CS 412 Introduction to Machine Learning, Spring 2018
University of Illinois at Chicago

Homework 5: Mini-project
Due: May 3, 2018, 11:59pm

Goal: Gain experience developing a machine learning project on a real-world dataset by utilizing the concepts and algorithms that you have learned in class. This is an individual project for graduate students, while undergraduate students are allowed (but not required) to work in pairs. You can choose one from three possible tasks:

Option 1: You are working for a non-profit that is recruiting student volunteers to help with Alzheimer's patients. You have been tasked with predicting how suitable a person is for this task by predicting how empathetic he or she is. Using the *Young People Survey* dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/), predict a person's "empathy" on a scale from 1 to 5. You can use any of the other attributes in the dataset to make this prediction.

Option 2: You have joined a startup that delivers healthy meals to people. You have been tasked with doing a marketing study and understanding how likely a person is to "pay more money for good, quality or healthy food" (on a scale from 1 to 5) using the *Young People Survey* dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/). You can use any of the other attributes in the dataset to make this prediction.

Option 3: You have been hired as the first data scientist at a news organization and tasked with creating value for the company from their terabytes of data. As a first task, you decide to use machine learning to do something interesting with the *Vox news corpus* (https://data.world/elenadata/vox-articles). You need to define a supervised or unsupervised learning problem and solve it. You must do something more complicated than binary classification.

For all tasks, you can use existing python packages, such as sklearn, libsvm, TensorFlow, keras, etc. but make sure you give credit in your write-up. In your evaluation, you will need to define simple classifiers as baselines and show that your proposed method is performing better than the baselines. Split the data into train/dev/test and tune hyperparameters on the dev data, and report final results on the test data. You are welcome to report on multiple methods that you have tried.

Be sure to answer the following questions: (a) what is your data and task? (b) what ML solution did you choose and, most importantly, *why* was this an appropriate choice? (c) how did you choose to evaluate success? (d) what software did you use and why did you choose it? (e) what are the results? (f) show some examples from the development data that your approach got correct and some it got wrong: if you were to try to fix the ones it got wrong, what would you do?

What to submit:

1) **Code**: Upload all your python files as a single zip file **hw5.zip** on Gradescope under *Homework 5*. Include a README that describes how to run your code. When running your code, it should print high-level information about what it is doing and also the results from the evaluation.

2) **Write-up**: Upload a *one-page* description as a PDF under *Homework 5 – Written Part* on Gradescope. If you choose to submit more than one page, keep in mind that we will not read anything beyond the first page and will grade the homework based on the first page only. You need to describe the problem

you are solving, the dataset preprocessing steps, your solution, your experimental setup (e.g., % train/dev/test), and evaluation. You will be graded based on creativity, clarity, completeness, and valid justification for all the steps in the project. We will not grade the project based on whether it achieved the best possible accuracy.

3) **Optional** (up to 20% extra credit): Students will be given up to 20% extra credit for creating a *private* github (or bitbucket) repository for their project and depositing their code, write-up, and an additional Jupyter notebook in the repository. We will only grade the notebook which should describe the steps in your project with a mix of code, narrative, and figures that provides more information than your one-page write-up. There is no length limit for this part. For some examples, see the notebooks in the github repository for the Hands-on ML book (https://github.com/ageron/handson-ml). Do not deposit the dataset in the repository, instead add a link to its online source in your repository README. If you don't have experience with github and bitbucket, take a look at these introductions: https://guides.github.com/activities/hello-world/ and https://confluence.atlassian.com/get-started-with-bitbucket. You need to share your repository with the Instructors (usernames to be shared later) and the repository link should be included in the write-up added to Gradescope. You are allowed to make your repository public only after the semester is over.

Your entire homework will be considered late if any of these parts are submitted late.