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

Homework 5: Mini-project
Due: December 7, 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" as either "very empathetic" (answers 4 and 5) or "not very empathetic" (answers 1, 2, and 3). You can use any of the other attributes in the dataset to make this prediction; however, you should not handpick the predictive features but let an algorithm select them.

Option 2: The Food & Drug Administration plans to conduct a survey among young people on the use of cigarettes, in order to decide whether it needs to introduce new regulations on the tobacco industry. You have been hired as a consultant and tasked with building a model that can predict how likely a person is to be in one of the following four categories: "never smoked," "tried smoking," "former smoker," and "current smoker." For this prediction, you will use the *Young People Survey* dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/). The target variable is "Smoking," and you can use any of the other attributes in the dataset to make this prediction; however, you should not handpick the predictive features but select them using an algorithm.

Option 3: You have been hired as a data scientist for a question-answering platform and tasked with using machine learning to predict whether some questions posted by users should be removed from the platform by using the Quora Insincere Questions dataset (https://www.kaggle.com/c/quora-insincere-questions-classification/data). You need to follow the Evaluation setup described here: https://www.kaggle.com/c/quora-insincere-questions-classification#evaluation. Note that this is an active Kaggle competition, and you are welcome to participate in the competition but you are not allowed to participate in the discussions until the semester is over.

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 should report on multiple methods that you have tried.

What to submit:

1) **Code (50%)**: 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. Make sure that your code is modular and that the training and test portions can be run separately, so that if we decide to run your test code, we can do it without running your training code.

2) **Write-up (50%)**: 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 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.

Be sure to answer the following questions:

(a) what is your data and task (including data preprocessing steps)?

(b) what ML solution did you choose and, most importantly, *why* was this an appropriate choice?

(c) how did you choose to evaluate success, including baselines, experimental setup (e.g., % train/dev/test), metrics?

(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?

3) **Optional** (up to 20% extra credit): Students will be given up to 20% extra credit for creating a *private* github repository for their project and depositing their code 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. Aesthetics and readability are important! 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, take a look at this introduction: https://guides.github.com/activities/hello-world/. You need to share your repository with all the Instructors (usernames: zynnel, chris-tran-16, zohrehovaisi), 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.

These three parts should be stand-alone and cannot be used as substitutes for each other. For example, when we grade part 2, we will not use any information from your github repository.

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

Honesty policy: While you are welcome to look at the discussions associated with the Kaggle tasks, copying code and solutions from Kaggle or any other source will be considered an academic integrity violation and will lead to 0% **in the course**, not just the assignment.