Pontifícia Universidade Católica de Minas Gerais
Instituto de Ciências Exatas e Informática
Curso: Sistemas de Informação
Disciplina: Ciência de Dados
Professor: Cristiano Rodrigues

Trabalho Final: Classificação de Exoplanetas

Objetivo

O objetivo do Trabalho Final é praticar os conceitos aprendidos na disciplina e adquirir experiência no uso de alguns dos principais métodos de classificação, na avaliação de modelos e na interpretação e apresentação de resultados de experimentos.

Para isso, você irá utilizar e comparar métodos de classificação baseados em princípios diferentes em um problema de classificação binária de candidatos a exoplanetas.

Tarefas

Neste trabalho, você deverá realizar uma comparação entre seis métodos de classificação:

- Naive Bayes
- Decision Tree
- k-Nearest Neighbors
- Support Vector Machines
- Random Forest
- Gradient Tree Boosting

Você deverá realizar os experimentos listados abaixo. Pode ser necessário normalizar os dados e testar diferentes valores para os hiperparâmetros dos métodos para se obter bons resultados (não é necessário entregar todas as combinações testadas, apenas a de melhor resultado, exceto os casos que foram pedidos abaixo). A avaliação dos métodos deverá ser feita usando a métrica acurácia e a estratégia de validação cruzada *k-fold* com $k = 5$.

- **Naive Bayes**: Experimento para servir de baseline.
- **Decision Tree**: Variar a altura máxima da árvore (incluindo permitir altura ilimitada) e mostrar os resultados graficamente.
- **SVM**: Avaliar os kernels linear, sigmoid, polinomial e RBF.
- **k-NN**: Variar o número k de vizinhos e mostrar os resultados graficamente.
- **Random Forest**: Variar o número de árvores e mostrar os resultados graficamente.
- **Gradient Tree Boosting**: Variar o número de iterações e mostrar os resultados graficamente.

Todos os métodos listados acima estão disponíveis na biblioteca *scikit-learn* (`https://scikit-learn.org/stable/`) da linguagem Python. Podem ser utilizadas bibliotecas auxiliares, como por exemplo, *matplotlib* (`https://matplotlib.org`), para gerar gráficos.

Para cada um dos experimentos realizados, explique qual o objetivo do experimento (qual o significado do hiperparâmetro que está sendo variado, por exemplo) e inclua uma interpretação dos resultados com base nos conceitos teóricos estudados na disciplina.

Ao final, deverá ser feita uma comparação entre o desempenho dos métodos. Se desejar, você pode incluir a curva ROC e as métricas de precisão e revocação (*precision* e *recall*).

Conjunto de Dados

Os métodos serão testados em um problema de classificação binária de candidatos a exoplanetas encontrados pela sonda espacial Kepler da NASA[1].

Um exoplaneta é um planeta fora do sistema solar (i.e., que não orbita o sol). A sonda primeiro identifica sinais de possíveis exoplanetas, chamados de *Kepler Object of Interest* (KOI). Porém, nem todos os KOIs são de fato exoplanetas; alguns se tratam de falsos positivos de origens diversas. A tarefa é, então, classificar os KOIs entre exoplanetas confirmados e falsos positivos. Cada observação corresponde a um KOI, e as *features* são características estimadas de cada (possível) exoplaneta (tamanho, temperatura, *features* da estrela hospedeira, etc.).

O conjunto de dados estará pronto para uso e está disponibilizado no arquivo `koi_data.csv`. O arquivo está no formato CSV separado por vírgulas. A primeira coluna identifica o KOI, a segunda traz a sua classificação correta (`FALSE_POSITIVE` ou `CONFIRMED`) e as demais colunas são *features* sobre o KOI extraídas de diversas formas. Para este trabalho, não será necessário entender o significado das *features*.

Entrega

O trabalho deverá ser entregue no formato de Jupyter notebook (`.ipynb`). O notebook deverá conter todo o código (devidamente comentado) necessário para executar os experimentos, a apresentação dos resultados por meio de texto, gráficos e tabelas, a explicação do que está sendo feito e a interpretação dos resultados. Apenas o notebook deve ser entregue. A organização e a clareza do notebook fazem parte da avaliação do trabalho. O notebook deve ser organizado de tal forma a ser possível reproduzir os experimentos apenas executando as células do notebook em ordem.

Critérios de Avaliação

- Implementação correta dos métodos, dos experimentos e da avaliação dos modelos.
- Organização das informações apresentadas no notebook.
- Apresentação dos experimentos e dos resultados de forma clara, concisa e não ambígua.
- Corretude conceitual das explicações e interpretações dos experimentos.

A avaliação não será feita pelos valores de acurácia obtidos, mas pelo processo em si.

[1] Dados retirados do NASA Exoplanet Archive (`https://exoplanetarchive.ipac.caltech.edu/`)