

Aprendizagem Automática II

Proposta para o Trabalho Prático

-

Mestrado em Engenharia Informática
Universidade do Minho

Grupo

PG41080	João Ribeiro Imperadeiro
PG41081	José Alberto Martins Boticas
PG41091	Nelson José Dias Teixeira
PG41851	Rui Miguel da Costa Meira

April 6, 2020

1 Introdução

Tal como foi requerido pelo docente da unidade curricular "Aprendizagem Automática II", este documento serve para fazer um levantamento preliminar do que será desenvolvido para o trabalho prático da respetiva UC. Exibe-se, de seguida, o repositório *GitHub* público associado a este projeto, onde será colocado todo o código desenvolvido pelo grupo bem como toda a documentação e recursos intrínsecos à implementação:

`https://github.com/Nelson198/AA2`

2 Tema

O tema escolhido pelo nosso grupo é o desenvolvimento de uma *framework* de AutoML. A *framework* visa obter o melhor modelo para problemas de *supervised learning* e *unsupervised learning*, de forma automática e com a menor intervenção possível por parte do programador. O objetivo final é colocar a *framework* disponível para os utilizadores da linguagem Python.

3 Planificação

Devido à complexidade do desenvolvimento deste projeto e atendendo ao curto espaço de tempo disponível, não serão incluídas opções de pré-processamento de dados. Com isto, o utilizador/programador deverá indicar qual o tipo de modelo (regressão, classificação ou *clustering*) que deseja obter, sendo depois da responsabilidade da *framework* a procura do melhor modelo desse tipo, visitando todos os algoritmos disponíveis. Se o utilizador preferir um algoritmo em

especial poderá indicá-lo, sendo da responsabilidade da *framework* a procura dos melhores hiperparâmetros.

A *framework* vai distinguir entre um problema de *supervised* e *unsupervised learning* consoante receba, ou não, as *labels/targets*. Todos os problemas de *supervised learning* vão ser distinguidos entre regressão e classificação, dependendo se o *target* é uma variável contínua ou discreta.

Para problemas de regressão, os algoritmos disponíveis serão:

- Regressão Linear
- Regressão Polinomial
- *Support Vector Regression*
- *Decision Tree Regression*
- *Random Forest Regression*
- Redes Neurais

Para problemas de classificação, os algoritmos disponíveis serão:

- Regressão Logística
- *k-Nearest Neighbors* (KNN)
- *Support Vector Machine* (SVM)
- Kernel SVM
- *Naive Bayes*
- *Decision Tree Classification*
- *Random Forest*
- Redes Neurais

Para problemas de *clustering*, os algoritmos disponíveis serão:

- *k-Means Clustering*
- *Hierarchical Clustering*

A procura de melhores hiperparâmetros para Redes Neurais vai ser realizada com o uso da biblioteca **kerastuner**. Para outros algoritmos, vai-se fazer uso do *GridSearchCV* e *RandomizedSearchCV* para esta escolha.

Na eventualidade da plataforma estar terminada e ainda existir tempo para tal, será estudada a possibilidade de se incluir o pré-processamento de dados, aumentando a complexidade mas também a flexibilidade da *framework*. Isto poderá permitir que sejam testados modelos de tipos distintos.

4 Objetivos

1. Permitir o teste de diferentes modelos, com diferentes algoritmos, para um certo conjunto de dados;
2. Comparar diferentes modelos, apresentando a suas métricas;
3. Encontrar o melhor modelo, com base nas métricas apresentadas;
4. Disponibilizar a plataforma *online*, para uso da comunidade;
5. Preparar o projeto para a inclusão de pré-processamento de dados.