**COMP257 Unsupervised and Reinforcement Learning**

**Section 003**

**Assignment 4**

Github: <https://github.com/venussirada/COMP257/tree/main/Assignment4>

**1. Output**

1. Dataset

A collage of different faces

Description automatically generated

1. Dataset shape after applying PCA

A number and a number

Description automatically generated with medium confidence

A graph with a curve

Description automatically generated

1. Covariance type and score

A white background with black numbers and text

Description automatically generated

A graph with numbers and a bar chart

Description automatically generated with medium confidence

1. Optimal number of clusters

A screenshot of a computer

Description automatically generated

A graph with a line and a red line

Description automatically generated

1. Hard clusters

A number grid with numbers and symbols

Description automatically generated with medium confidence

1. Soft clusters

A screenshot of a computer

Description automatically generated

1. New faces

A close-up of a person's face

Description automatically generated

1. Modify images

A collage of different images of a person's face

Description automatically generated

1. Log-likelihood scores of original and modified images

A number and numbers on a white background

Description automatically generated

1. Anomaly detection of the model

Answer The log-likelihood scores generated by the Gaussian Mixture Model shows that the model successfully distinguished between normal and modified images. The normal images yielded higher scores, indicating a stronger fit to the model’s learn distribution. While the modified images showed significantly lower scores, suggesting that they are different from the expected pattern.

**Code Explanation**

Answer I started by loading the dataset and splitting it into training, validation, and test sets using **StratifiedShuffleSplit** to ensure balanced labels across splits. Next, I applied **PCA** on the training data to retain 99% of the variance and reduce the dimensionality.

I then fitted the **Gaussian Mixture Model** with different covariance types (full, tied, diag and spherical) and calculated **AIC** and **BIC** scores. The ‘full’ covariance type had the best **AIC** score, while ‘diag’ type had the best **BIC** score. And I tested different numbers of clusters to find the optimal numbers of clusters for **AIC** and **BIC**.

After selecting ‘diag’ type and 2 clusters based on **BIC**, I fitted the model to find hard clustering and soft clustering assignments. I then generated 5 new samples using the model.

Lastly, I selected 5 samples from training dataset and applied modifications (rotation, flip, and darken). Then I computed log-likelihood scores for both original images and modified images to test the model’s ability to detect anomalies.

**Lesson Learned**

1. **Covariance type selection**: The optimal covariance type can vary between AIC and BIC. For example, ‘full’ may yield the best AIC score, while ‘diag’ might perform better for BIC.
2. **GaussianMixture sampling**: The method sample() can generate new random samples from the fitted Gaussian distribution. Ref: <https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html>
3. **Anomaly detection**: The log-likelihood score is a key metric in determining whether an image belongs to the model’s learn distribution. Significantly lower scores show that the image is an anomaly.