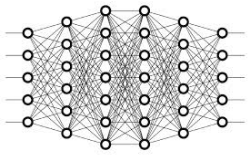
# Statistical Learning with Deep Artificial Neural Networks

Task 1



## Context and Data

Recent technological advances and international efforts, such as The Cancer Genome Atlas (TCGA), have made available several pan-cancer datasets encompassing multiple omics layers with detailed clinical information in large collection of samples. The need has thus arisen for the development of computational methods aimed at improving cancer subtyping and biomarker identification from multi-modal data.

In https://www.frontiersin.org/articles/10.3389/fonc.2020.01065/full three multi-modal cancer datasets generated by The Cancer Genome Atlas (TCGA) Research Network (https://www.cancer. gov/tcga) are considered. Protein abundance, gene expression and copy number variants are used to predict breast invasive carcinoma (BRCA) estrogen receptor status (0: negative; 1: positive).

In this task we ask you for integrating gene expression, protein abundance and clinical data. Data sets are available in `Breast.zip` file.

## Questions

1. Describe protein abundance and gene expression datasets. How many patients have data of both types available. Are there missing data from some of the datasets? Preprocess them if necessary.

With gene expression data.

2. Select the 25% of genes with the most variability.
3. Implement an stacked autoencoder (SAE) with three stacked layers of 1000, 100, 50 nodes. Provide in each case evidence of the quality of the coding obtained.
4. Using the SAE as pre-training model, couple it with a two-layer DNN to predict the state of the estrogen receptor. The DNN must have 10 nodes in the first layer followed by the output layer.
5. On the test set, provide the ROC curve and AUC and other performance metrics.
6. With `tfruns()` repeat points 4 and 5, exploring the configurations of the first layer of the DNN based on 5, 10 and 20 nodes. Determine which configuration is the best.

So far, we have two SAEs. One for the abundance of proteins (see class examples) and the other for gene expression we just built.

7. Split the set of patients with complete data (gene expression and protein abundance) in train and test sets.
8. Concatenate the two SAEs to fit, on the trainset, a DNN that integrates both data sources to predict estrogen receptor status. The DNN must have a dense layer (with the better number of nodes according with point 6) and the output layer.

9. On the testset, provide the ROC curve and AUC, and compare it with the model found in point 5.
10. Discuss the results of the analysis.

## Important remarks

- Answer the questions in a reasoned way, adding the necessary comments, not just only the code.
- A R markdown report as dynamic as you can.
- Use relative paths instead of absolute paths to read/write files, to make it easier to run the code outside of your computer.

## Delivery / Deadline

A zip file including:

- the Rmd file used as template for the report,
- the output reports in pdf and/or html files.

**Deadline**: April 10th, 2022.