algoritmos de classificação supervisionada no contexto do diagnóstico de câncer de mama utilizando conjuntos de dados reais. O foco está na avaliação da acurácia do algoritmo, quando submetido a diferentes níveis de ruído.

Dataset

Breast Cancer Wisconsin (Diagnostic):

Conjunto de dados reais para classificação binária.

- Número de amostras: 569
- Número de atributos: 10 atributos, incluindo medidas como raio, textura e simetria das células.

Dataset

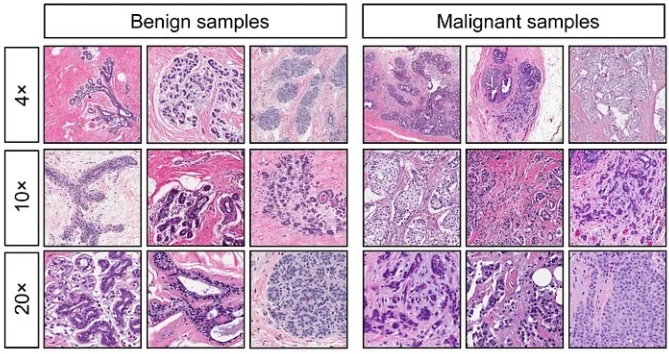

Figura: Exemplo de amostra.

Metodologia

O experimento será conduzido utilizando um Experimento Completamente Aleatorizado com Blocos (RCBD).

$$\begin{cases} H_0 : \tau_i = 0, \ \forall i \in \{1, 2, ..., a\} \\ H_1 : \exists \ \tau_i \neq 0 \end{cases} \tag{1}$$

- **Níveis:** Diferentes algoritmos de classificação supervisionada:
  - XGBoost
  - Random Forest
  - Support Vector Machine (SVC)
  - Multilayer Perceptron (MLP)
- **Blocos:** Níveis de ruído introduzido no treinamento do dataset.

Metodologia

A métrica de avaliação de desempenho do modelo será acurácia percentual do modelo.

$$\text{Acurácia } \% = \frac{\text{Número de acertos}}{\text{Número total de testes}} \times 100\% \tag{2}$$

Metodologia

Tamanho Amostral

- O tamanho amostral necessário pode ser calculado como o menor inteiro N tal que o poder do teste π^* seja igual ou maior que o poder desejado. No caso da hipótese alternativa unilateral, temos que:

$$N^* = \min N \left| \; t_{1-\alpha}^{(N-1)} \leq t_{\beta^*\,;\,|\mathsf{ncp}^*|}^{(N-1)} \right. \tag{3}$$

- A abordagem proposta para calcular o número de repetições por bloco é similar ao procedimento para um CRD (*Completely Randomized Design*).
- A variabilidade intra-grupo, necessária para calcular o tamanho amostral necessário, será estimada a partir de um estudo piloto.

Metodologia

- O treinamento do modelo será feito por meio de validação cruzada a fim de mitigar os efeitos estocásticos dos métodos de aprendizado de máquina.
- O conjunto de treinamento será dividido em blocos e serão treinados em conjuntos aleatórios. A proporção de classes Malígno e Beníngo serão mantidas para evitar desbalanceamentos no problema de classificação.

Resultados Esperados

- Identificar algoritmos mais robustos para classificação do *dataset*, considerando os diferentes cenários de ruído.
- Avaliação dos resultados por matriz de confusão e cálculo dos erros tipo I e II empirícos.
- Propor recomendações para aplicações práticas.

Referências

[1] Lea Eckhart, Kerstin Lenhof, Lisa-Marie Rolli, Hans-Peter Lenhof, A comprehensive benchmarking of machine learning algorithms and dimensionality reduction methods for drug sensitivity prediction, Briefings in Bioinformatics, Volume 25, Issue 4, July 2024, bbae242, https://doi.org/10.1093/bib/bbae242

[1] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System https://arxiv.org/abs/1603.02754