Project Description for Introduction to Machine Learning 2020

The final project is a chance for you to use the knowledge you have obtained during the course and in true machine learning spirit compete with your best algorithms. The competition is hosted on the platform Kaggle. For those of you not familiar with this platform it is a place for machine learning enthusiasts to exchange datasets and ideas and compete against each other, or as Wikipedia would state it:

"Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges"

The data that we are going to study comes from a project studying migration patterns of cancer cells. Cancer cells migration is an important topic for cancer research. There are different types of cancer, and migrating cancer are among the worst ones. This process is known as metastasis. Understanding how cancer cells migrate could potentially help researchers develop new drugs and schemes to fight metastatic cancers. There are two migration modes for cancer cells: continuous and discontinuous. Discontinuous cells are more invasive as compared to their continuous counterparts. All cancer cells that were studied are clones of one another, and the current state of knowledge cannot explain the differences of migration mode. The challenge consists of classifying cancer cells into 2 classes: continuous and discontinuous. The original data contained videos of the cells, and features have been extracted and compiled in a single table for your convenience.

You will have to train a classifier of your choice on the training data. Separating the train set into a train and validation set will help you generalize. The training set contains 109 columns, some are useful, but many contains no information. You need to clean the data and remove useless information. Finally, when you have a working classifier, you can predict the test set and the resulting file to Kaggle and compare your score to the others. At the Kaggle page for this project you can check out the current leaderboard, download the data or read more about the project/challenge, there is also some tips and tricks on how to get started.

At least one in each group would need to register your team on Kaggle (if none of you want to do this, you can send your results to us and we will submit them). First register for Kaggle, then under "teams" (on the competition page)

you come up with a team name and if you want invite the other persons in your team.

After this you are ready to submit your results. You can only make 2 submissions per day and you will receive your score directly after submission. Be careful though, there is a secret test set that will be revealed at the end, so don't overfit the test set! Submission is done by uploading a generated .csv file with your predictions on the test set. In the data folder there is a sampleSubmission.csv file which is your template for submission. Basically, you generate your own .csv file with one column called "id" and one called "Predicted". The id should be the same as the one in the test file and the "predicted" is the prediction for that row in the test file (the class labels 0, 1 or 2).

When you feel that you are happy with your results you write a report on you work ~ 5 pages. Here you describe and interpret your results (include lots of graphs and plots). What pre-processing steps have you done and what features did you use? If you picked any algorithm that was not described in the course you should describe this briefly.

You are evaluated on your final report and the way you chose to solve the problem and not on how you score in the competition, this is there to make it more fun and motivating.

If you have any problems don't be afraid to send an email to any of the lab assistance or come by Håkan and Nicolas office 2144 or Evas 2114, from 8.00-11.00 and 12.00 - 14.00.

Start by heading over to the competition page to learn more about the data and the challenge

`https://www.kaggle.com/t/caab1d32e4cf4e138d67fe3d464ebb4b`