Aprendizagem Automática II
Trabalho prático – ano letivo 2019/2020

O trabalho prático a realizar nesta UC terá duas opções distintas:

- Opção 1: Análise de um conjunto de dados selecionado, usando a linguagem Python e *packages* disponíveis.
- Opção 2: Trabalho de desenvolvimento de algoritmos/ software no âmbito da Aprendizagem Máquina

Opção 1:

Neste caso, o grupo deverá começar por escolher o conjunto de dados que irá ser usado na análise de dados. Os dados poderão ser retirados de algum tipo de publicação (e.g. artigo científico). Não se recomenda a escolha de conjuntos de dados de competições onde existam já muitas soluções disponíveis por parte da comunidade, mas participar em competições a decorrer com soluções inovadoras é uma opção. Serão indicadas em anexo algumas possibilidades sugeridas pelo docente, podendo ser escolhido pelo grupo outro conjunto de dados que terá que ser validado pelo docente.
A análise dos dados deverá ser adequada às características dos dados escolhidos e ao seu domínio de aplicação, sendo indicadas abaixo um conjunto de tarefas recomendadas (algumas poderão não fazer sentido dependendo dos dados escolhidos e outras não mencionadas poderão ser igualmente consideradas relevantes):

- carregamento e preparação dos dados (incluindo se necessário a geração dos atributos);
- exploração inicial,incluindo sumarização, visualização e pré-processamento;
- aprendizagem não supervisionada (redução de dimensionalidade, clustering);
- seleção de atributos;
- aprendizagem supervisionada (usando métodos "tradicionais") incluindo a otimização dos seus hiper-parâmetros. Poderão ser usados e avaliados (de forma robusta) diversos métodos de Aprendizagem Máquina mais tradicional, de acordo com as características dos dados;
- metodologias de *deep learning*, comparando os resultados com os métodos anteriores, devendo ser exploradas alternativas em termos das configurações destes métodos, podendo ser testadas também classes distintas de modelos (e.g. *feedforward*, recorrentes, convolucionais).

Deverá ser criado um (ou vários) *Jupyter Notebooks* contendo os scripts usados e os resultados obtidos, bem como incluindo a explicação das razões para cada um dos métodos e discutindo-se os resultados obtidos. Estes podem usar ficheiros de código adicional desenvolvido pelo grupo (módulos incluindo funções ou classes). Todo o código desenvolvido por terceiros que seja usado deve ser devidamente identificado.

Opção 2:

Neste caso, o grupo deverá começar por escolher o tema para o desenvolvimento a efetuar. Na proposta deve incluir os objetivos do desenvolvimento a atingir. Os temas sugeridos serão validados pelo docente.
No final, o grupo deve escrever um relatório sucinto do trabalho desenvolvido (máximo de 5 páginas), juntando os ficheiros relevantes em anexo (código, tabelas de resultados, etc). Nesta opção, não será imposta uma linguagem para o desenvolvimento de software.

Datas e condições de entrega:

Em qualquer das opções, o grupo deve colocar o código desenvolvido num repositório público (e.g. github), onde esteja a documentação e os recursos necessários, devendo na página de entrada identificar claramente o conteúdo das várias pastas. Este repositório deverá ser continuamente atualizado com os resultados do trabalho e será avaliado numa fase intermédia e no final

Prazos e condições de entrega:

- Constituição do grupo (nº e nome dos 3/4 elementos), escolha do tema para o trabalho incluindo um plano de trabalhos preliminar com os dados a usar/ recolher e os objetivos a atingir (máximo 1 página) – submissão no e-learning até dia **7 de abril**.
- Indicação do URL do repositório: **25 de abril**, sendo a avaliação inicial realizada na semana de 27 a 30 de abril
- Apresentação à turma: **15 maio** (10-12 minutos por grupo) – objetivos detalhados; metodologias a usar/ plano de trabalhos; resultados preliminares já obtidos
- Prazo final para atualização dos repositórios: **6 junho,** sendo a avaliação final realizada a partir de 8 de junho; os grupos poderão ser convocados para uma defesa do trabalho nesta fase

