Machine Learning in Biology

In this class, we will discuss the main concepts of Machine Learning and its applications to different bioinformatics and biomedical data science problems.

Learning Objectives

- Main concepts of Machine Learning including supervised and unsupervised learning
- Application of ML to gene expression analysis
- Neural Networks and Deep Learning architectures
- Applications to sequence analysis and image analysis in disease detection.
- Limitations and strengths of ML

1. Review the slides “ML”.

Quiz

Task 1 –

1. If you are predicting house prices, you are probably doing ... ?
- Reinforcement Learning
- Regression
- Classification
- Clustering

2. What are the main types/applications of Supervised Learning?
- Reinforcement and Unsupervised
- Classification, Labeling and Clustering
- Regression and Classification
- Unsupervised, Labeling and Regression

3. In ___ we try to find the decision boundary, which can divide the dataset into different groups.
- Classification
- Machine Learning
- Reinforcement Learning
- Regression

4. _____ Learning can be used for those cases where we have labelled input data!
- Supervised
- Unsupervised
- Machine
- Reinforcement

Discuta os seguintes cenários com os elementos do seu grupo e proponha soluções para a sua análise.

Task 2

In order to be used in analysis with Deep Learning methods biological sequences need to be converted into a suitable format. **One-Hot-Encoding (OHE)** allows a sequence of categorical values (e.g. nucleotides) to be converted into a binary/numeric format. Write a function that given a DNA sequence does the OHE conversion of it, *dna_ohe(seq*).

A -> [1, 0, 0, 0]
C -> [0, 0, 1, 0]
G -> [0, 0, 1, 0]
T -> [0, 0, 0, 1]

Task 3

Consider that you have a set of protein sequences (identified by their name) and you want to know how they are distributed and what kind of relationships and structure they have between each other. Assume that besides the sequence you have **no** information about the nature of the sequences (e.g. protein family, presence of functional motifs, species, etc.).

i) Write a function that given a sequence returns a data structure that contains the frequency of each k-mer in the sequence. Function: *word_to_kmer(word, k).*
ii) Write a function that given a file of sequences in fasta format, returns a table with the frequencies of all k-mers in each of the sequences: *file_to_kmer_table(file_name)*. You can return a table with one of the objects presented in class or a pandas-like table.

Task 4

Consider a scenario in which you have a set of proteins previously classified into 10 family types. The idea would be to be able to assign the new sequences that are being determined to one of the 10 types of families already known. Assume also that in this case, we do not have the sequences but rather a set of indicators about the composition of the sequences, such as: size, GC%, presence/frequency of more than 100 types of motifs from the PFAM database, number of alpha helices, beta sheets, etc...

Describe what kind of Machine Learning based approach you would use for each analysis above. Consider the following points:
- nature of the learning/analysis problem;
- objective;
- input data and output data?
- methodology for defining the relationship between sequences (measurements, distances, score functions, etc....);
- consider how your proposal would scale to n =100, 1000, 10000, ...
- other points that you consider relevant.