ETC 3555/5555 2024 - Assignment 3 - Group Project

In this group project you will practically apply the concepts you've learned in lectures and labs to a real-world dataset.

Objective

Your mission is to construct a series of models to predict a specific outcome of interest from a dataset you have been given. The emphasis isn't on achieving the best model but on the methodology you employ and your documentation of that process.

Groups

You have been randomly allocated to groups by Moodle according to whether you are taking ETC3555 or ETC5555 and which tutorial group you are in. Most groups have 3 participants. Marking adjustments will be made to account for smaller and larger groups.

Requirements

The project requires each group to complete the following tasks. The marks associated with each task are indicated below.

- Data Preprocessing: (10 points)
  - Download and manipulate your data into a suitable form for your machine-learning models
  - Do any required preprocessing, handling missing data, dealing with categorical variables or standardisation.
- Problem Specification (5 points)
  - Did you choose classification or regression

    - What error measure and/or evaluation metrics will you use

- Model Building (40 points):
    - Fit an appropriate standard linear model to your data - this will be your benchmark.
    - Extend your linear model to consider non-linear feature transforms. Did you improve upon your benchmark's out-of-sample performance? How did you modify the model fitting as a result of having a more complex model?
    - Consider a simple neural network model (one *hidden* layer). Can this improve upon your benchmark? What methods did you apply to make this happen?
    - Consider a deeper neural network with more than one *hidden* layer. Were you able to improve your benchmark even further and if so how?

  Each of these models may require and iterative model fitting procedure e.g. if you try a model and realise it is overfitting then you should try it again with regularsiation.

- Documentation (20 points):
    - Submit a concise report (up to 5 pages)
    - For your report, clarity and comprehensiveness are paramount. Ensure any figures are appropriately sized with clearly labeled, legible axes.
    - Marks will be awarded here based on how well you summarize what you have done and relate it to what has been seen in the lectures.

- Presentation (20 points)
    - Submit a presentation (up to 10 slides)
    - and deliver a presentation during Lab 12 (10-minute duration)
    - Every team member is expected to present. If unforeseen circumstances prevent you from presenting on the scheduled day, alternate arrangements will be made.

Project Report

Your data analysis report must be no longer than 5 pages and should adhere to the following sections:

1. Problem and Data
   - Describe the dataset you've been given and the goal of the analysis?
   - What minimal preprocessing was done to make the data suitable for machine learning?
   - What error measure will you use for learning and evaluating the model?
2. Models
   - Clearly define the four models/hypotheses you used and comment on the number of parameters they have and their flexibility
   - Outline the learning algorithm used in each case
   - Did your model require regularisation? If so what did you use, how did you estimate any hyperparameters
   - Mention how fast/slow your learning algorithm was for your specific model
3. Results
   - How did each of your models perform on the held-out test data?
   - Present your model fitting, diagnostics, and other relevant details.
4. Summarize your finding

Your presentation should follow a similar structure.

Deadlines

- Sunday 13th October at 23:55pm
- Your group needs to upload to moodle
  - Project report
  - Presentation slides

  - A zip file containing the source code and R files you utilised.

- During the lab sessions in Week 12 (Tuesday 15th October) your group will present your project to me, the tutors and your fellow students.

Remember: Procrastination can hinder quality. Start early to ensure thoroughness and clarity in your project!

Data Sets

The dataset comes from MIMIC project https://mimic.physionet.org/.

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Note: the dataset that you will receive has already been partially preprocessed for you.

The files `mimic_train_X.csv` and `mimic_test_X.csv` contain the training and the testing features for your model. Each row of `mimic_train_X.csv` and `mimic_test_X.csv` correponds to one ICU stay (`hadm_id` + `icustay_id`) of one patient (`subject_id`). The remaining columns correspond to vitals of each patient (when entering the ICU), plus some general characteristics (age, gender, etc.), and their explanation can be found at `mimic_patient_metadata.csv`.

Note that the main cause/disease of patient condition is embedded as a code at `ICD9_diagnosis` column. The meaning of this code can be found at `MIMIC_metadata_diagnose.csv`. But this is only the main one; a patient can have co-occurrent diseases (comorbidities). These secondary codes can be found at `extra_data/MIMIC_diagnoses.csv`.

The files `mimic_train_y.csv` and `mimic_test_y.csv` contain the training and testing response variables. Each row of these corresponds to the equivalent row in `mimic_train_X.csv` and `mimic_test_X.csv`. The response variable relevant to your group depends on if you wish to undertake a classification or regression task. Please choose one and only one of these.

- Classification: Column `HOSPITAL_EXPIRE_FLAG` is the indicator of death (= 1) as a result of the current hospital stay.

- Regression: Column `LOS` is the length of stay of the current hospital stay, equal to discharge time minus admit time.

You can use the training data in whichever way you like to *train* your models, but you should only use the testing data to *evaluate* each of your fitted model.

ETC3555 vs ETC5555

There are two differences in what is required of ETC5555 students compared with ETC3555.

Data: ETC5555 must incorporate the `extra_data/MIMIC_diagnoses.csv` data set containing the co-occurrent diseases (comorbidities) into their analysis in some way. This is not required for ETC3555.

Report: For marking ETC5555 greater onus is up on the students to demonstrate that they have understood how the hypothesis class, learning algorithms and methods for regulairsation work.

AI acknowledgement

Please provide a summary of how generative AI assisted you with the completion of this assignment. This should be provided as a final appendix section of your assignment file (`.Rmd`) and should include

- What generative AI tools did you use to complete this assignment?
- Which parts of the assessment did generative AI assist you with?
- How did you have to modify the output given to you by generative AI to answer the exercises?
- What did you learn from the generative AI you used that could be used in future assignments/projects?