[2021.1] Ciência de Dados

Comparação de diferentes algoritmos para criação de modelos

Trabalho 2

VISÃO GERAL

Ao longo das 10 primeiras aulas da disciplina de Ciência de Dados, vimos alguns algoritmos para criação de modelos, como Regressão Linear e Logística e k Vizinhos mais Próximos (kNN). Esses foram apenas três dentre uma grande variedade de algoritmos existentes. O objetivo do presente trabalho é fazer com que você busque e compare os algoritmos vistos em sala de aula com outros algoritmos não vistos, mas disponíveis na biblioteca `scikit-learn`.

OBJETIVOS

1. Buscar e utilizar algoritmos não vistos na disciplina para tarefas de regressão e classificação.
2. Comparar os algoritmos escolhidos com conjuntos de dados reais utilizando métricas de avaliação vistas ou não na disciplina.

ESPECIFICAÇÕES

Neste trabalho, você deverá escolher um ou mais conjuntos de dados para comparar diferentes algoritmos para criação de modelos. Não há uma quantidade máxima de algoritmos, mas você deve comparar pelo menos um algoritmo visto na disciplina com um algoritmo não visto, tanto para tarefas de regressão quanto para tarefas de classificação. Exemplo:

- Regressão: Regressão Linear e Naive Bayes
- Classificação: kNN e SVM

No relatório a ser entregue, escreva uma breve explicação dos algoritmos utilizados que não foram vistos em sala de aula.

Você pode utilizar as métricas de avaliação que desejar. Inclua uma breve explicação para a escolha das métricas apresentadas.

AVALIAÇÃO

Como não sabemos o dataset que será escolhido, decidimos não definir um número mínimo de análises a serem feitas. Porém, utilizaremos os seguintes critérios:

- **Organização do notebook** - seu relatório será todo feito usando o Jupyter Notebook (o qual será o artefato de envio do trabalho). Logo, ele precisa está bem organizado, combinando código python com markdown;
- **Organização dos experimentos** - não basta treinar os modelos e apresentar uma métrica qualquer. Procure métricas que possam ser mais adequadas para seu problema. Além disso, utilize recursos para deixar sua experimentação mais consistente, como *grid search*, *cross validation*, etc.
- **Reprodutividade do relatório** - seu notebook será executado no momento da avaliação. Garanta que as células estejam na ordem correta e que qualquer instrução adicional para a reprodutibilidade esteja devidamente apresentada no notebook;
- **Tamanho da equipe** - trabalhos feitos por equipes possuem avaliações mais rígidas, sendo a dificuldade proporcional (não necessariamente de forma linear) ao número de membros.

Referências

1. Documentação de algoritmos do `scikit-learn`: https://scikit-learn.org/stable/supervised_learning.html