Assignment 11
K-Means, GMMs
EECE-6822 Machine Learning

Cris Ababei
Electrical and Computer Engr., Marquette University

1. Objective

The objectives of this activity include: (1) run several code examples that illustrate k-means and GMM models; and (2) conduct an investigation on the use of GMM to generate new images.

2. Prerequisite Readings

Murphy
- K-means, GMM: Murphy 21.3-21.5

Geron
- Ch.9: Unsupervised learning techniques

Raschka
- Ch.10: Working with Unlabeled Data – Clustering Analysis

3. Code Examples

Example 1: Clustering in SciKit-Learn

This is the example code from **Ch.8** from Aurelian Geron's book.
[*B3-Geron] Aurelien Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly, 2022.
Open in your google colab and go through the code. You should read first the chapter itself from the book, before going through the code.
https://github.com/ageron/handson-ml3/blob/main/09_unsupervised_learning.ipynb

The chapter discusses clustering, anomaly detection (also called outlier detection), and density estimation (estimating the probability density function (PDF) of the random process that generated the dataset).

Example 2: Working with Unlabeled Data – Clustering Analysis ←

This code example is from **Ch.10** from:
[*B3-Raschka] Sebastian Raschka, Yuxi Liu, and Vahid Mirjalili, Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python, Packt Publishing, 2022.
The source code (both Python code an Jupyter Notebook) is located at the GitHub repository. You should read first the chapter itself from the book, before going through the code.
https://github.com/rasbt/machine-learning-book/tree/main/ch10

To execute the code, I found the easiest usually to run directly the Python code (same code as in the notebook) in Anaconda Spyder. However, because the source code was developed with earlier versions of various libraries, you may need to fix little things here and there to make it work.

In this chapter, there are several concepts discussed including:

- Finding centers of similarity using the popular k-means algorithm
- Taking a bottom-up approach to building hierarchical clustering trees
- Identifying arbitrary shapes of objects using a density-based clustering approach

Example 3: k-means examples

These are examples suggested by K. Murphy on his github repository for the textbook:

https://github.com/probml/pyprobml/tree/master/notebooks/book1/21

Look on the above page for the notebooks that have kmeans (such as **kmeans_yeast_demo.ipynb**) or clustering in their file name and study them.

Example 4: mlm tutorials

Here, we look at several tutorials from **machinelearningmastery (mlm)**. Read the following tutorials and run the code where applicable:

--10 Clustering Algorithms With Python ←

https://machinelearningmastery.com/clustering-algorithms-with-python/

--Revisiting k-Means: 3 Approaches to Make It Work Better

https://machinelearningmastery.com/revisiting-k-means-3-approaches-to-make-it-work-better/

--K-Means Clustering in OpenCV and Application for Color Quantization

https://machinelearningmastery.com/k-means-clustering-in-opencv-and-application-for-color-quantization

--K-Means Clustering for Image Classification Using OpenCV

https://machinelearningmastery.com/k-means-clustering-for-image-classification-using-opencv/

--A Gentle Introduction to Expectation-Maximization (EM Algorithm)

https://machinelearningmastery.com/expectation-maximization-em-algorithm/

4. Assignment

In this assignment, you must change again the example from lecture #6:

Code 1: mlm-face-recognition-system-using-facenet-in-keras-and-an-svm-classifier.py

In that example we saw a "face identification" pipeline, working with a dataset with images of 5 famous people.

First, you need to edit the dataset in the following way:

--Create a sixth folder inside the dataset "5-celebrity-faces-dataset/train" that you shall name using your last-name. Place in that folder 10 images of yourself. These images should be showing only you alone in the photo, and should show your face - in order for **Code 1** to work smoothly. Also, these images/photos of yourself should perhaps be rescaled so that they are not larger than 2000 pixels in either direction – you can use free image/photo editors to resize images or to crop yourself out of photos with more people in so that you create images with only you in.

--Create a sixth folder inside the dataset "5-celebrity-faces-dataset/val" as well, and place inside it just 1 photo of yourself.

--Delete from all the other folders inside train/ of the dataset images randomly so that you keep only 10 images in every person's folder. Also delete from all the other folders inside val/ images so that you keep only 1 image for every person.

--At this end of this step, rename the dataset folder to "**6-celebrity-faces-dataset**" – which will contain train/ with six folders, and 10 images in each folder; as well as val/ with six folders, and 1 image in each folder.

NOTE: I assume (given that we leave in an era when everybody has a smartphone with a camera…) you all have 11 photos of yourselves, preferably from different times and different light conditions.

Second, you must change the code of the original pipeline of the example **Code 1**, which for convenience is described again here:

1--**MTCNN** model to do face-detection and cropping of those detected faces into 160x160 images.

2--**FaceNet** model to pre-process detected faces to create face embeddings

3--**Linear SVM** to classify any given face-embedding as one of the 5 people

Your task here is to remove steps #2 and #3 above and replace them with the steps from the end of the following Jupyter Notebook (of the current lecture #8), from section "**Example: GMMs for Generating New Data**":

Code 2: 05_12_Gaussian_Mixtures.ipynb

Basically, you need to apply PCA so that you preserve say 95% of the variance, then GMM, then inverse PCA to generate new images.

At this stage, after you apply step 1 to crop all faces for all 6 people, your dataset should combine all 10 faces from train/ folder of all 6 people in one input dataset **X**. It is this dataset (of 60 face images) that we will use for the clustering.

You should also actually change the first step of the pipeline, "1--**MTCNN** model", so that the face images are cropped to initially 96x96 pixels. Hopefully, this image size will not increase too much computational runtime. If the runtime increases too much, feel free to decrease the image size to something smaller in steps, for example, reduce to 72x72, then, 64x64, etc.

In your report, you should have included at least:

--The plot like the one from **Code 2**, that shows "*the AIC to get a gauge for the number of GMM components we should use*", based on which you must select your own number of components. Please add axis labels to your plot (do not leave it without axis labels like it is done in the notebook)!

--A new generated image of yourself, and one new image generated of Seinfeld.

5. Deliverables

You must write (typed) a report and upload it as a PDF file on D2L. The report should be named "**hw11_report_LastName.pdf**". You should also create a .zip archive with all your code and implementations of all parts of the assignment. Upload also this archive .zip file with the name "**hw11_implementation_code_LastName.zip**" to D2L. Hence, your D2L should contain two items: the report and the .zip file. **Do not include the report inside the .zip and upload only the .zip. They should be two separate items!**

The report should include the following sections and subsections. Make sure section titles are in bold font and pages are numbered.

1) **Title + course info + your name**
2) **Summary.** Describe in one paragraph what the objective of the assignment is.
3) **Description of Experiments and Discussion.** Describe the experiments you did. All tables and figures should be numbered and should have captions. All plots in all figures should have axes labels and titles. Present a meaningful discussion with the interpretation of the results you obtained. Explain if you expected the results or not; discuss the intuition behind it.
4) **Conclusion.** Present your conclusions; highlight what are your main takeaways that you learned from this assignment. Describe what issues you encountered and how you solved them.
5) **References.** Include all references that you used, as a numbered list. Cite them in the report itself; do not just list them! If your report has References that are not numbered and cited in the report, points will be deducted!