Classificação de Exoplanetas

Trabalho Prático I - Aprendizado de Máquina

1 Objetivo

O objetivo deste trabalho prático é de praticar os conceitos aprendidos na disciplina e de 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.

2 Tarefa

Neste trabalho você deverá realizar uma comparação entre seis métodos de classificação: Naive Bayes, Decision Tree, k-Nearest Neighbors, Support Vector Machines, Random Forest e Gradient Tree Boosting. Além disto você deverá realizar os experimentos listados abaixo específicos para cada método. 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 feito usando a acurácia e validação cruzada k-fold com k igual a 5.

- **Naive Bayes**: Apenas um 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.

Você não precisa implementar os métodos listados acima. Todos os métodos listados acima estão disponíveis na biblioteca *scikit-learn* da linguagem Python. Você também pode utilizar bibliotecas auxiliares, para gerar gráficos e de operações matemáticas por exemplo, desde que elas não implementem o experimento em si.

Para cada um dos experimentos realizados você deverá explicar qual o objetivo do experimento (qual o significado do hiperparâmetro que está sendo variado por exemplo) e incluir uma interpretação dos resultados com base nos conceitos teóricos estudados na disciplina. Ao final deverá ser feita uma comparação entre a performance dos métodos, incluindo curva ROC e as métricas de precisão e revocação (precision e recall).

3 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). Porem 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 será disponibilizado no Moodle no arquivo koi_data.csv. O arquivo estará no formato CSV separado por virgulas. 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.

4 Entrega

O trabalho deverá ser entregue no formato de Jupyter (IPython) notebook. O notebook deverá conter todo o código (devidamente comentado) necessário para executar os experimentos, a apresentação dos resultados por meios 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 monitor deverá ser capaz de reproduzir os experimentos apenas executando as células do notebook em ordem.

O notebook deverá ser submetido no Moodle até o dia 22/05 às 23:59 (apenas o arquivo .ipynb deve ser submetido). **Trabalhos submetidos fora do prazo não serão avaliados**. **Não haverá extensões do prazo**. **Se for detectado plágio, todos os alunos envolvidos receberão nota zero e serão encaminhados para a coordenação do curso**.

5 Critérios de Avaliação

- Implementação correta do uso dos métodos, dos experimentos e da avaliação dos modelos (50%)
- Apresentação dos experimentos e dos resultados de forma clara, concisa e não ambígua (20%)
- Corretude conceitual das explicações e interpretações dos experimentos (20%)
- Organização das informações apresentadas no notebook (10%)

Note que você não será avaliado pelos valores de acurácia obtidos, apenas o processo em si.

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