Institute of Biochemistry and Biophysics, Department of Bioinformatics

Project 1

Detection of Tumor Type Using RNA-seq Data from Tumor-Educated Platelets

Background:

Cancer remains one of the most challenging diseases to diagnose and treat, with many types progressing silently until advanced stages. Traditionally, tissue biopsies and imaging techniques are used to diagnose different cancer types, but these methods can be invasive, expensive, and prone to complications.

However, recent research has demonstrated the potential of blood-based diagnostics, specifically by analyzing RNA from tumor-educated platelets (TEPs). Platelets play a critical role in blood clotting, but they are also affected by the presence of tumors in the body, taking on distinct RNA profiles. These changes can be detected using RNA-sequencing (RNA-seq), offering a non-invasive method to classify different cancer types from a simple blood draw.

The dataset used for this project contains RNA-seq data from 284 blood platelet samples, including samples from patients with six different types of tumors (breast cancer, hepatobiliary cancer, colorectal cancer, glioblastoma, non-small cell lung cancer, and pancreatic cancer) and healthy individuals. The RNA-seq data has been processed and normalized, providing counts that reflect gene expression levels in each sample. By analyzing this data using machine learning models, we aim to build predictive models to detect the presence of cancer and classify patients by tumor type.

1. Exploratory Data Analysis (EDA)

The first step in any machine learning project is to explore and understand the dataset. In this task, you will perform exploratory data analysis (EDA) on the cancer datasets and get ready for training machine learning models.

Tasks:

1. Class Distribution Visualization:
   - Plot a bar chart showing the number of samples for each class (cancer types and healthy controls).
   - Identify any class imbalances. Discuss how such imbalances might affect the performance of machine learning models.
   - For any classes that contain more than one mutational subclass, plot a bar chart showing the distribution of these subclasses
2. Gene Expression Distribution:
   - Identify the top 10 genes with the highest variance in expression across all samples.
   - Plot a boxplot to visualize the distribution of these 10 genes.
   - Discuss whether these genes might play an important role in distinguishing between the different cancer types. Do you think these genes will be helpful in the classification task? Why or why not?

---

2. Binary Classification - Tumor vs. Healthy Control

In this task, you will perform a binary classification to distinguish between tumor patients and healthy controls using RNA-seq data. You will train multiple classification models and evaluate their performance.

Tasks:

1. Data Splitting:
   - Split the dataset into training and testing sets. Use Cross-Validation for the split (e.g., K fold cross-validation).
2. Model Training:
   - Select 4 different machine learning algorithms (e.g., KNN, Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, etc.).
   - Train each of these models on the training data to classify patients as either "Tumor" or "Healthy Control."

3. Model Evaluation:
    - Evaluate each model using the following performance metrics:
        - Accuracy
        - Precision
        - Sensitivity (True Positive Rate)
        - Specificity (True Negative Rate)
    - Compare the models based on their performance in each of these metrics. Discuss which model performs the best overall and why.
4. Model Optimization:
    - Use either Grid Search or Random Search in combination with nested Cross-Validation to tune the hyperparameters of one model of interest.
    - After optimization, reevaluate the model on the testing set and compare its performance with the non-tuned model.

---

3. Multi-Class Classification - Predicting Tumor Type

This time, you will perform a multi-class classification to predict the specific type of tumor among the cancerous samples. Follow a similar approach as in the binary classification task and report your results.

Keynotes:

You are required to create a comprehensive report that includes your code, results, and answers to the provided questions. The report should also contain brief explanations of your code and any relevant analysis. Send your answers as a zip file via email to Mahdi.Anvari7@ut.ac.ir by **November 5, 2024**. Make sure to include your name and student number in the email subject line.

You can research your answers online or using textbooks, and discuss your solutions with your classmates, but you need to disclose all the resources that you used in your report. If you use tools like ChatGPT, include your prompt and the answer in your report.

Good luck