Trabalho prático de Aprendizagem Automática II – propostas para Opção 1

A. Propostas na classificação de compostos – Apoio: João Correia

Proposta A.1 - Desenvolvimento de uma abordagem de AutoML integrada numa ferramenta de machine learning existente (DeepMol).

Os resultados alcançados, nas últimas décadas, pelo uso de abordagens de inteligência artificial em áreas como visão computacional e processamento de linguagem natural levaram a um uso mais generalizado dessas abordagens em outras áreas. Um desses casos passa pela aplicação de abordagens de Machine e Deep Learning em tópicos de quimioinformatica onde o objetivo passa pela identificação e design de moléculas com propriedades específicas. No entanto, por vezes, a utilização deste tipo de abordagens não se revela tarefa fácil para utilizadores sem background computacional.
Desta forma, o objetivo deste projeto passa pela implementação e integração de uma abordagem de AutoML numa ferramenta de machine learning de classificação de compostos. A ideia passaria por implementar um módulo através do qual seja possível automatizar alguns dos processos da pipeline e comparar os resultados obtidos por diferentes abordagens tendo em conta um set de inputs cedidos inicialmente.
Inicialmente, as tarefas passariam por explorar a pipeline existente usando um dataset como case study***. Depois de perceber a estrutura e organização da pipeline iria proceder-se à implementação e integração do módulo de AutoML.
O projeto será desenvolvido usando a linguagem Python.

Proposta A. 2 - Desenvolvimento de abordagens para a criação de embeddings moleculares baseados em algoritmos de NLP para integração numa ferramenta de machine learning existente (DeepMol).

Os resultados alcançados, nas últimas décadas, pelo uso de abordagens de inteligência artificial em áreas como visão computacional e processamento de linguagem natural levaram a um uso mais generalizado dessas abordagens noutras áreas. Um desses casos passa pela aplicação de aborgadens de Machine e Deep Learning em tópicos de quimioinformatica onde o objetivo passa pela identificação e design de moléculas com propriedades específicas. Para isto, o processo de implementação de métodos que sejam capazes de criar representações vetoriais que caracterizem substruturas moleculares assume grande importância.
Hoje em dia existem múltiplas técnicas baseadas em modelos de NLP capazes de eficientemente criar embeddings moleculares a partir de representações como SMILES. Exemplos disso são os algoritmos Mol2vec, Seq2seq Fingerprint, SMILES Transformer, Message Passing Neural Networks for SMILES, SMILES-BERT entre muitos outros.
Desta forma, o objetivo deste projeto passa pela implementação e integração de um módulo de abordagens baseadas em algoritmos de NLP para a criação de embeddings moleculares numa ferramenta de machine learning de classificação de compostos.
Inicialmente, as tarefas passariam por explorar a pipeline existente usando um dataset como case study***. Depois de perceber a estrutura e organização da pipeline iria proceder-se à implementação e integração do módulo para criação de embeddings moleculares.
O projeto será desenvolvido usando a linguagem Python.

***Datasets propostos:

Dataset	Data Type	Task Type	Compounds	Link
HIV: Experimentally measured abilities to inhibit HIV replication	SMILES	Binary Classification	41127	https://github.com/GLambard/Molecules_Dataset_Collection
BBBP: Binary labels of blood-brain barrier penetration(permeability)	SMILES	Binary Classification	2039	https://github.com/GLambard/Molecules_Dataset_Collection
BACE: Qualitative binding results for a set of inhibitors of human β-secretase 1(BACE-1)	SMILES	Binary Classification	1513	https://github.com/GLambard/Molecules_Dataset_Collection
Bitter: Bitter or not?	SMILES	Binary Classification	2608	https://github.com/Niv-Lab/BitterPredict1
Tox21: Qualitative toxicity measurements on 12 biological targets	SMILES	Binary Classification (12 Tasks)	7831	https://github.com/GLambard/Molecules_Dataset_Collection
ClinTox: Qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons	SMILES	Binary Classification (2 Tasks)	1478	https://github.com/GLambard/Molecules_Dataset_Collection
ChEMBL benchmark dataset	SMILES	Binary Classification (1000+ Tasks)	456331	http://www.bioinf.jku.at/research/lsc/index.html

B. Propostas na classificação de proteinas – Apoio: Ana Marta Sequeira

Proposta B.1 - ML/DL with DNA using ProPythia

- Implement DNA descriptors from literature
- Perform a ML/DL pipeline using a case study

Machine learning methods have been applied to a huge variety of problems in genomics and genetics and can be used in a variety of problems such as identification of transcription sites, splice sites, TF binding sites or gene identification [1]. Most common approaches include the use of sequential data (hot encoded sequence) [1,2]. However, an alternative approach can be the calculation of descriptors for these sequences. Some packages have already implementations of these descriptors: Pybiomed [3] (example application predicts nucleosome positioning in genomes), Bio-Seq analysis [4,5] (example application is the identification of enhancers and identification of DNase I hypersensitive (DHS) prediction) and iLearn [6] (example application is the identification of m5C sites).

