Universidade Tecnológica Federal do Paraná
Especialização em Ciência de Dados
Métodos e Modelos Avançados em Ciência de Dados
"Atividade 05 - Ensembles"
Prof. Dr. Rafael Gomes Mantovani

Objetivo da atividade

Objetivo desta atividade é então avaliar as implementações existentes de algoritmos de ensembles em problemas reais, cujos dados podem ser encontrados em repositórios de dados públicos. Além disso, também deseja-se compreender melhor o comportamento do algoritmo. Opcionalmente, é sugerido implementar versões próprias dos algoritmos de Bagging, AdaBoost e Random Forest, e compará-los também com os resultados obtidos com as ferramentas.

1 O que fazer?

Encontre diferentes implementações de algoritmos de ensemble (Bagging, Boosting, RandomForest) disponíveis em R/Python e avalie o desempenho deste destes algoritmos em algum problema de sua escolha. Para a seleção de datasets, recomendo novamente o OpenML (https://www.openml.org), ou UCI (https://archive.ics.uci.edu/ml/datasets.php). Caso queiram usar algum outro conjunto de dados do UCI, Kaggle e afins, fiquem à vontade para explorá-los também.

Com o dataset selecionado, execute alguns testes/experimentos para avaliar os algoritmos que são nosso foco de estudo. Faça testes variando o número de algoritmo base (*base learners*) no comitê (ensemble) e verifique a mudança nos valores de desempenho obtidos nos conjuntos de teste.

Como trabalho, elabore um relatório técnico (documento pequeno, .pdf, ou notebook) que descreva todas as etapas realizadas:

- explicar o dataset/problema escolhido;
- mostrar algumas estatísticas interessantes do problema, e justificar porque ele foi escolhido como objetivo de investigação;
- explicar as implementações de ensembles usadas, correspondentes hiperparâmetros, e características gerais de uso;
- descrever a metodologia experimental: como os dados foram manipulados para indução dos modelos, qual medida de desempenho foi usada, e quais tipos de investigações/experimentos foram realizados. Se possível, usar validação cruzada;
- adicionar alguns algoritmos convencionais como *baselines* (árvore de decisão, knn, svm, etc);
- apresentar os resultados obtidos e descrever conclusões críticas com base nestes resultados observados.

2 O que também pode fazer? - Opcional

Se você conseguiu realizar todas as tarefas acima e ainda tem algum tempo para dedicar aos estudos:

- Implemente a sua versão dos algoritmos de *Bagging*, *AdaBoost (boosting)* e *Random Forest*;
- compare seus algoritmos com as implementações existentes;
- quão bem seus códigos foram em relação as ferramentas da literatura?
- como seu código poderia ser futuramente melhorado?

3 Alguns links que podem ser úteis

- **mlr3** - *Machine Learning in R*: *framework* em R para implementação de soluções de Aprendizado de Máquina. Ele é bem completo, quase tudo já está implementado, o problema pode ser entender como encaixar cada peça. Link: https://mlr3.mlr-org.com;
- **scikit learn** - *Machine Learning in Python* - contapartida do mlr mas em Python. https://scikit-learn.org/stable/