Bioinformatics for Translational Medicine
Group Assignment

Classification Assessment of Tumor Subtypes

April 4, 2022

1 Introduction

Breast cancer is a heterogeneous disease and classification of breast cancer tumors in their molecular subtypes has important implications for treatment and prognosis. Three receptors play a pivotal role in these subtypes: the Estrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal growth factor Receptor 2 (HER2). After removal of the tumor, the pathology department of the hospital tests these samples for presence of ER, PR and HER2.
The three main subtypes in breast cancer on which the treatment decision will be based are:

- HER2 positive: HER2+
- Hormone receptor positive (HR+): ER+ and/or PR+, and HER2-
- Triple negative (TN): ER-, PR- and HER2-

Each of the three subtypes reacts differently to different types of treatment.

In this group assignment, you will train a classifier that is able to predict subtypes of breast cancer tumors based on array CGH (aCGH) data.
First, you will build a model in R or Python and make an estimate of the accuracy of its predictions using a dataset of 100 aCGH samples with their associated clinical outcomes (subtypes) that we have provided for you. Then, you will use this model to make predictions for an independent test set for which the subtypes are unknown (57 samples). Finally, you will explain your approach in a report and a presentation, and perform peer review on the draft report of one of the other groups.

2 Deliverables

Assessment of this assignment is based on the following four components:

- Report written in the style of a short paper [70%]
- Peer review of the draft report of another group [pass/fail]
- Predictions + accuracy estimate (+ R/Python scripts, model file) [15%]:
  - Format correct of predictions + model file [5%]

  - Quality of predictions [5%]
  - Accuracy estimate [5%]
  - Full training and validation code - [pass/fail]
- Presentation [15%]

Please see Canvas for an up-to-date schedule and hand-in dates for each part of the assignment.

3 Building your classifier

Your task is to build a well-trained classifier for predicting the three breast cancer subtypes. To achieve this, you may need to perform some of the following steps.

1. Data purification and transformation, if necessary.
2. Feature selection, if necessary.
3. Choosing machine learning methods (classifiers).
4. Training and validation of the classifiers.

3.1 Data

We provide you with a dataset with 100 breast cancer samples from the three subtypes. These samples are analyzed on a high-resolution array CGH platform with 244,000 probes per array that measures the quantity of chromosomal DNA. The pre-processing of the data has been done for you.
For each of these regions and each sample we give the call whether that region is a gain, an amplification, a loss or normal (-1 for loss, 0 for normal, 1 for gain, and 2 for amplification of the DNA).
Two files are provided, one containing the preprocessed aCGH data of the cancer samples (`Train_call.txt`) and the other containing the associated clinical outcome of these samples — the subtypes (`Train_clinical.txt`).

3.2 Software

You are recommended (but not limited) to use R (3.3.2+) or Python (3.5+).

3.2.1 R

A number of machine learning R packages are available, such as `CORElearn`, `RWeka` and `caret`. You need to refer to the official documentation for usage.
If you are unfamiliar with R, you could start your work with the tutorial (`CATSRTutorial.pdf`) we have written for you, and we also advise you to use `caret` because we are able to provide some support for it.
You are required to save your trained classifier into a file by using `saveRDS`. This model file (`*.rds`) together with your R script needs to be submitted.

3.2.2 Python

If you opt to use Python, the Python library `scikit-learn` is recommended for the machine learning part. You may also need `NumPy` or `pandas` to transform the data; you can refer to the official documentation of these packages.

- scikit-learn: Quick start
- NumPy: the absolute basics for beginners
- pandas: 10 minutes to pandas

You are required to save your trained classifier into a file by using the `joblib` function of `scikit-learn`. This model file (`*.pkl`) together with your Python script needs to be submitted.

4 Classifier assessment

In order to test how well your classifier performs, you will be given another 57 samples for which you will need to predict their subtypes (HER2+, TN or HR+). These samples will be given to you towards the end of the assignment. In addition, you will need to estimate how many samples you classified correctly, which means you need a good benchmarking scheme to support your estimate.

