Sistemas Inteligentes para a Bioinformática
2022/2023

Trabalho prático – análise de dados usando aprendizagem máquina

Este trabalho consiste na análise de um conjunto de dados, através da utilização de algoritmos de aprendizagem máquina, usando o Python como linguagem de programação. Deverá ser elaborado um *Jupyter Notebook*, organizado em secções, que inclua os passos da análise realizada, e que explique muito sucintamente os procedimentos realizados e decisões tomadas ao longo da análise. Além disso, os grupos farão uma apresentação intermédia do trabalho realizado.

Os trabalhos poderão ser realizados em vários tipos de problemas a abordar (opção A):

A1. Classificação de proteínas, a partir da sua sequência de aminoácidos
A2. Classificação de compostos, a partir da sua representação em SMILES (ou outra)
A3. Classificação de fenótipos, a partir de dados ómicos
A4. Outros conjuntos de dados de interesse em Bioinformática
A5. Combinação de mais do que uma das anteriores

Além disso, o trabalho poderá ser realizado na análise de dados de competições atualmente em curso (opção B) e que se relacionam com os tópicos acima indicados. Foram identificadas as seguintes competições de interesse:

B1. Dream challenge Heart Failure and Microbiome - https://www.synapse.org/#!Synapse:syn27130803/wiki/619274
B2. Kaggle – Novozymes enzyme stability prediction - https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction

Os grupos de trabalho devem ser compostos por 3 elementos. Cada grupo deverá escolher uma das opções (A ou B) para a realização do trabalho. Deverá ser submetida no e-learning a proposta do grupo, indicando qual a opção. No caso da opção A, deverão indicar quais os dados que se propõem analisar. Para esta escolha poderão usar a literatura existente e procurar datasets em bases de dados existentes.

Genericamente, ao longo do trabalho, devem ser realizadas as seguintes etapas indicativas (podem ser adaptadas dependendo dos dados):

Fase 1:
Exploração inicial e pré-processamento

- Rever toda a documentação disponível sobre o conjunto de dados.
- Carregar o conjunto de dados e realizar uma análise exploratória do mesmo.

- Realizar os passos necessários de preparação dos dados e pré-processamento, incluindo possivelmente a geração de atributos, a sua seleção, o tratamento de possíveis valores em falta, etc.

Esta etapa deve corresponder à secção 1 do *Notebook* onde deverá:

- descrever e caracterizar os dados atribuídos de acordo com a documentação/ literatura existente;
- descrever sucintamente as características dos dados disponíveis a partir da análise exploratória inicial;
- descrever os passos de preparação dos dados e pré-processamento que efetuou, justificando as suas escolhas;
- incluir os gráficos exploratórios iniciais que ilustrem as principais características dos dados.

Análise não supervisionada

- Usar as técnicas de redução de dimensionalidade adequadas aos seus dados;
- Usar as técnicas de visualização de dados multivariadas adequadas aos seus dados;
- Aplicar métodos de clustering que considere adequados aos seus dados.

Esta etapa deve corresponder à secção 2 do *Notebook* onde deverá:

- Reportar/ analisar os resultados obtidos para as técnicas de redução de dimensionalidade e visualização de dados;
- Reportar/ analisar os resultados obtidos a partir dos algoritmos de clustering.

Aprendizagem máquina

- Comparar o comportamento de diversos modelos/ algoritmos de Aprendizagem Máquina no conjunto de dados. Deverá analisar o comportamento dos algoritmos calculando métricas de erro apropriadas e usando métodos de estimação do erro adequado. Deverá ainda apresentar o melhor modelo a que consiga chegar para os dados disponíveis, usando todos os exemplos, interpretando-o quando tal for possível.

Esta etapa deve corresponder à secção 3 do *Notebook* onde deverá reportar e analisar criticamente os resultados da etapa 3.

Fase 2:

Deep learning

- Utilizar métodos de *deep learning* de forma semelhante à etapa 3, comparando os resultados com os métodos aí apresentados.

Esta etapa deve corresponder à secção 4 do *Notebook* onde deverá reportar e analisar criticamente os resultados da etapa 4.

Datas importantes:

- Submissão da constituição do grupo e escolha do tema: **25 novembro 2022**
- Submissão dos notebooks:

    - Etapas 1 a 3 – **21 dezembro 2022**
    - Etapa 4 – **2 fevereiro 2022**
- Apresentação: **18 janeiro 2022**

Avaliação:
A avaliação dos trabalhos será feita recorrendo a 3 elementos:

- Avaliação do notebook da fase 1 – peso de 40%
- Avaliação do notebook da fase 2 – peso de 30%
- Apresentação – peso de 30%

O docente reserva-se o direito de **justificadamente** poder atribuir classificações distintas aos vários elementos de cada grupo. Os elementos de cada grupo poderão ser chamados a avaliar o desempenho dos colegas no trabalho, bem como a comentar/ avaliar o trabalho de outros grupos.