CSE343/ECE343: Machine Learning
Assignment - 4: CNN & KMeans
Max Marks: 25 (Programming: 15, Theory: 10) **Due Date**: 30/11/2022, 11:59 PM

Instructions

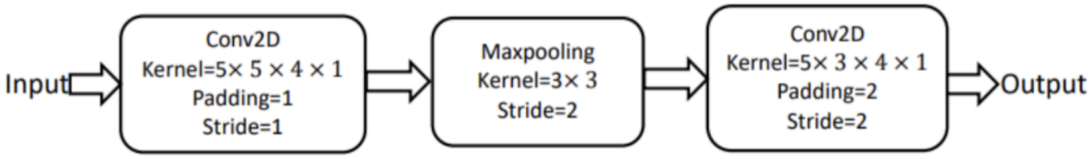
- Keep collaborations at high level discussions. Copying/Plagiarism will be dealt with strictly.
- Late submission penalty: As per course policy.
- Your submission should be a single zip file **2020xxx_HW4.zip** (Where *2020xxx* is your roll number). Include **all the files (code and report with theory questions)** arranged with proper names. A single **.pdf report** explaining your codes with results, relevant graphs, visualization and solution to theory questions should be there. The structure of submission should follow:

  2020xxx_HW4
  |– code_rollno.py/.ipynb
  |– report_rollno.pdf
  |– (All other files for submission)
- Anything not in the report will **not** be graded.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours at-least **two days before the deadline.**
- Your code should be neat and well commented.
- **You have to do either Section B or C.**
- **Section A is mandatory.**

---

1. (10 points) **Section A (Theoretical)**

   (a) (5 marks) Suppose you are given an input image with the dimensions of $15 \times 15 \times 4$, where 4 denotes the number of channels. The same is passed to a CNN shown below:-



The kernels are of shape h × w × I × O, representing height, width, number of input channels, and number of output channels, respectively.

(i) What is the output image size? **[2]**

(ii) What is the significance of pooling in CNN? **[1]**

(iii) Compute the total number of learnable parameters for the above CNN architecture (ignore bias) **[2]**

(b) (5 marks) You need to cluster these points in 3 clusters:

- point(3, 12),
- point(3, 7),
- point(9, 6)
- point(6, 10),
- point(8, 7),
- point(7, 6),
- point(2, 13)

These three points are the initial cluster points:

- point(3, 12),
- point(8, 7),
- point(2, 13).

Use k means to find the three cluster centers after the second iteration.

Defined as the distance function: D((pointX1,pointY1),(pointX2,pointY2))= mean average distance between the two points .

2. (15 points) **Section B (Scratch Implementation)**

Convolutional Neural Network (CNN)

Following are the packages you may need:

- NumPy
- Matplotlib
- np. random.seed(1)

A convolutional neural network is an artificial neural network that can be used in image recognition and also for processing pixel data.

(a) For this question, you are asked to implement a Convolutional Neural Network with the following functions. Following are the required functions to be implemented for the assignment from scratch:-

- Convolution functions: Convolution forward, backward, window, and zero padding [**2,2,2,2**]
- Pooling functions: Creating mask, distributing values, pooling forward and backward [**1,2,2,2**]

For every function, use appropriate inputs for the functions to run, and show the outputs according to your understanding for grading.

For all the above functions, you need to report in detail the working and use of each in the convolutional neural network.

OR

3. (15 points) **Section C (Algorithm implementation using packages)**

   You can use any library of your choice.

   Dataset folder: **Folder**

   For this question, use some unsupervised learning techniques where you would be required to find out which population segment has earnings that are greater than 50,000 every year.

   The goal of the question is: (i) to analyze which clusters will be *over-represented* in the *general population* (ii) which clusters are *over-represented* in the *more_than_50k* population after feature extraction, clustering, and feature selection.

   (a) Dataset Folder contains **population.csv** : General Population Data; **more_than_50k.csv** : Dataset for Population having more than 50k Annual Income; **Data Description.csv** : Contains descriptions of the features in the dataset. You are required to perform the steps below on the Dataset **Population.csv**:

   (b) (1 mark) **Preprocessing** : There are some missing data as '?', replace them with NaN, and remove columns with 30 percent more data missing. Make sure to use bins and convert numerical data into categorical data.

   (c) (2 marks) **Imputation, Bucketization, One-Hot Encoding** : Using mode for each feature, replace the missing values in each column in both the datasets. Bucketize and one hot encode the required features.

   (d) (4 marks) **Clustering** :Using k median clustering in the range[10,24] as values of k, draw the average within-cluster distance vs. a number of clusters graph. Using the graph, find the best value of k. Give a valid reason as to why it is the best value of k.

   (e) (3 marks) **Handling *more_than_50k* data** : Now, repeat all the same steps on the *more_than_50k dataset*.

   (f) (5 marks) **Compare the Population dataset with the *more_than_50k dataset*** :

- Compare the proportion of data in each cluster for the *more_than_50k* data to the proportion of data in each cluster for the general population.
- What kind of people are part of a cluster overrepresented in the *more_than_50k* data compared to the general population? For this, you may need to inverse transform PCA to map to original features and then analyze the value of the centroid of the clusters. You may use features with the highest magnitude for the first principal component to analyze the values for the centroid.
- Find out which clusters are over-represented in the *general population* vs *more_than_50k* population and vice versa.
- Similarly, analyze a cluster overrepresented in the *more_than_50k* data compared to the *general population*. (Give valid justifications for all these parts)