ECS708 Machine Learning Assignment 2: Clustering and MoG

Aim

The aim of this assignment is to become familiar with clustering using the Mixture of Gaussians model.

1 Introduction

For this lab, we will use the Peterson and Barney's dataset of vowel formant frequencies. (For more info, look at Classification of Peterson & Barney's vowels using Weka. - a copy of this article is at QMplus)

More specifically, Peterson and Barney measured the fundamental frequency $F0$ and the first three formant frequencies ($F1 - F3$) of sustained English Vowels, using samples from various speakers.

The dataset can be found in the QMplus, in the files of Assignment 2, in the folder named "data", at the file "PB_data.npy". Load the file. In your workspace, you will have 4 vectors ($F0 - F3$), containing the fundamental frequencies ($F0$, $F1$, $F2$ and $F3$) for each phoneme and another vector "phoneme_id" containing a number representing the id of the phoneme. The arrangement of the data is as follows:

phoeneme ID	$F0$	$F1$	$F2$	$F3$
1	xxx	xxx	xxx	xxx
...	xxx	xxx	xxx	xxx
10	xxx	xxx	xxx	xxx
...	xxx	xxx	xxx	xxx
N	xxx	xxx	xxx	xxx

In the exercises that follow, we will use only the dataset associated with formants $F1$ and $F2$.

2 MoG Modelling using the EM Algorithm

Recall the following definition of a Mixture of Gaussians. Assuming our observed random vector is $\mathbf{x}$, a MoG models $p(\mathbf{x})$ as a sum of weighted Gaussians. More specifically:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \frac{p(c_k)}{(2\pi)^{D/2} \det\left(\mathbf{\Sigma}_k\right)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k{}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right) \quad (1)$$

Where D is the dimension of vector $\mathbf{x} \in \mathbb{R}^D$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $p(c_k)$ are the mean vector, covariance matrix, and the weight of the k-th gaussian component, and K is the total number of gaussian components used.

Task 1:
Load the dataset to your workspace. We will only use the dataset for F1 and F2, arranged into a 2D matrix where the first column will be F1 and the second column will be F2. Produce a plot of F1 against F2. (You should be able to spot some clusters already in this scatter plot.).

Include in your report the corresponding lines of your code and the plot. [20 points]
Task 2:
Train the data for phonemes 1 and 2 with MoGs. You are provided with python files task_2.py and plot_gaussians.py. Specifically, you are required to:

1. Look at the task_2.py code and understand what it is calculating. Pay particular attention to the initialisation of the means and covariances (also note that it is only estimating diagonal covariances).
2. Generate a dataset X_phoneme_1 that contains only the F1 and F2 for the first phoneme.
3. Run task_2.py on the dataset using K=3 Gaussians (run the code a number of times and note the differences.) Save your MoG model: this should comprise the variables mu, s and p.
4. Run task_2.py on the dataset using K=6
5. Repeat steps 2-4 for the second phoneme

Include in your report the lines of code you wrote, and results that illustrate the learnt models. [20 points]
Task 3:
Use the 2 MoGs (K=3) learnt in task 2 to build a classifier to discriminate between phonemes 1 and 2. Classify using the Maximum Likelihood (ML) criterion (feel free to hack parts from the MoG code in task_2.py so that you calculate the likelihood of a data vector for each of the two MoG models) and calculate the miss-classification error. Remember that a classification under the ML compares $p(\mathbf{x}; \boldsymbol{\theta}_1)$, where $\boldsymbol{\theta}_1$ are the parameters of the MoG learnt for the first phoneme, with $p(\mathbf{x}; \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_2$ are the parameters of the MoG learnt for the second phoneme.

Repeat this for $K = 6$ and compare the results.

Include in your report the lines of the code that your wrote, explanations of what the code does and comment on the differences on the classification performance [20 points]

Task 4:

Create a grid of points that spans the two datasets. Classify each point in the grid using one of your classifiers. That is, create a classification matrix, $\mathbf{M}$, whose elements are either 1 or 2. $M(i, j)$ is 1 if the point $x1$ is classified as belonging to phoneme 1, and is 2 otherwise. $x1$ is a vector whose elements are between the minimum and the maximum value of F1 for the first two phonemes, and $x2$ similarly for F2. Display the classification matrix. Include the lines of code in your report, comment them, and display the classification matrix. [20 points]

Task 5: In the code of task_5.py a MoG with a full covariance matrices is fit to the data. Now, create a new dataset that will contain 3 columns, as follows:

$$X = [F1, F2, F1 + F2] \tag{2}$$

Fit a MoG model to the new data. What is the problem that you observe? Explain why.

Suggest ways of overcoming the singularity problem and implement them.

Include the lines of code in your report, and graphs/plots so as to support your observations. [20 points]

Write a report about what you have done, along with relevant plots. Save the solution in a folder with your ID. Create and submit a .zip that contains:

1. **all of your code and**
2. **a copy of your report. The report should be in .pdf format named as ml_lab1_part1_StudentID.pdf**