Algoritmos Bioinformática /Bioinformática 2024/2025

Group Assignment 2

Cancer subtype classification

1. Background

In this assignment, you will explore the challenges and opportunities of classifying molecular subtypes of **Glioblastoma Multiforme (GBM)**, one of the most aggressive brain tumors. In this assignment, you will explore **both supervised and unsupervised machine learning techniques** to analyze high-dimensional gene expression (Gexp) data. Your focus will be on distinguishing between two molecular subtypes of cancer: **Classical** (prognostically favorable) and **Mesenchymal** (prognostically adverse). These subtypes are associated with different patient outcomes, and accurate classification is essential for prognosis and the development of personalized therapeutic strategies.

- **Unsupervised learning -** Apply unsupervised learning to explore whether the intrinsic structure of the gene expression data naturally reflects the known cancer subtypes. Use dimensionality reduction techniques (e.g PCA or UMAP) for visualization and clustering algorithms (e.g K-means, hierarchical clustering). Are the subtypes of cancer (**Classical** and **Mesenchymal**) recapitulated using these methods? Generate visualizations using the results from these methods.

- Build a binary classifier capable of distinguishing between the **Classical** (prognostically favourable) and **Mesenchymal** (prognostically adverse) subtypes. A possible pipeline for this task is outlined below and should guide you in building and evaluating your subtype classification model using gene expression data:

    1. Data splitting: Since the dataset has a relatively low number of samples, the whole pipeline should be evaluated with a cross-validation scheme. This means all steps of the pipeline (including pre-processing and feature selection) should be performed on the training folds and applied to the test/validation fold. For computational reasons keep the value of K (number of folds) under 5.

    2. Data preprocessing: Consider applying data transformation such as feature scaling to the data (e.g using the **StandardScaler from scikit-learn package).**

    3. Feature selection: The provided dataset contains 5000 features. Consider performing feature selection selecting the most informative features (e.g

minimal variance, Anova F-test or Correlation with the target variable). Keep in mind the **StandardScaler** will change the scale of your data, and some methods might not be applied.

4. Model training: Train at least 2 types of classifiers (e.g SVM, RandomForest, Logistic Regression, NaiveBayes, etc). Experiment with different hyperparameters combinations. Evaluate the model performance using multiple metrics (accuracy, precision, recall and F1-score). The performance should be the average + standard deviation across the 5 folds.

5. Write a small discussion on the results. Remember that there is no *a priori* expectation on the results, so these will be totally novel results requiring interpretation.

2. Data Description

You will receive two files:

- **Gene Expression data:** A matrix representing mRNA levels across protein-coding genes. Values are normalized counts from RNA sequencing (logTPMs).
- **Labels:** a tab delimited file containing the labels "Mesenchymal" or "Classical" for each cancer sample. Classical are coded as 0 while Mesenchymal samples are coded as 1.

For this assignment, you should submit:

- Report – a small report with 2-3 pages, describing your approach, the tools and packages used and the main results. You should include the table with the results and finish with a conclusion. Submit as a .**pdf** file. You can use the same templates as group assignments 1 to build your report.
- Code – submit the developed code in a python notebook. Note that your code should be reproducible. Include dependencies on external packages in a requirements.txt file

Submit a **zip** file with **2 files**: report and code as described above.

Do not forget to mention in the report all the students who contributed for the assignment and any particular notes on their specific contributions.