11/08/21

Project Submission and Validation Tests Guideline

Remaining Events:

- 21/8: **Final Submission**:
  - **Final Submission (30%)**
  - **Performance on an external validation set (30%)**
- 23/8: **Model Evaluation meeting**
- 25/8: **Final oral Presentation (10%) -** Final joint meeting of all groups, where each group will have 7 minutes to present its project. You will be judged based on your presentation and understanding in "check point" building, room 420 @ 17:00.

*See guidelines below, grading mechanism was described in the project description document.

Submitted files

All files should be emailed to dancoster@gmail.com until **21**/08/2021 at 23:59.

The submitted files should include:

1. **Final submission document** summarizing your analysis and results. The document is limited to 10 pages. It should be self-contained, explaining your approach and methods without relying on previous presentations or documents, and contain the most prominent results and conclusions. You can add an appendix of additional results or method details. Please explain any method that you used/developed scientifically (e.g. *"we used a logistic regression a model with norm 2 for penalization"* rather than *"we used the classifier sklearn.linear_model.LogisticRegression with penalty=l2"*). Any method/package/function you used should be explained in a manner that will enable the reader to understand the concept of this method and repeat your work.
2. **The code of your analysis**. The code should be clear and we will evaluate it for readability, documentation, and code quality. Your code will be evaluated on an external validation dataset (see exact requirements below).
3. **The presenation you plan to present on the 25/8/21**.

Final Submission Requirements:

A. A **word/PDF File with detailed project** description, its main analyses, and results:
   (1) **Project Introduction**
   (2) **Cohort Description** – An explanation of the datasets - their characteristics, size, missing rates, etc. Add a table of characteristics that compares the two groups in each data set including an appropriate statistical hypothesis test.

(3) **Methods:**

a. **Data Preprocessing** – Explain the feature engineering process – normalization, transformations, shapelets, etc.

b. **Models** - Succinctly describe your models. For example, suppose you select only continuous features that were significantly different between the 'Positive Culture' vs. 'Negative Culture' groups on the training set based on student's t-test. Then, mention how you compared the groups (=t-test), but do not explain or repeat what is a t-test and how it is computed. If you are unsure – provide a reference to the method. The same goes for algorithms – when you use existing models, such as Random Forest, XGBoost, Logistic Regression, mention, explain briefly, and reference them but you do not need to explain their theory. If you develop something new – describe it in detail, in an appendix if needed.

c. **Evaluation Process** - how did you evaluate your model. K-fold validation? Bootstrapping?

(4) **Results:**

a. **Describe your results with relevant figures**. The figures should be informative with a clear legend to describe each one.

b. The results should be described based on the datasets you got at the beginning of the workshop.

(5) **Discussion:**

a. Summarize your project results, your model's limitations, and your main conclusions.

(6) **Refrences**

a. In case of Algorithm, please refer to the article that presented it. In case of a package, please reference for an article/link as well.

* It is highly recommended to use many figures to validate your conclusions.

* The size of the doc should include ~8-10 pages of text (without references but including figures and tables). If needed, use supplement for additional details and results.

* This document should be written so that someone could replicate your work based on it.

<u>Code Submission Requirements for External Validation Set</u>

Code should be written in python. We will evaluate the final submitted models by running them on datasets disjoint from your training data, but with similar characteristics. The validation file formats will be identical to the formats of *'model_a_mimic_cohort_v2'* and *'model_b_mimic_cohort_v2'* files.

Each module will get as an input the model type (e.g. model_type= {“a”,”b”}, as a string) plus additional parameters as described below.

1. **Requirements file** with a list of packages and their versions, use a comment for python version.
2. **module_1_cohort_creation(*file_path*, *db_conn, model_type*):**

   Output: ***external_validation_set.csv***

   db_conn is psycopg2.connect object (see this link), all the SQL queries should be incorporated into python code.

```
conn = psycopg2.connect(
    host="localhost",
    database="suppliers",
    user="postgres",
    password="Abcd1234")
```

   The ***file_path*** will contains a CSV file with the same formats as the *‘model_a_mimic_cohort_v2’* and *‘model_b_mimic_cohort_v2’* files (except for the target columon). The DB that you will query via ***db_conn*** will have exactly the same schemas and tables as in MIMIC (set tables to the name ‘mimiciii’, editing permissions will be provided for temporary tables). That means, that you will be able to create any feature that you created on the MIMIC database in the same manner. We put much effort into the mimicry (☺) of these tables. The output is a CSV file with all the raw features (not processed!) you’ll use in the model. The output could have multiple CSV files as needed (with any names as you wish). If you conducted any feature engineering in this module (such as non-human values removal, generation of time-series features etc.) please mention it explicitly in the documentation, notebook and submission.
3. **module_2_preprocessing(*external_validation_set.csv, model_type*):**

   Output: ***cohort_exclusion.txt***, ***processed_external_validation_set.csv***

   In this module, all the feature engineering processes will be performed (normalization, transformations, shapelets, imputation, non-human values removal, etc.).

