The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

COMP3009/COMP4139 Machine Learning 2024-25

Assignment 2

Machine Learning for Breast Cancer Treatment Response Prediction

Dr Xin Chen

1. Introduction

This assignment assesses your practical skills in applying machine learning methods to a real-world problem. The implementation will be based on Python and third-party Machine Learning libraries. Same as assignment 1, you must work in the same group and submit your work by **13th December 2024 at 3 pm** UK time on Moodle by member 1 of each group. You can split and distribute the work to individual members, but each individual is expected to understand every aspect of the work.

2. Background

Breast cancer is the most common cancer in the UK for women. Chemotherapy is a commonly used treatment strategy to reduce the size of locally advanced tumours before surgery. However, chemotherapy is a toxic process to the human body and it is not always effective for everyone. Complete tumour resolution at surgery, known as pathological complete response (PCR), has a high likelihood of achieving a cure and longer relapse-free survival (RFS) time. RFS is the length of time after primary treatment for cancer ends that the patient survives without any signs or symptoms of that cancer. However, only 25% of patients receiving chemotherapy will achieve a PCR, with the remaining 75% having residual disease and a range of prognosis. Better patient stratification and treatment could be achieved if PCR and RFS could be predicted using information prior to chemotherapy treatment.

3. Aim

You are asked to use advanced machine learning methods to predict PCR (classification) and RFS (regression) using both clinically measured features and features derived from magnetic resonance images (MRI) prior to chemotherapy treatment.

4. Data

Based on the public dataset from The American College of Radiology Imaging Network (I-SPY 2 TRIAL), a simplified dataset is generated for this assignment.

Each patient in this dataset contains 11 clinical features (Age, ER, PgG, HER2, TrippleNegative Status, Chemotherapy Grade, Tumour Proliferation, Histology Type, Lymph node Status, Tumour Stage and Gene) and 107 MRI-based features. The image-based features were extracted from the tumour region of MRIs using a radiomics feature extraction package (known as Pyradiomics: https://pyradiomics.readthedocs.io/en/latest/). You do not need to understand the meaning of these clinical features and image-based features to complete this assignment but worth reading background information on the I-SPY 2 Trial website. **"999" in the spreadsheet means a missing data value**. A training dataset (**trainDataset.xls**) is provided and available on Moodle that contains 400 patients. A test dataset that contains *N* patients is reserved (**hidden from you**) for the final performance evaluation. You can assume that the test set and training set are sampled from the same data distribution, but the ratio of PCR positive and negative could be different.

5. Implementation Requirement

You are asked to build a machine-learning model for each of the PCR (classification) and RFS (regression) predictions. You need to consider and implement methods for data pre-processing (e.g. how to handle missing data, outlier, normalisation, etc, if needed), data imputation, feature selection, machine learning modelling, hyperparameter tuning (if applicable) and method evaluation. There is no restriction or requirement for the selection of methods. However, you will likely need to compare several methods to pick the best one with the best parameter setting. When you perform feature selection, ER, HER2 and Gene are very important features that must be retained and used in the modelling process.

Your code will be finally tested on a reserved test set after your code is submitted. An example test file is provided (**testDatasetExample.xls**) that only contains 3 examples. It is your responsibility to ensure your code can run on a test file in a similar format but contains more patients. You must name your final test code "FinalTestPCR.py" or "FinalTestPCR.ipynb" for PCR prediction, and "FinalTestRFS.py" or "FinalTestRFS.ipynb" for RFS prediction so that they can be tested on the test dataset. The code for method development needs to be in a separate file, not in the "FinalTestXXX" file.

The test set will be released on 12th December 2024 at 9 am and you need to run your code to produce the predictions for the test set and submit on Moodle by 13th December 2024 at 3 pm together with other deliverables (section 7). One spreadsheet for PCR and one for RFS must be generated to store the prediction outcome. The output files must be a spreadsheet (.csv) that contains the predicted outcome for each tested patient (i.e. the first column is the patient ID, and the second column is either the predicted PCR or RFS outcome). Name the files: PCRPrediction.csv and RFSPrediction.csv. Balanced classification accuracy will be used to evaluate PCR prediction. Mean Absolute Error will be used to evaluate RFS estimation.

All implementations need to use Python programming language. Any machine learning libraries are allowed (e.g. Scikit-learn, Scipy, Pandas, Tensorflow, Pytorch, etc.). Grid search for automatic hyperparameter tuning is allowed. **However, any autoML based package or Large Language Models**

(e.g. ChaptGPT or other methods that accept the raw data and automatically select the best ML method and optimise the parameter for you) are NOT allowed.

