Lab Assignment - 7

CSL2010: Introduction To Machine Learning

Principal Component Analysis (PCA)
AY 2021-22, Semester-I

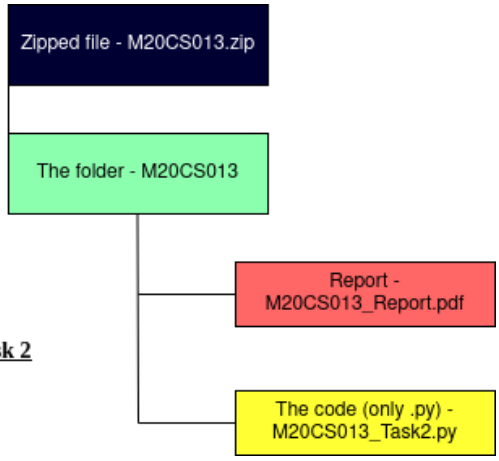
General Instructions

1. Clearly mention the assumptions that you have made, if any.
2. Make sure to add references to the resources that you have used while attempting the assignment.
3. **Any submission received in another format or after the deadline will not be evaluated.**
4. Plagiarism of any kind will not be tolerated and will result in zero marks.
5. Please do not copy paste code or screenshot, etc. in the report. Report should look like a technical document, containing plots, tables etc whenever necessary.

Instructions regarding the submission

1. There will be 2 different submissions.
2. In the first submission, named 'Assignment 7 - Task 1', you are supposed to answer question no. 1 and upload the same in ***.py*** format. Name the file as ***<Your_Roll_No>_Task1.py***. For eg, **M20CS013_Task1.py**. *Do not upload in any other format as it will not be evaluated*.
3. In the second submission, named 'Assignment 7 - Task 2', you need to upload a zip, which contains two files - question no. 2 in ***.py*** format, and the report for the entire assignment in ***.pdf*** format. *Again, do not upload in any other format as it will not be evaluated*. See the attached image to get a better clarity.

Format for Submitting Task 2



4. All are expected to follow the naming convention as given in the above image.
5. ***Do not*** download the .ipynb file, rename it as .py, and upload it. .ipynb files are not exactly in a readable form, and hence uploading it will only result in you receiving 0 marks for the same. You have an option to download .py file in google colab. Please use it to get the .py format.
6. Provide your colab file link in the report. Make sure that the file is shareable as view .
7. The report should include both task 1 and 2.

Task 1 (Due: 11:59 PM, 29 Sep 2021)

1. PCA (API)

   1.1 Download the MNIST dataset. You can use the mnist package for the same.

   1.2 Describe the downloaded dataset.

   1.3 Visualize any one of the images in the dataset by reshaping the data (28 x 28).

   1.4 Perform dimensionality reduction using the inbuilt PCA function from sklearn library by -

   i) Passing in the number of principal components. Find the amount of variance contributed by each component.

   ii) Passing in the variance to be retained. Find out the corresponding number of principal components required to have that amount of variance to be retained.

Task 2 (Due: 5:30 PM, 06 Oct 2021)

2. PCA (from scratch)

   2.1 Calculate the covariance matrix of the input dataset. (Since the dataset might be having a very large number of rows, you can use a randomly sampled subset of 5000 samples/images from the original dataset)

   2.2 Obtain the eigenvectors (or principal components) and eigenvalues from the covariance matrix. Report the first five eigenvalues.

   Take the following values - 10, 50, 100, 300, 700 as the number of principal components.

   Perform the following experiments on any two randomly sampled images from the dataset:

   2.3 Reconstruct the image for different values of principal components.

   2.4 Visualize the reconstructed images made from the previous step and compare them with the original image.

   2.5 Visualize the residual images by subtracting the reconstructed image from the original image (for each case).

   2.6 Find the reconstruction error (pixel-wise root-mean-square) for each sample and plot them for a different number of principal components.