To make it clear here, you need to submit

1. The **predicted labels** for each of the samples (you need to follow the file format on Canvas) saved as `prediction.txt`.
2. An **estimate** for the number of correctly labeled samples (out of 57), saved as a `.txt` file containing only a number (`estimate.txt`).

5 Report

For your (short) research paper, you need to choose a research question. We ask you to submit this question early on in the course, mainly to ensure that you are on the right track. Make sure to get feedback on this, before you write your draft. Please have a look back at the writing lecture from the course Fundamentals of Bioinformatics.
You should follow the Bioinformatics guidelines for writing an 'original research paper': `http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/general.html`
Please find Word and LaTex templates here: `http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/submission_online.html`
In addition, you should follow the outline below:

5.0.1 Contents of paper (5-6 pages A4)

Abstract (max. 200 words)
- Motivation
- Results and impact

Introduction
- Include references to other papers
- Explain background of the data set (both experimental & preprocessing)
- Explain the context (biomedical) of the research
- Explain if any similar bioinformatics approaches have previously been described
- Set out the research question you are trying to answer in this paper

Methods
- Explain which methods you have used to build your classifier

- Explain how you have cross-validated your data

Results

- Explain how well your classifier performs, based on your own benchmarking
- You may compare multiple classifier approaches, or multiple feature selection methods.
- If you were to mark a single region (biomarker) for classification, what would it be, and how well would you do?

Discussion and conclusion (should be kept concise: no more than half a page)

- Discuss any issues that may affect the results
- Discuss the single best region you found (see above)
- Clearly state which research question you have answered in this paper
- Explain what impact your research has on future research

Tables and figures

- around 2-4 (in total) figures and tables
- Make sure to explain all axes, labels, lines and points in the caption of your table / figure
- Make sure to refer to each table / figure in the main text, and explain in the main text what can be seen from the figure

Author contributions

- Describe what each group member contributed to the project. See e.g. this PLOS example.

5.0.2 Grading of draft report (5-6 pages A4)

Your draft will be peer-reviewed by 4 other students. They will score you on the elements listed below, and give suggestions on how to improve the draft.

- Context
- Research question
- Methodology
- Contents/results
- Tables/figures

scoring: good / satisfactory / sufficient / insufficient

5.0.3 Grading of final paper

The grade for the final report will be based on the rubrics for the master research project, which you can find on Canvas.

scoring: 1-10

6 Presentation

For the presentation you should prepare exactly 4 slides per group:

1. Cross validation/ testing scheme to estimate accuracy
2. (Comparative) table of accuracy estimate(s) for submitted methods (and others if you wish)
3. Methods (feature selection and classification) used for submitted prediction

4. Rationale why you think the submitted method performed best

The presentation should last no longer than 8 minutes. Slides must be submitted via Canvas.

7 Submitting your classifier and report

7.1 Report and presentation

Four submissions should be made via Canvas:

1. Draft of a manuscript describing your research
2. Peer review on the draft of another group
3. The final manuscript
4. Presentation slides

7.2 Code and output files

Submit your scripts and output files via Canvas in a folder in `.zip` or `.rar` format. The name of the folder should be `GroupXX`, e.g. `Group05.zip`. The compressed folder must contain the following directories:

`results/`
- `estimate.txt`: a file containing only a number (0-57)
- `prediction.txt`: predictions following the format as specified on Canvas

`model/`
- `model.rds` or `model.pkl`: a file containing your model
- `run_model.R` or `run_model.py`: a script which reads the model file and outputs your prediction

`code/`
- `***.R` or `***.py`: all other scripts

`run_model.R` or `run_model.py` is the script that reads the model file and that outputs your prediction. You need to finish this script by filling in the template script provided on Canvas. When you finish the script, you need to run the fixed command line below to make sure the script works (file names in the command line can be different). You also need to check whether `output.txt` has the same content as `prediction.txt` you will submit.

```
Rscript run_model.R -i unlabelled_samples.txt -m model.rds -o output.txt
```

OR

```
python3 run_model.py -i unlabelled_samples.txt -m model.pkl -o output.txt
```

The folder `code` contains all the other scripts you have written for this project. Please keep them neat. They will be graded only as pass/fail, but should be able to reproduce your estimate.