

Trabalho Prático 4 - Aprendizagem de Máquina¹

Objetivos

O objetivo deste trabalho é 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.

Descrição

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 feita 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 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 eles estão disponíveis na biblioteca *scikit-learn*² da linguagem Python. Você também pode (deve) 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 o desempenho dos métodos, incluindo reporte das 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 NASA³. 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 sua 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 PVANet Moodle 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*.

¹Elaborado com base no material do professor Adriano Veloso, DCC/UFMG.

²<https://scikit-learn.org/>

³Dados retirados do NASA Exoplanet Archive (<https://exoplanetarchive.ipac.caltech.edu/>)

Entrega

A entrega deve ser efetuada conforme agendado no PVA^{Net} Moodle. 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, se necessário, além da explicação do que está sendo feito e a interpretação dos resultados. Apenas o notebook deve ser entregue em uma pasta compactada (.rar ou .zip) também através do PNA^{Vet} Moodle. A pasta compactada deve conter informações do aluno (ex.: julio_reis-tp4.zip). Se preferir pode enviar apenas um arquivo contendo o link.

A organização e a clareza do notebook fazem parte da avaliação do trabalho, permitindo a reprodução dos experimentos apenas executando as células do notebook em ordem. Por fim, perceba que você não será avaliado pelos valores de acurácia obtidos, apenas o processo em si.

Comentários Gerais

- O trabalho é individual (grupo de UM aluno);
- Trabalhos copiados serão penalizados (NOTA Zero).