Undergraduate AI Capstone NYCU Spr2025 Project #1 Due 3/18/2025

In this project, the objective is for you to go through the process of constructing and analyzing a dataset. Most of the time (including in our own classes), algorithms are talked about more than the data. However, to build good models in machine learning, you need good data to learn from. As a result, this is one aspect that usually consumes a large portion of efforts and resources when building models.

Some requirements and suggestions on your dataset:

- You cannot use a dataset that is already available. You need to create one yourself.
- However, you can look at public datasets and their documentations to get an idea of how to construct datasets.
- Choose a topic such that (1) you can obtain reasonable labels and/or target outputs, and (2) it does not require a huge amount of data to produce reasonable results.
- It is allowed that two of you collaborate to compose a single dataset. Just indicate name of your collaborator in your report. Such a dataset will be expected to be of better quality.
- The grade will include part on the "creativity" or "novelty" of your datasets. Very trivial datasets, like those for digit recognition or dogs-vs-cats types, will get fewer points on this part. There are many different things you can try: text, sound, all kinds of sensors on your smartphones, open data sources (such as population or weather or stock markets), etc.
- You can try automatic or semi-automatic ways of data collection, such as web crawlers or simulations. However, such datasets usually require some level of cleanup, an aspect that you should address in your report.
- You can choose to construct your dataset for classification or regression.
- Think carefully about the characteristics, constraints, and composition of samples to be included in your dataset.

Algorithms:

- At least two methods for supervised learning and one for unsupervised learning. At most one method can be deep-learning based.
- Pretrained models can be used.
- You can use publicly available codes for your algorithms or models.
- You can use publicly available codes or sources to convert your raw data to features (such as HOG for images or MFCC for audio) or "embeddings" (such as word2vec).

Analysis:

- Evaluate the performance (supervised learning) using tools and metrics such as accuracies and confusion matrices for classification, and MSE or similar metrics for regression. For yes/no classification problems, measures such as precision/recall/F1/AUROC should be considered.
- Be sure to use cross-validation to obtain the evaluation metrics (supervised learning).
- If you do clustering (unsupervised learning), labels can be used as external evaluation metrics.

Experiments:

- You need to do experiments on different aspects related to the datasets. The following are just examples; try to come up with other "research questions". Some experiments can be done using partial datasets.
- How are the results affected by the amount of training data? Are different methods and/or hyper-parameters affected by the amount of data differently?
- How are the results affected by the composition/balance of data? Are the results improved with class/sample weighting or resampling (like SMOTE)?
- Try data augmentation if applicable, and compare the results with and without data augmentation.

- Compare the results with and without dimensionality reduction (such as PCA) of the features when using high-dimensional data.

Discussion:

- Based on your experiments, are the results and observed behaviors what you expect?
- Discuss factors that affect the performance, including dataset characteristics.
- Describe experiments that you would do if there were more time available.
- Indicate what you have learned from the experiments as well as your remaining questions.

Submission:

- A report in PDF format (maximum 10 pages single-spaced). It should contain the following sections:
  - Web link to your dataset (see below).
  - A brief statement of you research question in plain text.
  - Documentation of your dataset, include the data type, external source (if any), amount and composition, the conditions you set for data collection, the process of data collection (including any hardware or software you used), and examples.
  - Description of the supervised and unsupervised methods used. Provide references for public libraries, open-source codes, and/or pretrained models you used. Also include other methods used in your experiments, such as feature extractor, data resampling, dimensionality reduction, etc. Note: You should expect the reader to read this part of your report without the code listing. Do NOT write your report like a documentation of your program.
  - Description of your experiments, include evaluation results and examples.
  - A "discussion" section as described above.
  - List of references.
  - Include your program code as an appendix (not counting toward the 10-page limit), starting from a separate page. You can use C/C++ or Python to write your program. In general, the TAs will not actually compile or run your programs. The code listing is used to understand your thoughts during your implementation and to find problems if your results look strange. Therefore, the code listing should be well-organized and contain comments that help the readers understand your code; this will also affect your grade.
- Additional notes about the report:
  - Submit your report through E3. Late submissions are accepted for up to 5 days, with a deduction of 10% per day.
  - Experimental results should come with sufficient description/summarization and discussions. When you list your results, always ask yourself this question: "What message do I want my reader to learn from these numbers?" And then put your answer to text.
  - For presentation: Tables drawn in you report are so much better than screenshots. Charts or plots, when appropriate, are even better.
- Dataset submission:
  - Upload your dataset to GitHub. You need to include your dataset documentation (as described above) as well.
  - Include the link to this GitHub page in the beginning of your report.