Università degli Studi di Torino - M.Sc. in Stochastic and Data Science - A.Y. 2021/2022

Final project of Statistical Machine Learning (MAT0043)

Gene selection for cancer type classification

In recent years, medicine has made a great step forward in finding new and efficient therapies for different diseases. Thanks to up-to-date technologies, collecting huge amount of data is no longer an issue, so that one can exploit them to define personalised treatments for patients. In particular, cancer genome scale screens are just one example of these applications. In particular, they provide valuable information about the role of genes in driving cancer growth. Thus, researches has developed a Cancer Dependency Map in order to identify genetic and pharmacologic dependencies. However, this is quite a challenging aim: the dataset is not at all easy to handle (high-dimensional, over than ~ 17.000 features) and the picked drug-targetable genes should only rely on a specific cancer type, thus imbalanced classes.
In this project, we apply cutting-edge Statistical Machine Learning algorithms to classify different cancer types and select the most relevant genes. After a quick exploratory analysis with PCA, we try Random Forest, Lasso-SVM and Neural Network classifiers and see how the same technique performs differently according to which tumour we are focusing on. In fact, our classification accuracies range from 45% to 98%. **Aggiungere altre conclusioni**

1 Introduction

Scrivere che il cancro e' una malattia molto brutta e brevemente come funziona (aka le cellule impazziscono: svillupano mutazioni, diventano imprevedibile e causano problemi negli individui...). Il compito della medicina e' trovare delle cure (banana). E qua entra in gico DepMap: DNA arrays che raccolgono mutationi dei geni delle cellule cancerogene.

Developing new cancer therapies is based on finding processes that selectively kill cancer cells. In particular, the Achilles project uses genome-wide screens to systematically identify essential genes and report vulnerabilities across hundreds of human cancers.

2 Dataset

We use two publicly available datasets, both found on the DepMap Portal website:

- *CRISPR_gene_dependency.csv*, which contains 1.032 cancer cell lines characterised by 17.393 gene scoring results;
- *sample_info.csv*, which contains cell lines information, such as primary disease and sample collection site.

Data were collected from real patients and successively processed, so that each element of this (1.032×17.393)-matrix is the probability that knocking out a gene has a real depletion effect on the cell.
First of all, we look for any missing values: only 10 rows have empty columns, specifically either 678 or 1.285 Nas. At first impact, this could seem a big deal, but is actually the 4% and 7% of the total genes. Moreover, these cells come from different tumours, so we decide to simply remove them all.
Before proceeding with our analysis, we also notice some weird observations: 2 of them are labelled as "Non-Cancerous" and 6 as "Engineered". The first can be reasonably discarded as our goal is classifying cancer cells, whereas the latter requires a little care. Engineered cells are synthetically modified sample in lab and, here, they are manly associated

to the Eye sample collection site. We keep them and we associate the cancer type according to the sample.
One could ask: why do we focus only on the primary disease and not also on the sample collection site? If we count the number of cells with respect to cancer type and collections site, we find many peculiarities. For instance, some Brain-cancer cells have been picked from the abdomen, whereas Lung-cancer cells comes from a variety of different places. This is because of the nature/curse of cancer: metastasis are ill cells identifiable as the original tissue but found on a different site. Therefore, this subdivision would only complicate our task.
Finally, we group the various cancer types in 10 classes according to common medical knowledge (https://www.cancer.gov/types/by-body-location)
and we obtain the following classes:

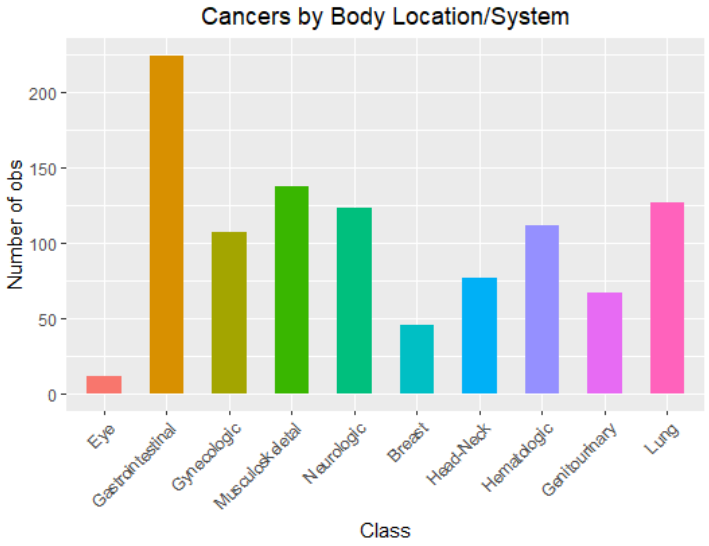

Figure 1: Cancer classes

Figure 1 shows our classes.

3 Methods

4 Results

5 Conclusion and future works

References

[1] DepMap Portal: https://depmap.org/portal/

[2] DepMap, Broad (2021): DepMap 21Q3 Public, figshare. Dataset: https://doi.org/10.6084/m9.figshare.15160110.v2

[3] Project Achilles: https://depmap.org/portal/achilles/