ISTA 421 / INFO 521 – Final Project OPTION B

Due: Monday, December 13, 5pm

Total 20% of final grade

Forward

Final Project Option B is a project of your choosing. If you have chosen to do Option B, it is assumed you have already spoken with us (Cristian and Clay) about your plans. If you did not talk with us already, you can still submit the for this option, but you are NOT guaranteed that we will accept full 20% credit if the project is not deemed substantial enough; if you have not already spoken with us, I recommend that you do Option A.

The final project is due Monday December 13, 5pm.

Option B: Project of your choosing

In this option for the final project, you are proposing your own project that involves any machine learning method, including methods we have not covered in class. This can include contrasting 2 or more methods. You will preferably demonstrate the performance of the method(s) on real-world data, but could also use synthetic data if it does a good job demonstrating the strengths and weaknesses of the method(s). You must include in your project an evaluation of methods that involves appropriate training and testing.

You will include the following three types of material in your submission:

1. Any data you used to train/test your method. The data in your submission must be open source (no license restrictions). If you are using data that you did not generate or collect yourself, you must clearly specify where it came from and make appropriate attribution / citations. If the data is very large (more than 20MB) and is hosted on a server, you can submit instructions for how to retrieve it; in this case, you should still include in your submission a small portion of the data (< 20 MB) on which you can still run your code, so that your submission can be run stand-alone. If the data you are interested in working with is proprietary, you must talk to me ahead of time. (In this case, one option is to create synthetic data similar to your proprietary data and use that for your final submission.) One general source of data is the UC Irvine Machine Learning Repository (`https://archive.ics.uci.edu/ml/`), and you can also use Google's datasets search: `https://datasetsearch.research.google.com/`

2. Code you used in your project. All code must be open source. It *is* ok for you to use code written by others as long as it is open source. **All code must be well-documented**. If you are using code that an individual or research group developed, you MUST ensure it is well-documented; if it is not, you must document it yourself, clearly describing what the component parts do. You can use larger libraries, like SciKit-Learn – in this case, you do not have to submit the code base or document it (that is sufficiently done already). If you have any questions about this, be sure to talk to me before your submission. In all cases, you should have a "top-level" script that you run that will perform training, perform evaluation, and generate summary results. This must also be well-documented and instructions for its use will be included in your report; from these instructions, I must be able to reproduce all of the results you report.

3. A **concise and clear** written report (submitted as pdf). The report should follow the same format as a typical conference paper and include the following:

   (a) The pdf submitted in the Project B repository should be named: `final_project_OptionB_report.pdf`

(b) The written report will be single-spaced sentence spacing and around 8 pages (if including many figures, it can be longer, but I want your writing to be *clear* and *concise*). Include what is needed to clearly communicate that you understand the methods you used, explain what you did, and describe your results.

(c) Most reports will include the following sections (depending on your topic):

i. Introduction: Provide a concise description your project, including what the task is – what is it that your proposed method(s) is accomplishing. You do not need to include an exhaustive literature review, but should cite relevant sources and demonstrate you understand what the task is.

ii. Models: Describe the machine learning models/methods (including relevant equations and what they mean, description of the algorithm, etc.) that you are working with in this project. Describe the type of problem the method(s) is designed for. Include citations with summary descriptions of any relevant background material you are deriving this work from.

iii. Data: Describe the data you are using, including where it came from, what it's structure is, what it represents.

iv. Procedure: Describe the procedure you followed to train your models and generate any results. This should include instructions for how to run your top-level script/code to train, test and generate summary results.

v. Evaluation: Describe your evaluation method and rationale for its use. Describe all measures/statistics that you use to evaluate your method's performance.

vi. Results: include a clear description of the results of applying your method to the data; include any figures/tables of data summarizing the results. All plots/tables must have labeled axes and clear, concise and informative captions explaining what the plots/tables represent.