Classification Assessment of Tumor Subtypes

3 Apr 2018

In this group assignment, you need to train a classifier that is able to predict subtypes of breast cancer tumors based on array CGH (aCGH) data.

Background

Breast cancer is a heterogeneous disease and classification of breast cancer tumors in their molecular subtypes has important implications for the treatment. 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+
- Triple negative (TN): ER-, PR- and HER2-
- Hormone receptor positive (HR+): ER+ and/or PR+, and HER2-

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

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 probability of that region being a gain/amplification, loss or normal, and we give the actual call (-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).

Building your classifier

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

1. Data purification, transformation, if necessary.
2. Feature selection, if necessary.
3. Choose machine learning methods (classifiers).
4. Train and validate the classifiers.

Programming and software

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

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 R script needs to be submitted.

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 to refer to the official documentation of these packages.

scikit-learn: Quick start
NumPy: Quick start
pandas: 10 minutes to pandas

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

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 at a later stage (see Schedule below). In addition, you will need to estimate how many labels 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) and 2) an estimate for the number of correctly labeled samples (out of 57).

Schedule and key dates

Please see Canvas for an up-to-date schedule and hand-in dates for this assignment

Deliverables

- Report written in the style of a short paper [70%]
- 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%]

Submit your group assignment

This submission consists two parts:

Part A.1: draft of a manuscript describing your research.
Part A.2: the final manuscript.
Part B: predictions + accuracy estimate + scripts + model files

Part A

A.1: Draft of report can be submitted through Canvas
A.2: The final report is also submitted via Canvas.

Part B

You can submit this also via Canvas: all your files should be made into a .zip or .rar file with a file name 'groupXX'. For example, if you are in Group 5, your submission file is 'group05.tar.gz'.

In this tarball file, you need to stick to the follow file structure in the folder 'groupXX'. We reuse the example here, so the folder 'group05' should have the following structure:

- results
  - estimate.txt
  - prediction.txt
- model
  - model.rds or model.py
  - run_model.R or run_model.py
- code
  - ***.R or ***.py

'results', 'model' and 'code' are all subfolders.
estimate.txt is a text file only containing a number (between 0 and 57).
prediction.txt contains your prediction, the format of which you need to refer to is on Canvas.
model.rds is the file that saves your R model, model.pkl is the one for Python.

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.

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 with '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.

Paper

For your (short) research paper, you need to choose a research question. This question should be discussed in one of the CATS practicals. 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:

Contents of paper (5-6 pages A4)

- Abstract (max 200 words), sections:
  - Motivation
  - Results & 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 section
  - 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 + 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 what research question you have answered in this paper
  - Explain what impact your research has on future research
- Tables & Figures
  - around 2-4 (in total) Figures and Tables
  - Make sure to explain all axes, labels, lines 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

Grading of draft paper (5-6 pages A4)

Your draft will be checked for the elements listed below. It will be scored as if it was a paper submitted to a journal, and judged if it can be accepted for publication.

Abstract:
Context:
Literature:
Research question:
Methodology:
Contents/Results:
Structure:
Lay-out:
Tables / Figures:
Style:
Length:

scoring: good / satisfactory / sufficient / insufficient

Grading of final paper (5-6 pages A4)

The final grade for the final paper will be based on:
- Research Question (originality, context, clarity, answer to question)
- Structure (abstract, length, figures/tables, structure of text, style)
- Contents (Quality of research performed & analysis)

scoring: 1-10

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. At the same meeting we will also announce the winner of the contest.