Compulsory Assignment #1

Data Mining, Machine Learning and Deep Learning [KAN-CDSCO1004U]

Somnath Mazumdar
sma.digi@cbs.dk
Copenhagen Business School, Denmark

Deadline: {Refer digital exam for exact date and time}

Instructions

1. This compulsory assignment contains three main parts. Some parts may have more than one question. Please answer all the questions.
2. You must upload your solutions before the deadline to the digital exam `http://exam.cbs.dk/`.
3. Please use Python 3 to answer the below questions whenever necessary.
4. It is always a good practice to use comments extensively in your code so that it will be easy for other people to understand it.
5. If your answers involve any math, please feel free to use whatever format you like [Latex, Word, plain paper and pen], but make sure that it is readable.
6. The Python code for answering the questions can be either submitted as one single *jupyter* notebook or multiple *jupyter* notebooks. In addition to submitting your *jupyter* notebook/notebooks, please also export the Python code as a file from the *jupyter* notebook ('*File*' –> '*Download as*' –> '*Python (.py)'*) and upload the exported Python code file, in addition to *jupyter* notebook. Alternatively, suppose you are using any other IDE (Integrated development environment). In that case, you can also submit your code as one single python file (with .py extension) containing all the source code from different classes/modules/functions etc.

Assignments

Question 1 : Exploratory Data Analysis and Clustering

1.1 Exploratory Data Analysis (EDA)

Perform EDA on the given data set covering the following items.

1. Write a python code to count the total number of rows and columns in the given data. Your code also must find if there are any missing values in any of the columns or rows.
2. Write another python code to visualise data. Hints: You might look into *histograms*, *boxplots*, *scatterplots* few to name.

1.2 Clustering

Choose one of the clustering algorithms that were covered during lectures. Then choose desired columns from the data that you think is suitable for clustering. Explain the results, and you are free to use the graphs/plots and any other sort of visualizations.

Use the below data for **both** sub-questions.

- Data: `https://raw.githubusercontent.com/nick-edu/dmmldl/master/MobilePrice.csv`
- Description: `https://raw.githubusercontent.com/nick-edu/dmmldl/master/MobilePriceColumns.txt`

Question 2 : Principal Component Analysis

2 Principal Component Analysis (PCA)

1. Apply PCA on Olivetti faces dataset, while preserving 99% of the variance. Then compute the reconstruction error for each image.
2. Next, take some of the images you built using the PCA (previous step) and modify/add some noise to some of the images using techniques such as rotate, flip, and darken (use libraries such as scikit-image [1]) and look at their reconstruction error. You should also notice how much larger the reconstruction error is.
3. Finally, plot all the three respective reconstructed images side-by-side (original image, image after PCA, image after PCA + noise) and compare the results.

Download Olivetti faces data set from AT&T using below code

```
from sklearn import datasets
faces = datasets.fetch_olivetti_faces()
```

Question 3 : Written Assignment

Write an approximately 2-pages extended abstract to discuss the applications of **one** of the following techniques.

1. Dimension reduction
2. Clustering

To answer the above question, you can also do a tiny literature review to find out how your chosen techniques are used across various application domains. Also, note that these techniques are old and trendy and have laid the foundation for many other machine learning techniques.

Feel free to choose whatever direction you want to explore. Furthermore, based on your experience using the selected techniques in your assignment, you can also reflect on their capabilities and limitations in applying them to various datasets. Finally, write your reflections and critical comments rather than just reproducing what you found in the literature review.

Your report should conform to academic standards, and use APA referencing/citation style for your report.

References

[1] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014.