PROJECT 7

Project name: Classification of normal vs tumor samples

Programming language(s): Python

Short description: Analysis of the TCGA gene expression data with classification techniques

Expected outcome: Download data of TCGA GRCh38 Breast Cancer gene expression data and analyze it.

Provide a detailed self-documented Jupyter Notebook:

1) Analyze the input file, such the percentage of the healthy and tumoral samples, ...
2) Split the data into train and test sets. If you need a validation set, further split the train into train and validation sets.
3) Run different classification techniques and compare them by using different evaluation metrics (accuracy, precision, recall, …).
4) Optional: Use k-fold cross validation for the analysis.

Hints/Notes:

The GenoSurf(http://geco.deib.polimi.it/genosurf/) interface can be used to download the related data. You need to select:

- project_name: ['tcga-brca']
- assembly: ['grch38']
- data_type: ['gene expression quantification']
- is_healthy:
    - ['false'] for tumoral data
    - [true] for normal data.

You can use DOWNLOAD LINKS in RESULTS ITEMS tab for downloading the samples.
Each sample has one region file (with extension ".gdm") and one metadata file (with extension ".meta.gdm") as a couple. The columns of the GRCh38_TCGA_gene_expression_2019_10 dataset is as below:

1. chromosome STRING
2. start LONG
3. end LONG
4. strand CHAR
5. ensembl_gene_id STRING
6. entrez_gene_id STRING
7. gene_symbol STRING
8. type STRING
9. htseq_count LONG
10. fpkm_uq DOUBLE
11. fpkm DOUBLE

You can use *ensembl_gene_id* as header and *fpkm* as value. And then you can pivot it to create a single matrix.

Literature/Resources:

Recommend using the methods that you learned in the Machine Learning course.