6. Assessment

Assignment 2 weighs 80% of the coursework mark (i.e. 24% of the whole course mark). The marking will be performed based on the objective performance on the test set, the quality of code and the quality of technical writing. The marking criteria are provided in section 8. A single mark and feedback will be given to each group. The final mark for individual students will be calculated based on the contribution table described in section 7.

7. Deliverables

COMP3009 Students:
For the completion of Assignment 2, the following have to be submitted on Moodle. One report (.pdf) and one zipped code file need to be submitted per group.

1. The Python code for implementing the two tasks (PCR and RFS prediction). Besides the code for method development, two files "FinalTestPCR" and "FinalTestRFS" must be included for testing the test set. The two .cvs files for PCR and RFS predictions of the test set should also be included in the code folder (note: the test set will be released on 12th December 9 am on Moodle).

2. A report in the format of an IEEE conference paper. Technical paper writing will be introduced in one of the lectures. A template of the required format will be provided in Word and Latex. Based on the given format, a maximum of 4 pages is allowed, excluding references (references can be on the 5th page).

3. At the end of the paper (excluded from the 4 pages), the following contribution table needs to be completed and agreed upon by all members, which will be used to calculate individual student's final marks.

Task and Weighting	**Data pre-processing (10%)**	**Feature Selection (25%)**	**ML method development (25%)**	**Method Evaluation (10%)**	**Report Writing (30%)**
Name of member 1	30%	15%	20%	20%	20%
Name of member 2	0%	25%	30%	0%	20%
Name of member 3	30%	20%	20%	10%	20%
Name of member 4	0%	10%	30%	30%	20%
Name of member 5	40%	30%	0%	40%	20%

The percentage of contribution in the above table is an example, which will be different for each group depending on the true contribution of each member. However, the task names and their weighting highlighted in red in the table should NOT be changed, and the sum of the contributions from all members for each task (i.e. each column) should be 100%. Note that each student can contribute to multiple tasks and each task can involve multiple students.

COMP4139 Students:

If you are enrolled under COMP4139, besides the report and code required for COMP3009 students, you also need to submit **a recorded video presentation** to present your work **as a group**. The content of the presentation should cover background, a literature review on existing solutions, proposed method, evaluation results and conclusions & discussion. **The presentation should be less than 10 minutes and involve all group members (preparing the slides, presenting, or both).** Save the video in .mp4 format and submit it on Moodle (file size should be less than 250MB).

8. Marking Criteria

COMP3009 Students: **Elements**	% mark
Performance on test set (objective)	25%
Code quality (e.g. comments, easy to read, robustness, etc)	10%
Description of Method	25%
Explanation and presentation of the results obtained	15%
Discussion of the strengths and weaknesses of the chosen method	15%
Scientific writing and clarity	10%

COMP4139 Students: **Elements**	% mark
Performance on test set (objective)	25%
Code quality (e.g. comments, easy to read, robustness, etc)	10%
Description of Method	25%
Explanation and presentation of the results obtained	10%
Discussion of the strengths and weaknesses of the chosen method	10%
Scientific writing and clarity	10%
Presentation	10%

Plagiarism check will apply, meaning that high similarities across different groups are not expected. Late submissions in each assignment will result in a 5% penalty per day (days rounded up to the next integer).

9. Common Q&As

- **What is the performance of each task we are expecting to achieve?**

  It is a real-world dataset for a challenging clinical task, hence I don't have an estimation of performance. However, a >90% classification accuracy is too good to be true for this task. For the RFS estimation is even more challenging. The performances are expected to vary across groups. You need to consider practical issues, including missing data in both training and testing sets, data imbalance issues, etc. You have the freedom to use any machine learning methods that are not restricted to the methods introduced in the lectures.

- **Why don't we use an anonymised peer-assessment form to score the contribution of each member?**

  Anonymised peer-assessment form was used in previous years. Occasionally, members can not settle on an agreed distribution and it may involve several rounds of interviews to decide the final percentage of contribution. Hence, it is changed to a more transparent and quantitative contribution table.

  You should split the tasks and agree on the percentage of contributions before starting the assignment, then add/reduce the percentage depending on the final delivery and quality of completion by each member. Therefore, no surprises when you see your individual mark. Remember that each group is a team rather than individual competitors. An ideal case for a group of 5 students is that each member contributes to ~20%, but I don't expect it to happen for all groups. Please split the tasks depending on your group experience learned from Assignment 1. The highest mark a member can get is the group mark, which is based on the quality of the work. Hence marking down the contributions of other members won't get the top performer a higher mark. So help each other rather than kill each other 😊.