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.

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

COMP3009/COMP4139 机器学习 2024-25

作业 2

运用机器学习预测乳腺癌治疗反应

陈新博士

1. 简介

这次作业评估你在将机器学习方法应用到实际问题中的实践技能。实现将基于 Python 和第三方机器学习库。和作业 1 一样，你必须与同一小组合作，并由每小组成员 1 于 2024 年 12 月 13 日下午 3 点（英国时间）在 Moodle 提交你的作业。你可以拆分和分配给各个成员，但每个成员应了解工作的各个方面。

2. 背景

乳腺癌是英国女性中最常见的癌症。化疗是一种常用的治疗策略，旨在手术前缩小局部晚期肿瘤。然而，化疗对人体具有毒性，并非对每个人都有效。手术时肿瘤完全消退，即病理完全缓解 (PCR)，有很高的可能性实现治愈和更长的无复发生存期 (RFS)。RFS 是指癌症的初次治疗结束后，患者在无任何该癌症的症状或体征的情况下存活的时间。然而，只有 25% 接受化疗的患者会达到 PCR，其余 75% 患者将存在残留疾病和不同的预后。如果可以通过化疗前的信息预测 PCR 和 RFS，则可以实现更好的患者分层和治疗。

3. 目标

在化疗前，您被要求使用高级机器学习方法来预测PCR（分类）和RFS（回归），使用临床测量特征和从磁共振成像（MRI）中提取的特征。

4. 数据

根据美国放射学会影像网络（I-SPY 2 试验）的公共数据集，为本次作业生成了一个简化数据集。

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 13th December 2024 at 4 pm and you need to run your code to produce the predictions for the test set and submit on Moodle by 14th December 2024 at 3 pm. One spreadsheet for PCR and one for RFS must be generated to store the prediction outcome. The output files must be a spreadsheet (.cvs) 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). 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,

这个数据集中的每个患者包含11项临床特征（年龄、ER、PgG、HER2、三阴性状态、化疗等级、肿瘤增殖、组织学类型、淋巴结状态、肿瘤阶段和基因）和107项基于MRI的特征。图像特征是通过使用一种名为PyRadiomics的放射组学特征提取包（https://pyradiomics.readthedocs.io/en/latest/）从MRI的肿瘤区域中提取的。你不需要理解这些临床特征和图像特征的含义来完成这个作业，但值得阅读I-SPY 2试验网站上的背景信息。电子表格中的'999'表示缺失数据。一个包含400名患者的训练数据集（trainDataset.xls）已提供并可在Moodle上访问。一份测试数据集包含N名患者，被保留用于最终性能评估（对你隐藏）。你可以假设测试集和训练集来自相同的数据分布，但PCR阳性和阴性的比例可能不同。

5. 实施要求

你需要为PCR（分类）和RFS（回归）预测分别构建一个机器学习模型。你需要考虑和实现数据预处理的方法（例如，如何处理缺失数据、离群值、归一化等，如有需要），数据插补，特征选择，机器学习建模，超参数调优（如适用）和方法评估。在方法选择方面没有限制或要求。然而，你可能需要比较几种方法以选择最佳的参数设置及最佳的方法。当你进行特征选择时，ER、HER2和基因是非常重要的特征，必须保留并在建模过程中使用。

您的代码在提交后将最终在保留的测试集上进行测试。提供了一个示例测试文件（testDatasetExample.xls），其中仅包含3个示例。您有责任确保您的代码能够在格式类似但包含更多患者的测试文件上运行。您的最终测试代码必须命名为“FinalTestPCR.py”或“FinalTestPCR.ipynb”（用于PCR预测），以及“FinalTestRFS.py”或“FinalTestRFS.ipynb”（用于RFS预测），以便可以在测试数据集上进行测试。方法开发的代码需要放在一个单独的文件中，而不是“FinalTestXXX”文件中。

测试集将于2024年12月13日下午4点发布，您需要运行代码生成测试集的预测，并在2024年12月14日下午3点之前提交到Moodle。必须分别生成一个PCR和一个RFS的电子表格来存储预测结果。输出文件必须是电子表格（.csv），其中包含每个测试患者的预测结果（即第一列是患者ID，第二列是预测的PCR或RFS结果）。平衡分类准确度将用于评估PCR预测，平均绝对误差将用于评估RFS估计。

所有实现都需要使用Python编程语言。允许使用任何机器学习库（例如，Scikit-learn、Scipy、Pandas、Tensorflow等）

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.