   In case you remove part of the patients in the cohort (for example, patients with more than 80% of missing values), you should create a ***cohort_exclusion.txt*** file with an explanation of your exclusion process. You should also mention the number of patients that were excluded due to this criterion on the files *model_b_mimic_cohort_v2*, *model_a_mimic_cohort_v2*. The ***cohort_exclusion.txt*** file will contain:

   “Exclusion Criteria are: ….”

   “On Model *model_type*: Y patients were excluded (X% of the cohort)”

   “Z patients were excluded in the external validation set (M%)”

Since you don't have the validation set and we can not limit the magnitude M. Hene, we require that based on the MIMIC training dataset, In your exclusion criteria, X should be < 4.

The input could have multiple CSV files as needed (aligned to the number of CSV files that were the output of module_1).

The output will be the processed cohort, which should be the input for your model.

4. **module_3_model(*processed_external_validation_set.csv, model_type*):**
   This module contained your trained model. The input will be *processed_external_validation_set.csv* and the output will be a continuous predicted risk score (see attached file *model_a_mimic_cohort_risk_score_group_N.csv*, where N is your group number). Please notice that you do not train any model at this phase, but will only use the model that you trained based on *model_a_mimic_cohort_v2, model_b_mimic_cohort_v2,* and *model_a_eicu_cohort_v2*. We will have a file with the target of each patient (see example *model_a_mimic_cohort_target.csv*, and we will use the code *"validation_set_evaluation.ipynb"* to evaluate your model and calculate your model AUPR and AUROC).
5. **module_4_model_a_creation(*model_type,model_a_mimic_cohort_v2.csv, model_a_eicu_cohort_v2.csv*):**
   In this module, you will need to create model A, based on the cohorts that were avaible to you during the workshop. Please supply a path for the cohort.
6. **module_5_model_b_creation(*model_type,model_b_mimic_cohort_v2.csv*):**
   In this module, you will need to create model B, based on the cohort that was avaible to you during the workshop. Please supply a path for the cohort.
7. **Jupyter Notebook** - Your submission should include a jupyter notebook that will generate a CSV file with a risk score for each patient (as in *model_a_mimic_cohort_risk_score_group_N.csv)* with the abovementioned functions (see the format in *MLHC_notebook.ipynb*).

*You are more than welcome to generate any other modules to import from them specific functions.

* The target will be defined in model a as target=0 -> negative, target=1 -> positive
* The target will be defined in model B as target=0 ->appropriate, target=1 -> Inappropriate
* The submitted model should be trained both on eICU and on MIMIC cohorts.
* Modules 4 and 5 are optional, and groups that will submit it will get a bonus.

Our Validation Process:

We changed the plan so that you can run the code on the validation set on your laptop, rather than converting the code and running it on the nova environment. This will save you a lot of time, but you musht ensure that your code is robust enough to use on new data.

1. You will send us your code by August 21 @ 23:59 with all the above requirements in a zip file.
2. We will meet each group separately in person in TAU on August 15 in order to enable you to run your code locally on your laptop.
3. We will supply to you a portable USB that will have:
    a. The code (= that you already sent to us the day before).
    b. A DB with the relevant schemas that you'll be able to query via ***db_conn***.
    c. CSV files with the validation sets ({a,b}), that has same format as *model_a/b_mimic_cohort_v2*.
4. The meeting with each group will have the following steps:
    a. Sanity check – you'll have to run the *MLHC_notebook.ipynb* with the following parameters:
        i. file_path= *model_a_mimic_cohort_v2.csv* (without the target column).
        ii. db_conn=MIMIC details (on your local PC).
        iii. Model_type= 'a'

    You will need to generate a *model_a_mimic_cohort_risk_score_group_N.csv* file. Then you will need to execute *"validation_set_evaluation.ipynb"* with the target column (as in *model_a_mimic_cohort_target.csv*). These results won't be part of your grade, but only a sanity check that your code is ok.
    The running time of this process should be < 20 minutes.
    b. Model A evaluation – we'll evaluate your model on model A but now with *db_conn* parameters for the DB, and with validation set path (on the portable USB)
    c. Model B evaluation – The same for model B.
8. Please notice the following issues:
    - You will be able to run only the code that was sent to us on August 14 (you will need to execute your code on your PC but from the portable USB).
    - A group that won't pass the sanity check will not be graded for the external validation set part. Pay attention that we will need to use MIMIC DB as part of the sanity check so please do not erase MIMIC from your laptop.
    - If module_2_preprocessing X>4, your final AUPR / AUROC score will be reduced by 15%.

- We trust you on this point and wish to enable you to focus on the more computational (and interesting!) part of the workshop and not in the technical issues of submission. During the day of evaluation, do not share any information with the rest of the groups. Cheating will be addressed in all severity.
- It was not be required to run module 4 and 5 during our meeting, so you should run this module in advance, and be ready to use module 3. We will review these modules so the code should be readable and clear.