Avaliação:

Peso na avaliação:
- Avaliação inicial – 20%
- Apresentação – 30%
- Avaliação final (incluindo defesa se necessário) – 50%

Notem que a avaliação é individual podendo, em casos justificados, haver diferenças de notas entre os vários elementos de um mesmo grupo

Sugestões de conjuntos de dados (opção 1):

Conjuntos de dados relacionados com Covid-19:

- **Kaggle global forecasting**: https://www.kaggle.com/c/covid19-global-forecasting-week-1/overview/ - pretende prever casos em várias partes do mundo, tendo como objetivo principal identificar os fatores mais importantes na previsão podendo ser recolhidos dados de outras fontes (assim, mais do que um grupo poderá escolher este trabalho). Há também outros sites globais como: https://github.com/CSSEGISandData/COVID-19, https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
- Exploração de dados de países individuais e eventualmente suas regiões: exemplo de dados de Itália (e.g. https://github.com/pcm-dpc/COVID-19), Portugal (https://github.com/dssg-pt/covid19pt-data), Espanha (https://github.com/datadista/datasets/tree/master/COVID%2019). Pode escolher-se um país (ou vários) e explorar as suas regiões, procurando correlacionar dados epidemiológicos da doença com outros dados geográficos, demográficos, climáticos, etc. Note que uma pesquisa por exemplo no Kaggle por Covid19 retornará um conjunto de datasets que incluem muitos países (Espanha, Brasil, China, Indonésia, Suiça, India, etc.) e dados epidemiológicos sobre estes, em alguns casos com valores discriminados por região.
- **COVID-19 Open Research Dataset Challenge (CORD-19)** – também disponível no Kaggle; dataset mais recente: https://pages.semanticscholar.org/coronavirus-research. Conjunto de dados de artigos sobre Covid19 – objetivo aplicar técnicas de mineração de textos para diversas tarefas listadas aqui: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks (como há várias tarefas e é um desafio aberto, também aqui pode haver escolha de vários grupos sobre os mesmos dados); outros recursos de mineração de textos: https://www.ncbi.nlm.nih.gov/research/coronavirus/
- Criação de possíveis modelos para classificar compostos antivirais. Deverão ser recolhidos datasets positivos de bases de dados como ChEMBL ou DrugBank e treinados modelos que possam ser aplicados a outros compostos para possível reposicionamento de fármacos. O docente poderá ajudar na definição destes datasets.
- Análise de imagens pulmonares de doentes Covid19 e casos normais. Estes dados podem ser de raios X ou TACs. Para já há poucos dados de Covid19, mas podem treinar-se modelos com outros dados e depois aplicar a Covid19 (*transfer learning*). Alguns dados em: https://www.kaggle.com/theroyakash/covid19; https://github.com/ieee8023/covid-chestxray-dataset
- Dados de Tweet Ids referindo-se a Covid19 - https://github.com/echen102/COVID-19-TweetIDs
- Dados de sequenciação de genomas (mais de 1000 genomas do vírus já disponíveis): https://www.gisaid.org/;

Para além dos referidos aqui, existem muitos outros datasets e recursos sobre Covid19 e o vírus disponíveis e que podem ser propostos pelos grupos de trabalho.

Sites genéricos:

- Kaggle: competições ativas - https://www.kaggle.com/competitions; podem procurar também outros datasets não ligados diretamente a competições
- Competições do site DrivenData: https://www.drivendata.org/competitions/
- UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets (verificar algumas das entradas recentes e com dados de dimensão suficiente)

Existem muitos outros sites e fontes de dados que pode escolher, dando-se apenas a título de exemplo a API do IPMA (Instituto Português do Mar e da Atmosfera) - https://api.ipma.pt/