COMP 4980: Machine Learning

– Final Project –

Summary

The final project for this course will be a scientific report of an in-depth machine learning experiment.

Groups

If you wish, you may do this project in a group. Expectations will be higher for a group submission, so there is no advantage or disadvantage either way. Let me know within one week if you are planning to make a group submission.

Deliverables

1. A PDF report of your work.
    a. Make sure you include a working link to the dataset you use.
    b. If you used a notebook for your project, include a web link to the notebook. Alternatively, if you have a larger project stored on Github, please include a link to the repository.
2. A copy of the python file associate with your code: either the Python file output (.py) of the notebook, or the main python script of a larger program (you do not need to submit libraries).
3. If the data set you use is less than 50 MB in size, please also upload that to Moodle.

Data

You will need to select a data set to use for this project. You will receive feedback on the ***Data Discovery*** Exercise on whether or not the data set you selected is an appropriate choice. Part of your mark will be based on the complexity of the data set you have chosen and your project plan. Generally speaking, a larger data set (more columns and/or more rows) is better.

I ***highly*** suggest working with tabular data, i.e. a set of individuals records each containing entries for n variables. I will not be providing support for data sets of image or audio files, although you may work with those if you are willing to do the extra work on your own that will be required to use them.

Format

The PDF should be clearly formatted as a report. It should include the following sections:

i. Cover page
ii. Data description
iii. Data analysis
iv. Data exploration
v. Experimental method
vi. Results and analysis
vii. References

Cover Page

This should include the title of your project, your name, the course name and code (COMP 4980, Special Topics: Machine Learning), and the date.

Data Description

This section should include a description of the data, including its name, contents, where you acquired it, size, etc. Please refer to the ***Data Discovery*** exercise for the expectations of what is included this section. You *may* freely reuse your work from the ***Data Discovery*** exercise, although you may want to expand it or polish it for the final report.

Data Analysis

This section should include some general statistical analysis of the data. This can include descriptive statistics (e.g. mean (average) values, max/min values), correlations, and anything else that is useful (such as the type of distribution of a variable). Visualizations are a very good thing to include, such as histograms and scatterplots to show variables and their relationships. You do not need to include every possible visualization: instead, identify a few that help to understand the data and your project plan.

Data Exploration

Perform initial exploration of the data using:

i. Principal Component Analysis (PCA)
ii. Decision trees

For the PCA analysis, find out the relative importance of the principal components. How many are necessary to explain 96% of the variance? Does the PCA analysis tell you anything about the data?

For the Decision Tree exploration, please use simple (i.e. single) decision trees (not an ensemble method). You can use Classification and/or Regression trees, as appropriate. Please include at least one visualization of a small (depth 2-5) decision tree that provides insight into the data and the relationships between the variables. Include a text explanation of this insight and anything else you noticed in this phase.

Experimental Method

Develop a machine learning hypothesis and test it using Random Forests (and/or ExtraTrees), AdaBoost, GradientBoost, and Multi-layer Perceptron (MLP, a type of neural network). Choose at least one of the ensemble methods in addition to MLP and refine your experiment until you produce good results (using cross-validation). This could involve changing your hypothesis, feature selection, feature scaling, otherwise modifying the variables, computing new values from the data, tuning the hyperparameters of your ML methods (e.g. with GridSearch, etc.), and other changes. Please record the major changes and why they were tried.

It is a very good idea to keep a lab diary for this phase. This is a standard part of experimental science: you record the hypothesis of each experiment and the method chosen. In our case, the method is the machine learning method, the hyperparameters, and any other details necessary to retry the experiment. **BIG HINT:** it is straightforward to include text output in your program code that captures these details so you can copy and paste it into your report! It is also possible to output it to a file so that you have a permanent record.

Results & Analysis

Once you are satisfied with your results, present them clearly in this section. This will include a clear description of the final method or process, along with an in-depth description of the performance. This will include the primary scoring method along with any other relevant evaluation metrics (such as a classification matrix, precision, etc.). Make sure you are using cross validation for your primary score to demonstrate its reliability.

Include a write-up of what you think this result means in terms of the data. What patterns or relationships have you found in the data, if any? Also discuss what practical applications your machine learning model could be useful for, if any.

References

Include references for any website or text you refer to in your report. Vancouver style references are the best, but you can use another style (like APA) if you are consistent.

Academic Integrity

Don't cheat or copy text without providing a citation. Claims that require support should also be cited (e.g. this method was developed by X; this method is 250% more efficient, etc.). Don't copy large amounts of other people's code. If you submit the same work as another student, you will receive an F in the course—please work on it as a group project instead.

As we've seen in this course, many of the most common lines of machine learning are the same across programs, and it is fine to use those. If you want to include a function, file, or library from another author, that is fine as long as it is referenced.