Foundations of AI
Instructor: Rasika Bhalerao
Assignment 2
Due September 23

Your second homework assignment is to choose a supervised machine learning task and use it to practice training and using supervised machine learning models.

Learning goals:

- Practice choosing the right machine learning model for the task
- Use a dataset to train a machine learning model and make predictions
- Determine the accuracy of the supervised machine learning model

What to do:

1. Choose an image-based dataset that interests you. You'll need one dataset which has images labeled by category, and one dataset which has a numerical label for each image
    a. These two can be the same dataset, if, e.g., each image has a category and a number. For example, you may choose a dataset of images of birds which are labeled by species, and a dataset of images of people's faces which are labeled by age (a number).
    b. If the dataset is very large, please use a small subset of it. You will not be graded on the size of the dataset (as long as there are at least 10 images).
    c. Kaggle's search is a good place to start.
    d. We'll spend some time in class looking for datasets. Post to Piazza if you found a good dataset!
2. For each dataset, split it into a training set and a test set (if it isn't already split for you).
    a. The validation set will be part of the training set.
    b. Note: the next two steps are very repetitive. Try to find ways to factor it out to minimize and simplify the code.
3. For the classification task:
    a. You will get practice with five models:
        i. Logistic regression
        ii. Support vector machine
        iii. Decision tree
        iv. Multi-layer perceptron
        v. (At least) one more model of your choice from Scikit-learn's available models
    b. For each model, take the training set and perform K-fold cross validation to choose the best values for the hyperparameters. Then, train the overall "final" model using those hyperparameters.
        i. Train the model using a few values for each hyperparameter, and calculate the performance on the validation set

        ii. Once you have the performance for a few different hyperparameter values, choose the best hyperparameter values, and train the overall "final" model using those hyperparameter values
    c. For each model, once you have trained the overall "final" model using the best hyperparameters, make predictions using the test set.
        i. Calculate the precision, recall, and F1 score (to be reported below)
        ii. Generate a confusion matrix to analyze the performance for each category
4. For the regression (numerical output) task:
    a. You will get practice with five models:
        i. Linear regression
        ii. Polynomial regression
        iii. Support vector machine
        iv. Multi-layer perceptron
        v. (At least) one more model of your choice from Scikit-learn's available models
    b. For each model, take the training set and perform K-fold cross validation to choose the best values for the hyperparameters. Then, train the overall "final" model using those hyperparameters.
        i. Train the model using a few values for each hyperparameter, and calculate the performance on the validation set
        ii. Once you have the performance for a few different hyperparameter values, choose the best hyperparameter values, and train the overall "final" model using those hyperparameter values
    c. For each model, once you have trained the overall "final" model using the best hyperparameters, make predictions using the test set.
        i. Calculate the Mean Squared Error (to be reported below)
        ii. Generate a scatter plot depicting the ground truth labels and the predicted labels, to find any patterns in the errors

What to turn in:

Please submit these via Gradescope:

- The Python with your code for training and testing the models
- Your confusion matrices and scatter plots
    - (These can be all in one file integrated among the questions below.)
- A text or pdf file with your answers to these questions:
    - For each classifier:
        - Report the precision, recall, and F1 score
        - Describe what the confusion matrix tells you about the performance (1 or 2 sentences)
        - Write a sentence or so about possible reasons why this model may or may not have been the right model for the task
    - For each regression (numerical output) model:
        - Report the Mean Squared Error

        - Describe what the scatter plot tells you about the performance (1 sentence)
        - Write a sentence or so about possible reasons why this model may or may not have been the right model for the task
    - How long did this assignment take you? (1 sentence)
    - Whom did you work with, and how? (1 sentence each)
        - Discussing the assignment with others is encouraged, as long as you don't share the code.
    - Which resources did you use? (1 sentence each)
        - For each, please list the URL and a brief description of how it was useful.
    - A few sentences about:
        - What was the most difficult part of the assignment?
        - What was the most rewarding part of the assignment?
        - What did you learn doing the assignment?
        - Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.