UNIL | Université de Lausanne

Project - Data Mining and Machine Learning

Professor: Michalis Vlachos

Note: Ask questions only on the #project slack channel.

Deliverable due: 10 Dec 2020, 23.59pm.

Project description

Real or Not? NLP with Disaster Tweets: In this project you are challenged to build a Machine Learning model that can predict which tweets are about a real disaster and which are not. The project topic is based around a Kaggle competition.

In this project, you will have the chance to compare your prediction results with your fellows. Create an account for yourself in Alcrowd and join the competition. Here's the link to the competition page:

https://www.aicrowd.com/challenges/final-project-of-the-data-mining-and-machine-learning-course

As soon as you make a submission you can see the prediction accuracy and your ranking on the leaderboard. Note that you can only make 5 submissions per day. To know more about the competition rules, check out the rules tab.

Data

You can find the training data and the unlabelled test data under the Resources tab.

Submissions

As you build your model and train it on the training data, you can generate predictions for the (unlabelled) test data. Make sure that your submission file has the same format as the example submission file on the Resources tab. Once you are sure about your model and satisfied with the prediction accuracy you got (on your own test data), you can try to generate predictions for the actual test data and submit in the competition.

Deliverables

I. Create a GitHub project for your project. The Github should have the following structure:

/code *(this should hold all your code)*
/data *(this should hold all your data)*

/documents (*this should have any related documents, reports etc)*
readme.md

In the readme.md you should a) mention the team name, b) describe the project c) your solution and some results (using figures). d) include also a link to the video that showcases your solution.

II. **/code**
Create a **Python notebook** that explains every step of your pipeline, from loading the data and preprocessing to building the model. Your notebook also stands as your report as well, so make sure that you add sufficient explanations to it. Add appropriate plots and/or tables to your notebook in case you think it can make your notebook more comprehensive. It is very important also to document, the **progress** of your submissions. Eg you should keep track of the reported accuracies of the different submissions and also what changes you introduced to have some improvement in your score.
Try to separate the preprocessing into a separate python file that you import from the main python file. In your notebook we expect to see at least:
   A. What is the baserate of the problem
   B. A table with all of the classification techniques that we saw in the class (logistic regression, kNN, Decision trees,…), the parameters used and the achieved accuracy.
   C. Your progression of accuracies in a graph, and which technique (with what parameters) achieved it.
   D. Go deeper. How can you improve the model? Do both of the following.
      1. Feature engineering.
      2. Hyper-parameter optimization.

Note: You should ONLY use techniques that we used in the class/lab. No other techniques are allowed. We should first master those techniques!

III. Create a **short video** of your solution (duration is up to you, but not more than 15mins) and report also in the video your best rank in the leaderboard. Post your video in slack channel #project, before the deadline.

IV. After the project description is available, each team is expected to do a **weekly stand-up** during the lab lecture. You report your progress. There needs to be something said every week, i.e., teams that have nothing to report for some week will experience some small penalty in the grade.

V. During the last class you will give a short **presentation**, summarizing what you did and your best result. Note, the important is to focus on the content of your project, and not to have a super-perfect presentation that has little content.

Logistics and deadline

1. Create an account in AIcrowd and join the competition. Here's the link to the competition page:
   https://www.aicrowd.com/challenges/final-project-of-the-data-mining-and-machine-learning-course

   After entering the competition page, you would see a "create team" button on the top right. Click on that button and select your team name. Your team name should be the same as the name which was assigned to you in Moodle. This name will be shown on the leaderboard. Once your team is created you can invite your team-mates to the team. Remember each team could only have 3 participants.

2. Make sure to mention your team name in your notebook.

3. There will be gifts for the champion! (team with the highest ranks in the leaderboard) :)

   The deadline for the submission of your material is **10th of Dec 2020** at 23h59.

Grading

1. Notebook quality (clean code, nice graphs, sufficient explanations, etc): 1/3 of the grade. Your code should be separated in logical blocks. You cannot have all your code in one big notebook!
2. Depth of solution. Experimentation with feature engineering and parameter optimization: 1/3 of the grade.
3. Presentation in the class: 1/3 of the grade (but focus on the project content first!)

Good luck with the project and the competition. We look forward to seeing your solution!