CSC311 Machine Learning Challenge: Group Formation and Getting Started

Form groups

Groups are to be formed by Nov 17th 10pm, which is the day after the drop deadline. To form groups, please submit a group name in a file called `groupname.txt` consisting of your group name, to be used for the leaderboard. Group names should follow the usual academic standards (e.g. no offensive language). Please make sure that you accept invitations to the groups prior to the deadline.

The groups will be persisted in the remainder of the project. If you do not declare a group, then you will not be eligible to participate in the ML Challenge.

Getting Started

The following files have been posted on Quercus:

- `clean_quercus.csv`: This is a CSV file containing the training data. The column "label" is what we wish to predict, given the remaining columns.
- `challenge_basic.py`: This is a python file containing code that performs some data transformation (e.g. create indicator variables) and build a rudimentary kNN model using sklearn. You may use any part of this code for your own exploration.
- `example_pred.py`: This Python file is example of how your `pred.py` script should take in the input, and produce an output. Your `pred.py` script can use different methods to process the input data, but the format of the input it takes and the output your script produces should be the same.
- `example_test_set.csv`: This is a CSV file that demonstrates how the test set file will be formatted.

We recommend that you take a look at all these files before getting started. Keep in mind, though, that these basic files are meant to be a starting point. Not all features are considered in the code.

Computer Resources

You may use your own laptop or the department computer labs. You should be able to remotely access the PCS in DH2020, DH2026, and DH2010. Choose any machine to use. The machines are named

- dh2020pcXX.utm.utoronto.ca
- dh2026pcXX.utm.utoronto.ca
- dh2010pcYY.utm.utoronto.ca

where XX is between 01 and 30 inclusive and YY is between 01 and 50 inclusive.

These machines should have `numpy`, `scipy`, and `sklearn` installed. If you need other software, please let Lisa know and she will relay the request to our system administrator. **Request for software should be received by Nov 17th.**

To access the machine via SSH, open up the terminal and enter the following command, replacing `<utorid>`, `<lab>`, and `<machine>` with your utorid, lab room, and the machine you wish to use:

```
ssh <utorid>@dh<lab>pc<machine>.utm.utoronto.ca
```

Once you are connected, you can use an editor of your choice to write out files (vim, nano etc). If you prefer writing with an IDE on your local machine, you can also transfer files from your local machine to the DH labs using the command `scp`.

For example, the following command will transfer the file project-a.hs from your local machine to your home directory in the dh2020 pc01.

```
scp project-a.hs <utorid>@dh2020pc01.utm.utoronto.ca:/student/<utorid>/
```