Propythia is a package developed to do machine and deep learning classification using protein sequences. The goal of this project is to implement the DNA descriptors in this package and perform a ML/DL pipeline using a dataset as a case study.

[1] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/
[2] Eraslan, G., Avsec, Ž., Gagneur, J. *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 20, 389–403 (2019). https://doi.org/10.1038/s41576-019-0122-6
[3] Dong, J., Yao, ZJ., Zhang, L. *et al.* PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *J Cheminform* 10, 16 (2018). https://doi.org/10.1186/s13321-018-0270-2
[4] Bin Liu, Xin Gao, Hanyu Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, *Nucleic Acids Research*, Volume 47, Issue 20, 18 November 2019, Page e127, https://doi.org/10.1093/nar/gkz740
[5] Bin Liu, BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, *Briefings in Bioinformatics*, Volume 20, Issue 4, July 2019, Pages 1280–1294, https://doi.org/10.1093/bib/bbx165
[6] Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, Zhu Y, Powell DR, Akutsu T, Webb GI, Chou KC, Smith AI, Daly RJ, Li J, Song J. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2020 May 21;21(3):1047-1057. doi: 10.1093/bib/bbz041. PMID: 31067315.

Proposta B.2 - Word embeddings with proteins

- Implement a word embedding class applied to proteins
- Apply this to a case study using DL

Adopting natural language processing (NLP) to discover functions encoded in biological sequences was a strategy firstly described by Asgari et al. [1].
In NLP, each word is embedded in a vector in an n dimensional space (similar words have closer vectors) and the meaning of this word being characterized by its context. In proteins case, the sequence can be separated in several words (for ex. trigrams of aminoacids), map this continuous representation into an embedding layer and train a RNN model. Using word embeddings to describe protein sequences has obtained good results in protein family classification [1], transporters classification [2], property prediction [3], protein receptor identification [4], identification of antimicrobial peptides [5] among others.
Propythia is a package developed to do machine and deep learning classification using protein sequences. The goal of this project is to implement a word embedding class for proteins and perform a deep learning pipeline using a dataset as a case study.

[1] Asgari E, Mofrad MRK (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLOS ONE 10(11): e0141287. https://doi.org/10.1371/journal.pone.0141287
[2] Nguyen TT, Le NQ, Ho QT, Phan DV, Ou YY. Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. Anal Biochem. 2019 Jul 15;577:73-81. doi: 10.1016/j.ab.2019.04.011. Epub 2019 Apr 22. PMID: 31022378.
[3] Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 2018 Aug 1;34(15):2642-2648. doi: 10.1093/bioinformatics/bty178. Erratum in: Bioinformatics. 2018 Dec 1;34(23):4138. PMID: 29584811; PMCID: PMC6061698.
[4] Le NQK, Huynh TT. Identifying SNAREs by Incorporating Deep Learning Architecture and Amino Acid Embedding Representation. Front Physiol. 2019 Dec 10;10:1501. doi: 10.3389/fphys.2019.01501. PMID: 31920706; PMCID: PMC6914855.
[5]Hamid MN, Friedberg I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics. 2019 Jun 1;35(12):2009-2016. doi: 10.1093/bioinformatics/bty937. PMID: 30418485; PMCID: PMC6581433.

C. Propostas em text mining / COVID – Apoio Rúben Rodrigues e Nuno Alves

Proposta C. 1 Detecção de notícias falsas

A desinformação presente nos meios de comunicação social tem vindo a ser um problema à escala mundial. Apesar dos profissionais de saúde promoverem fontes de informação fidedignas sobre doenças e tratamentos, as redes sociais, por vezes, disseminam informação falsa que prejudica a promoção da prevenção e tratamento de uma doença. Recentemente, a vacinação contra COVID tem sido um tema muito debatido e as notícias presentes nas redes sociais têm um impacto constante na promoção da vacinação.

Neste trabalho pretende-se que os alunos integrem métodos de deep learning capazes de identificar notícias falsas sobre COVID com recurso à framework desenvolvida em python no grupo BioSystems (BioTMpy). Os datasets de auxílio a serem utilizados para o desenvolvimento da ferramenta encontram-se nos links: https://github.com/cuilimeng/CoAID, https://www.kaggle.com/c/fake-news/data, https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php.

