Project - 2
CSE 456: Machine Learning Lab
Summer 2020

Al Mehdi Saadat Chowdhury

Assigned Date: 3rd October, 2020
Due Date: 10th October, 2020; 11.58PM

Basic Instructions:

In this section, you will find a general overview about the project, along with its grading policy.

- **Exam Format** Each project will have three components:
    1. A single code file that you will submit
    2. A report describing your project
    3. A viva-voce based on the project and everything taught in the class before the project's assigned date

  Marks Distribution 100 points for the project and report combined, 300 points for the viva. If project is incomplete, then points will be deducted both from the project and the viva.

- In here, general instructions for submitting your project is listed.

  What to submit You will submit a .zip file (not .rar, .7z etc) that contains a single code file (.py / .java / .cpp) and a PDF file containing the report. Details about each file is given in the Tasks sections below.

  Submission Deadline You are required to make **two** submissions.

    1. Temporary submission: 7th October 2020, before 11.58PM.
    2. Final submission: 10th October 2020, before 11.58PM.

  Submission Grading Only the final submission will be graded. You can also use **at most one bonus points** in this project. If you do not submit your project in the temporary submission date, then I'll assume you are using your **bonus point**. Your temporary submission date will automatically shift to 10th October 2020 then, and your final submission date will be 14th October 2020. After 14th October 2020, submissions will not be accepted. Please note that, there is no penalty grading deadline for this project.

  Number of Questions to be Submitted For each deadline, you will submit the following:

    1. Temporary Submission: Complete 50% of the task.
    2. Final submission: Submit the complete task.

- All submissions must follow the honor code mentioned in the class, which in summary is: **Do not copy your assignment from any sources including online resources, your classmates, friends, seniors and so on. You are NOT PERMITTED to do group study for this project**. Of course, you can use the textbook or any other books available to you **(except any solution manual)** for solving your project. If you are using any book, mention the name of the book in the acknowledgement section. Violating the honor code will incur the following penalties:

    1. The project will not be graded. You will loose 40% marks of this project.

2. In addition to the above penalty, you loose an additional 15% marks from the final grade. You will also be reported for further disciplinary actions to the head of the department.

- Understand that, even 1% copy will be regarded as violation of honor code. **All parties involved in the process will be penalized**. Also remember that, any suspected work will only be penalized after a thorough discussion with you. This discussion will start by giving you a chance to accept or reject my suspicion. If you actually were involved in this violation in any way and you accept it, then the second deduction of 15% marks will be waived for you. If you reject my suspicion, then further investigation will be conducted and you will also face a viva related to the **TOPICS** of the assignment. I kindly request you to obey the honor code so that we both can be saved from this unwanted hassle.

Project Questions

Project Description

Task 1 is a real-world project that asks students to compress images using unsupervised clustering methods, specifically k-means clustering. Compressing data is an active field of study, addressed both by researchers and software engineers in academia and in industry. Thus, this project I believe would boost your motivation to go further and explore more on machine learning.

However I also understand that, not everybody will be interested to a project of this sort that takes a lot's of dedication and hard work. Therefore, there will be an alternative problem that asks you to select a dataset on your own and apply some common classification algorithms.

Thus, through this project you will be exposed to both classification and clustering problems. You can choose any one task to answer. A student who solve both tasks will get some additional points at their viva-voce.

Task 1: Compress Images (100 points)

With this project file, you will be given two images.

1. For each image, apply k-means clustering to compress them. The value of k are given as: $2, 5, 10, 15, 20$. Thus, for each image you will have 5 compressed images as outputs (Input: 2 images, Output: $5 \times 2 = 10$ images). Your code should save all these compressed images to the current directory where the code has been run.

[Hints]

1. For basic image processing tasks, use the "PILLOW" library of python.
2. I hope you remember from our class that, the k-means clustering algorithm is sensitive to initialization. Therefore, you may need to run the same algorithm **for any particular k value** for several times, with different initial pixels as the cluster centroid. This is very very important. I suggest you to randomly initialize cluster centers (with a predefined random seed) taken from already existing image pixels and check if the image looks ok or not. If the compressed image doesn't look good, then take different cluster centers and repeat the process until you are satisfied with your compressed pictures. Remember to do this for every value of k and for both images.

Task 2: Classification (60 points)

This is the easiest task and completely open for you. For this task, do the following:

1. Download any classification dataset, either from UCI machine learning repository or from Kaggle or you can choose any other source for the dataset, as long as the dataset is about a real-world problem with at least 1000 training samples and 10 features. Make sure your dataset is different than any datasets picked by any of your classmates.

2. Use “scikit-learn” package of python to run the following classification algorithms to your dataset:
   (a) Logistic Regression
   (b) Support Vector Machine
   (c) Artificial Neural Network
3. In the project report, for each classification algorithm, you have to mention at least 10 different parameter settings that you have tried and you have to answer the following questions:
   (a) Write all different parameter settings that you have tried in your report.
   (b) Which classification algorithm worked best? For which settings?
   (c) How many hidden layers and how many hidden neurons is the best setting for your dataset?
   (d) Give links from where you have collected your dataset. Describe your dataset in complete details.

[Links]

1. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
2. Kaggle Datasets: https://www.kaggle.com/datasets