Trabalho 2
Wine Quality Dataset

1 Descrição do Dataset

A qualidade de um vinho depende de diversos fatores, como a forma que as uvas são cultivadas (viticultura) e como elas são transformadas em vinhos (vinificação). A forma como todo o processo é feito tem relação direta com diversas características químicas da bebida e, consequentemente, com o seu sabor. Uma medida da qualidade de um vinho pode ser feita baseando-se em medições químicas (pH, densidade, acidez, porcentagem de álcool) e testes sensoriais feitos por especialistas.

Nestre trabalho, você irá classificar a qualidade de um vinho a partir de medições químicas. As medições disponíveis são:

- **Fixed acidity**
- **Volatile acidity**
- **Citric acid**
- **Residual sugar**
- **Chlorides**
- **Free sulfur dioxide**
- **Total sulfur dioxide**
- **Density**
- **pH**
- **Sulphates**
- **Alcohol**

- **Quality**: a qualidade do vinho, com valor "0" (ruim) ou "1" (bom).

2 Tarefas

Pedimos que você:

1. Inspecione os dados. Quantos exemplos você tem? Qual o intervalo de valores de cada feature?
2. Inspecione o número de exemplos de cada classe (*Quality*) no conjunto de treinamento:
   (a) Há aproximadamente a mesma quantidade de exemplos de cada classe? Ou seja, as classes estão balanceadas?
   (b) Como você lidaria com classes desbalanceadas?
3. Normalize os dados de modo que eles fiquem todos no mesmo intervalo.
4. Treine uma regressão logística com todas as features para predizer a qualidade dos vinhos (*baseline*).
5. Implemente soluções alternativas baseadas em regressão logística (através da combinação dos features existentes) para melhorar os resultados obtidos no baseline.
6. Classifique os dados de validação.
7. Calcule a matrix de confusão e acurácia no conjunto de validação.
8. Escreva um relatório de no máximo 3 páginas:
   (a) Descreva o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;
   (b) Reporte os resultados no conjunto de validação e de teste (este último será disponibilizado alguns dias antes do prazo final de submissão) e compare possíveis diferenças nos resultados. Ocorreu overfit?
   (c) Escreva pelo menos 1 parágrafo com as conclusões tiradas na atividade;

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *wineQuality_train.data*: conjunto de dados para treinamento;
- *wineQuality_val.data*: conjunto de dados para validação;
- *wineQuality_test.data* **(será disponibilizado na sexta-feira anterior ao prazo final da submissão)**: conjunto de dados retido pelo professor;

4 Avaliação

O dataset foi previamente dividido aleatoriamente em três conjuntos — treino, validação e teste — e apenas os dois primeiros serão disponibilizados para que você implemente as suas soluções.

Na sexta-feira anterior ao prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-lo pelo canal da disciplina no Slack. No relatório, você deve reportar os seus resultados no conjunto de validação e no conjunto de teste.

A avalidação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foi feito, os resultados reportados e as conclusões feitas.

Observações sobre a avaliação:

- O trabalho poderá ser feito individualmente ou em duplas, podendo haver repetição das duplas a cada trabalho;
- O código e o relatório deverão ser submetidos no Moodle por **apenas um integrante da dupla**;
- Não se esqueçam de listar os nomes dos integrantes da dupla no início do relatório;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;