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%

允许使用Pytorch等工具进行网格搜索以自动调优超参数。然而，不允许使用任何基于autoML的包或大型语言模型（如ChaptGPT或其他方法，这些方法接受原始数据并自动选择最佳ML方法并优化参数）。

6. 评估

第二次作业占课程作业成绩的80%（即整个课程成绩的24%）。评分将基于测试集的客观表现、代码质量和技术写作质量。评分标准在第8节中提供。每组将获得一个成绩和反馈。每个学生的最终成绩将根据第7节中描述的贡献表进行计算。

7. 提交物

COMP3009 学生：
为了完成第二次作业，以下内容必须在Moodle上提交。每组需提交一个报告（.pdf）和一个压缩代码文件。

1. 实现两个任务（PCR和RFS预测）的Python代码。
   除了用于方法开发的代码外，还必须包含用于测试测试集的两个文件 'FinalTestPCR' 和 'FinalTestRFS'。

2. 以IEEE会议论文格式撰写的报告。技术论文撰写将在其中一节课中介绍。将提供所需格式的模板，包括Word和Latex格式。基于给定格式，最多允许4页内容，参考文献除外（参考文献可以在第5页）。

3. 在论文末尾（不计入4页内），需要填写并由所有成员同意以下贡献表，该表将用于计算每个学生的最终成绩。

任务及权重	数据预处理 (10%)	特征选择 (25%)	机器学习方法开发 (25%)	方法评估 (10%)	报告撰写 (30%)
姓名 的成员 1	30%	15%	20%	20%	20%
姓名 的成员 2	0%	25%	30%	0%	20%
姓名 的成员 3	30%	20%	20%	10%	20%
姓名 的会员4	0%	10%	30%	30%	20%
姓名 的会员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).

上述表格中的贡献比例只是一个例子，根据每个成员的实际贡献，每组可能会有所不同。但是，表格中用红色标出的任务名称及其权重不应更改，每项任务（即每列）所有成员的贡献总和应为100%。注意，每个学生可以参与多个任务，每个任务也可以涉及多个学生。

COMP4139学生：

如果你注册了COMP4139课程，除了COMP3009学生需提交的报告和代码外，你还需要提交一段录制的录像演示，展示你们作为一个小组的工作。演示内容应包括背景介绍、现有解决方案的文献综述、提出的方法、评估结果以及结论与讨论。演示应少于10分钟，并需涉及所有小组成员（准备幻灯片、演示，或两者兼顾）。将视频保存为.mp4格式，并在Moodle上提交（文件大小应小于250MB）。

8. 评分标准

COMP3009学生： 要素	% 评分
测试集上的表现（客观）	25%
代码质量（例如注释、易读性、健壮性等）	10%
方法描述	25%
获得结果的解释和展示	15%
所选方法的优缺点讨论	15%
科学写作与清晰性	10%

COMP4139 学生： 元素	% 分数
测试集上的表现（客观）	25%
代码质量（例如注释，易读性，健壮性等）	10%
方法描述	25%
获得结果的解释和展示	10%
选定方法的优缺点讨论	10%
科学写作和清晰度	10%
演讲	10%

将进行抄袭检查，这意味着不同组之间不应有高度相似性。每次作业的迟交将导致每天（天数向上取整至下一个整数）扣除5%的分数。

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 😊.

9. 常见问答

- 我们期望达到每个任务的表现是什么？这是一个具有挑战性的临床任务的现实世界数据集，因此我无法估计其表现。然而，对于这个任务来说，超过90%的分类准确率是不现实的。对于RFS的估计，更加具有挑战性。表现预计会在不同的组之间有所差异。你需要考虑实际问题，包括训练集中和测试集中缺失的数据，数据不平衡问题等。你可以自由使用任何机器学习方法，而不仅限于讲课中介绍的方法。

- 为什么我们不使用匿名的同侪评估表来评分每个成员的贡献？

  前几年使用的是匿名同侪评估表。有时，成员无法就分配比例达成一致，可能需要几轮面谈来决定最终的贡献比例。因此，改为使用一个更透明和量化的贡献表。

  你们应该在开始任务之前分配任务并商定贡献比例，然后根据每个成员的最终交付和完成质量来增加或减少贡献比例。因此，当你看到你的个人评分时不会感到惊讶。记住，每个小组是一个团队，而不是个人竞争者。对于一个5个学生的小组来说，理想的情况是每个成员贡献约20%，但我不指望所有小组都能实现这一点。请根据从第一个作业中学到的小组经验来分配任务。一个成员可以获得的最高分是小组分数，这取决于工作的质量。因此，降低其他成员的贡献不会获得最高分数。

  给表演者一个更高的评分。所以要互相帮助，而不是互相残杀。 .