

Universidade do Estado do Amazonas

Escola Superior de Tecnologia

Data: 31 de Março de 2018

Professora: Elloá B. Guedes

Disciplina: Tópicos Especiais III *Machine Learning* – 2018.1

Curso: Engenharia de Computação

MINI-TESTE #2

Para esta tarefa prática será considerado o [Wine Quality Dataset](#), o qual contempla vinhos verdes e tintos do norte de Portugal. As equipes de quatro estudantes devem produzir um relatório de até 3 páginas, em formato de texto corrido que descreva os resultados do que se pede a seguir.

O problema a ser considerado será de classificação multi-classe, em que cada classe corresponde à qualidade do vinho, atributo alvo desta tarefa de aprendizado. O objetivo será comparar diferentes modelos de Aprendizado de Máquina já vistos na disciplina para esta tarefa.

Primeiramente, dê uma visão geral do dataset disponível, descrevendo os atributos preditores, a quantidade de exemplos e o atributo-alvo, se este é balanceado ou não. Para comprovar esta afirmação, apresente um histograma do atributo-alvo.

A métrica de desempenho a ser considerada neste problema será o F-Score. Se o atributo-alvo for balanceado, a métrica a ser usada é resultante do *macro-averaging*. Em caso contrário, será o *micro-averaging*. Consulte a [documentação](#) do sci-kit learn para obter corretamente as métricas. Além disto, iremos utilizar $k = 3$ folds como método de validação cruzada para analisar e comparar os modelos obtidos.

Construa árvores de decisão para este problema, considerando diferentes critérios de separação (critério de Gini e ganho de informação), bem como diferentes valores de profundidade máxima e quantidade máxima de atributos. Proponha, no mínimo, 12 árvores de decisão diferentes para este problema. Descreva quais árvores a equipe decidiu considerar.

Além das árvores de decisão, as equipes devem construir também modelos baseados em *ensembles*, tais como [florestas aleatórias](#), [Adaboosting](#) e [Bagging](#). Nestes últimos tipos de modelos, em particular, não é necessária uma grande variação nos parâmetros. Reforça-

se apenas a necessidade de descrição dos mesmos. A mesma estratégia de avaliação dos modelos se aplica a este cenário.

As equipes devem descrever um sumário de quantos modelos serão treinados e testados para o problema em questão. Utilizando as tecnologias apropriadas, devem proceder com a implementação dos modelos, treinamento e testes.

Uma tabela deve sumarizar os resultados obtidos, ordenando os modelos de acordo com o maior F-Score. O modelo com melhor desempenho deve ser apresentado em detalhes, incluindo o tempo necessário para treinamento e teste do mesmo e uma matriz de confusão dos seus resultados para um dos folds. Se possível, no caso de ser uma árvore, uma representação gráfica do modelo deve ser apresentada.

Cada equipe deve finalizar seu texto com uma justificativa, baseada em elementos do Aprendizado de Máquina, que explique por que o modelo em questão capturou adequadamente as características do problema, obtendo desempenho superior aos demais modelos.

A avaliação deste mini-teste levará em conta a apresentação dos elementos obrigatórios, a correteza na condução da metodologia e a qualidade na descrição e justificativa dos resultados obtidos.

Prazos importantes

Apresentação do problema. 03/04/2018

Atendimento de dúvidas. Toda a aula do dia 03/04/2018 e metade do tempo de aula do dia 05/04/2018.

Entrega. 10/04/2018, 15h, via Google Classroom.