Proposta C.2 - Previsão do número de infetados/mortes de doenças pulmonares

A criação de modelos de previsão sobre uma doença pode auxiliar uma melhor resposta no combate à doença. Atualmente a pandemia COVID tem sobrecarregado os sistemas de saúde consumindo a maioria dos seus recursos. Porém existem outras doenças pulmonares como a pneumonia que continuam a provocar mortes.
Neste trabalho, pretende-se que os alunos desenvolvam uma ferramenta, com recurso a métodos de deep learning e à linguagem Python, que seja capaz de prever o número de infectados e mortos provocados por doenças pulmonares. A previsão será feita com base nos dados fornecidos pelas entidades de saúde sobre o número de pessoas com doenças pulmonares ao longo de um período de tempo.
Os datasets de auxílio a serem utilizados para o desenvolvimento da ferramenta encontram-se nos links:
https://healthdata.gov/dataset/provisional-death-counts-influenza-pneumonia-and-covid-19, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Proposta C.3 - Identificação de documentos similares/relevantes para cancro do estômago

Repositórios como o PUBMED contêm mais de 30 milhões de documentos que estão indexados a termos chave. Estes repositórios possibilitam a identificação de documentos relevantes sobre um determinado termo. Porém existem documentos que são relevantes a um determinado tópico, mas que não estão devidamente identificados.
Como solução, existem abordagens baseadas em aprendizagem máquina não supervisionadas como a Latent Dirichlet Allocation (LDA) em que é possível aplicar o cálculo da distância Jensen-Shannon para identificar documentos similares a um determinado tópico. Exemplos de pesquisas que podem beneficiar de um sistema orientado a similaridade de documentos incluem a pesquisa sobre uma determinada doença como o cancro do estômago e um tratamento que não foi indexado pelo repositório.
Neste trabalho, pretende-se que os alunos implementem uma pipeline em python que calcule a similaridade de documentos recorrendo ao LDA e à fórmula de Jensen-Shannon e integrem a pipeline desenvolvida na framework do grupo BioSystems (BioTMpy).
Os dados fornecidos para este trabalho estarão no formato JSON e serão disponibilizados a partir da plataforma de conhecimento sobre cancro gástrico desenvolvido no grupo BioSystems.

Proposta C.4 - Identificação de evidências similares/relevantes para cancro do estômago

A descoberta de evidências textuais sobre tratamentos de cancro do estômago em documentos biomédicos é um processo moroso. Existem métodos de processamento de linguagem natural (NLP) que recorrem a abordagens de aprendizagem não supervisionadas para identificar similaridade entre frases que contenham evidências biológicas. Implementações como o Doc2Vec, SentenceBERT, InferSent e Universal Sentence Encoder, requerem um conjunto de documentos devidamente processados para construir um modelo não supervisionado. A partir destes modelos é possível agrupar e identificar frases similares, permitindo a criação de um sistema que dado uma frase de entrada (*query*) possibilita a descoberta de frases similares que estão presentes nos documentos processados.
Neste trabalho, pretende-se que os alunos implementem uma pipeline em python que faça uso de ferramentas de NLP para processar documentos de cancro de estômago. Esta pipeline tem o propósito de criar modelos capazes de identificar frases similares mantendo o mapeamento dos documentos com as frases identificadas.
Os dados fornecidos para este trabalho estarão no formato JSON e serão disponibilizados a partir da plataforma de conhecimento sobre cancro gástrico desenvolvido no grupo BioSystems.

D. Propostas com apoio do Vítor Pereira

Proposta D.1

Deep learning methods, and in particular deep generative models, such as Autoencoders (AEs) and Generative Adversarial Networks (GANs), have been gaining popularity in drug discovery, aiming to accelerate the discovery of new therapeutical molecules while reducing costs.
In our research group, we have developed a framework, DeepMolGen, that allows to use some of these methods to create new molecules that can have interesting properties.
The aim of the project is to, given a training set of molecules, implement a latent based GAN able to generate new ones. Variants of the original GAN, proposed by Goodfellow et. al, may be considered, such as Wasserstein GAN (WGAN) and Least-Squares GAN (LSGAN)), that use distinct probability divergence functions and/or additional methods (e.g., gradient penalty (WGAN-GP) and spectral normalization (SN)) to improve the GAN's training. The objective is to add at least one of these approaches to DeepMolGen and apply to a given case study in molecule generation.

Proposta D.2

Network Function Chaining (NFC) is present in many of the tasks performed on communication networks, such as firewalls, deep packet inspection (DPI), video encoding and decoding, traffic load balancing. These functions, implemented on virtualized environments dispersed in the network, consume computational and forwarding resources that need to be optimized. As such, when new requests enter the network, careful decisions are made to select the best node or nodes where they should be processed to keep the infrastructure working properly. Decision-making involves optimization tasks that take too long to be conducted online. Machine learning classifiers present a good solution to overcome this obstacle by learning the intrinsic relations between the network state, requests and the selected network nodes. The aim of the project is to develop a classifier able to select

the best processing nodes for each new request. A dataset with traces, network states, and labels will be provided to train the classifier.