EECS4404/5327 Project Component

1 Overview

Instead of a final exam, this term you are asked to apply what you've learned in a small group-project on a supervised learning task. Please read these instructions fully before starting out. The overall procedure will be as follows (see below for details on each step):

Groups You will be assigned to a "super-group" (or simply group) of 9 or 10 students. There will be 7 such groups in total.

Data-set Each group should choose a dataset to work with, and agree on a train-test spit.

Train predictors The groups should then designate three sub-groups to separately train three different types of predictors (eg. a neural network, a k-nearest neighbor predictor, and a kernelized support vector machine) on the train data.

Test and combine the models Once each group converged on a predictor, these should be tested on the test data. Additionaly, you should test the performance of a (weighted) ensemble of the three predictors.

Report Each group needs to submit a report on the project (with sub-reports from the sub-groups) by December 20.

Individual task In the end, after your final report is submitted, you will be asked to individually fill in a small questionnaire on e-class.

The goal of the project is to *understand a toolbox* for one type of model, *work together in a team* and *train a model* that performs *reasonably* well on a dataset and most importantly, *adequately report* on your approach, the steps you took and your findings. You are not expected to achieve state-of-the art performance on the data you choose.

2 The Tasks

2.1 Planning

Choose a dataset, assign roles, and set time-lines

The first tasks for each group is to choose a dataset and assign tasks to all group members. For the dataset, I recommend choosing a classification dataset from the UCI repository, but you may choose a dataset from another public repository. You should ensure that there at least 500 datapoints, and that the features are numerical or a combination of numerical and and categorical. Examples (and you may just choose on of these) of suitable datasets from UCI are:

- Bank Marketing Data Set https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- Breast Cancer Wisconsin https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
- Adult Data set http://archive.ics.uci.edu/ml/datasets/Adult

Next, assign subgroups and roles. You may just choose to split into three sub-groups of three, then have each subgroup work on learning one type of predictor jointly and do the write-up jointly. Or you may choose to have the three predictors train by pairs of students, and have several students be assigned the task of building the ensemble in the end and prepare the report. You have some flexibility here, but you should work together as a team in the super-group to assign each member a task that they will best contribute on. Keep a record of the tasks assigned. Also sketch a time-line for when each task should be completed. Again, keep a record of this. Plan for some buffer time!

Also choose a contact-person. That is, assign one student to be the point of contact for me to the group. Once you decided on the dataset and point of contact, let me know by email.

2.2 Learning predictors

On the super-group level, agree on a train-test split of the data. Choose (and save) a random permutation of the data-points and set 20% of the data aside for testing in the end. Also decide on an evaluation metric. If the classes are imbalanced 0/1-loss may not be suitable, and you may choose to rather evaluate your model with an F1-score. Or, you may subsample a class-balanced sub-dataset and work only with these data-points. Record your choices.

Train predictors in sub-groups

In this phase, you are only working with the training data the super-group agreed on.

Assign three sub-groups to train three different types of predictors, such as a fee-forward neural network, a random forest, a k-nearest neighbor classifiers, a support vector machine (using a kernel). If you are working with python, I recommend you use scikit-learn library. For matlab, you can use the statistics and machine learning or the neural network toolboxes. To not spend too much time, it is important that you just make a choice, get yourself familiar with the type of predictor you want to learn and how to use the toolbox for that, and then train a predictor. Find out (and keep notes of) what type of parameters you can choose (ie which hyper-parameters you get to set). You may decide to only work with a subset of the features. Report on whether you chose to do so and how.

Use an additional split into train and validation data (or use cross-validation if you are working with a smaller dataset) to make a selection of potential hyper parameters. Again, keep a record of the methodology and the choices you are making. Finally, train a predictor on the training data.

Testing and building an ensemble

Once each group has trained a classifier, these should be evaluated on the test data. Additionally evaluate a majority vote (ensemble) over the three predictors you trained. You may try different weightings for the majority votes and see which single or weighting over the three predictors works best.

Note that for this part, each group needs to just provide the vector of predictions on the test data points. This suffices to also get the predictions of weighted majority votes over the test points.

2.3 Report

You will submit **one report per super-group**. The submission deadline is **December 20**. The goal of the report is to provide an accurate summary of how the project went.

Format

Aim for 5 pages maximum, 11pt or 12pt font, 1-1.3 inch margins. An additional page can be used for a list of references for any resources you consult.

Content

I suggest the following structure:

1. **Report on the planning and its execution (about 0.5-1 pages).** Use an introductory section to report on your plans, who was assigned which roles/tasks, what was the time-line you had planned for, how did it work out? How did you adapt in case there were delays or tasks took longer than anticipated? In case there were any conflicts, how did you resolve them?

2. **Three sub-reports for the three models trained (about 1-1.5 pages each).** Here describe which type of predictor you chose to train, which toolbox you chose to use, what was your overall methodology and what was the training performance (ie. report on whether you chose to use cross validation etc).

3. **Summarize your findings in a conclusions section (about 0.5-1 page).** Here report on the performance on the test data of each of the predictors. If there are any differences clear differences, can you explain them? Also evaluate the majority voting and discuss what you are finding.

Evaluation

I will evaluate the projects (the super-group report and individual questionnaires) according three criteria:

1. **Teamwork (40%)** How did you plan and work together as a group?

2. **Accurate reporting (40%)** Do the different part of the report paint a consistent picture of the work that was done?

3. **Overall evaluation and presentation (20%)** Is the report well structured and written? Are the main findings adequately summarized and explained?

3 Two pieces of important advice

Avoid plagiarism: Always clearly indicate when you are summarizing or citing content from a paper, website or documentation as opposed to stating your own ideas or results. **Never, under no circumstances, copy and paste any parts from a research paper or websites or anywhere else!** It is entirely legitimate for this project to consult any number of resources, but you have to properly acknowledge them. If in doubt, please ask!

Enjoy the process, don't get stressed out :) Always remember that this goal of the project really is to try things out, not to obtain the best possible performance in the end. If, in a reasonable amount of time (say each team member should expect to invest three full days), you don't get a satisfactory performance, just report this accurately and explain what you did and tried.

If you have any questions, would like to discuss intermediate results or considerations, send me an email!