P556 Bonus Project

Fall 2022

Due: Sunday December 4 11:59pm, 2022

After exploring different machine learning models, this project will allow you to implement machine learning techniques on realistic classification problems. You will be able to understand how to select and train a model.

- This project is optional, which will be counted as bonus credit. Everyone will work individually. The last submission day is December 4 (there is no late submission).
- Object detection is a well-explored topic. In this project, your task is to recognize the object in the images, i.e., to create a classifier that decides the correct category of a given image. To make the task easier, we did some preprocessing on the original images: grayscale and downsize. Each image has one object only.
- The training set contains 100 categories with labels from 0 to 99. Each category has 5000 images. Each image is formatted as 28×28 pixels. There are 500K images in total. You can download the dataset from `https://bit.ly/p556bonus`.
- The test set covers the same categories in the training set without any unseen categories. The test set has 100K images. We didn't provide the labels of the test set.
- To help you get started, we provided a demo code using KNN($k = 5$) which can achieve 44.48% accuracy on the test set.
  - Hint 1: since the training set is very large, if your computer couldn't support this large size, you can do the sampling first.
  - Hint 2: since the dataset came from images, you can use image features to improve the accuracy.
- The score will be given based on your accuracy on the test set. The highest one will get full credit (see announcement for updated grading schema). Submissions with running errors, invalid submissions, or no submissions will get 0. All the others will be scaled by the MinMaxScaler. In other words, this project is a mini-contest. The better accuracy you achieve, the higher score you get.

- All the submissions have to rely on your own work. If the plagiarism system detects your code majorly from a public resource, you will get 0 and be reported.
- You need to submit 3 files on Canvas, not IU GitHub (username is your email ID).
  - One report document, name it as **project_username.pdf**: contain (1) a description of how you formulated the problem; (2) an illustration of your preprocessing, assumption, and model design; (3) results or observations; (4) any ideas of how to improve the accuracy.
  - One Python script, name it as **project_username.py**: we will run it on the test set. If your code relies on a pre-trained model, contact Keith (xiaoq@indiana.edu) for submission instructions.
  - One output file, name it as **project_username.txt**: contain your prediction on each image in the test set. In the txt file, each line has one label, so there should be 100000 lines. If the predicted label is 1, your output should be recorded as 1, not 01.
  - Check the filename before submission. If your filename is not matched, the auto-grader won't be able to run your script and you won't get credit.
- We may update further instructions, please follow the posts on Canvas.