CO–495: Advanced Statistical Machine Learning & Pattern Recognition

Coursework #1: Mixture Models

Exercise I

The purpose of this exercise is to get you acquainted with Gaussian Mixture Models (GMMs) and in particular with how to properly programme a GMM. You are given a set of unlabelled, i.i.d. points $\texttt{X}$ which are drawn from a GMM. The points are depicted in Fig.1:

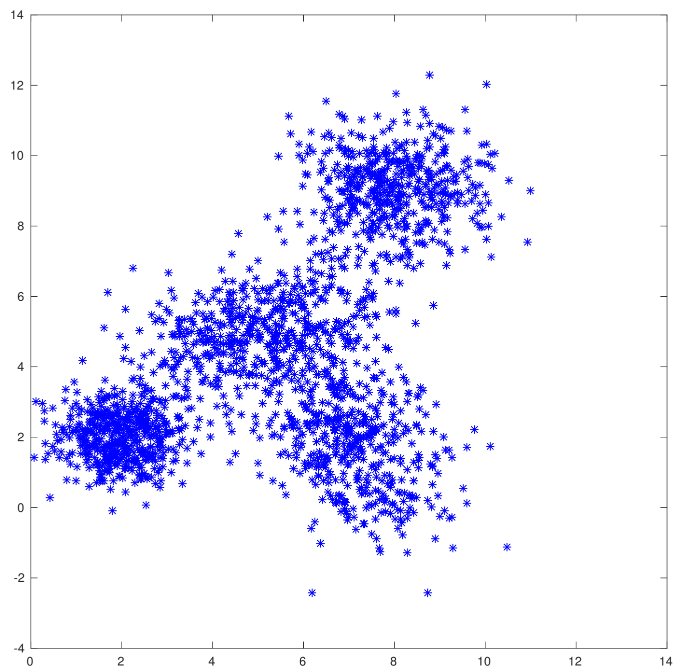

Figure 1: Scatter plot of the 2-D data $\texttt{X}$

We can infer from Fig.1 that there most probably exist 4 Gaussians in the graph. Your task is the following:

a) From left to right in Fig.1 and without doing any calculations, what is the form of the covariance matrix for each Gaussian (i.e., is it an isotropic[1] or an anisotropic one)? Why? Briefly explain.

[1]a covariance matrix $\mathbf{A}$ is called isotropic if it is proportionate to the identity matrix, i.e., $\mathbf{A} = \lambda\mathbf{I}$

b) For the coding part of the coursework, you may work *either* in matlab *or* in python. Below follow instructions for matlab and python submissions, respectively.

matlab instructions: Attached in this `zip` file there is also the dataset `X.mat` and four matlab files: `GMMDemo.m`, `GMMInit.m`, `GMMEStep.m` and `GMMMStep.m`.

i) `GMMDemo.m` contains the skeleton for the GMM you are about to code and defines the form of the final output. You should read it carefully but you *should not* edit it.

ii) `GMMInit.m` is a function that *needs to be filled in* and will be used for initialisation of the parameters. You may at first try to initialise the parameters randomly and see how the EM algorithm behaves. What can you notice and why? Briefly explain. The best option would be to use `k-means` on the existing dataset to find the initial values for the mixing coefficients and sample means and covariances.

iii) `EMEStep.m` is a function that *needs to be filled in* and will be used for the E-Step of the EM algorithm.

iv) `EMMStep.m` is a function that *needs to be filled in* and will be used for the M-Step of the EM algorithm.

After the execution of `GMMDemo.m`, a new spreadsheet file named `params.xslx` will be saved in your current directory. It will contain the final values for the parameters. Every function to be filled in by you should be accompanied with brief comments where you will concisely explain your steps. For this part, together with your brief explanations when requested (which you will include in a pdf named "answers.pdf"), you should *only* additionally provide the following files: `GMMInit.m`, `EMEStep.m`, `EMMStep.m`, `params.xslx`.

python instructions: Attached in this `zip` file there is also the dataset `X.mat` and a python notebook named `CW1.ipynb`. Your task is to fill in the empty functions in the notebook and add brief comments where you will concisely explain your steps. For the function `GMMInit`, you may at first try to initialise the parameters randomly and see how the EM algorithm behaves. What can you notice and why? Briefly explain. After the execution of function `GMMDemo` a new `csv` file named `params.csv` will be saved in your current directory. For this part, together with your brief explanations when requested (which you will include in a pdf named "answers.pdf"), you should *only* additionally provide the following files: `CW1.ipynb`, `params.csv`.

(80 marks)

Exercise II

Consider a special case of a GMM in which the covariance matrices Σ_k of the components are all constrained to have a common value Σ. Derive analytically the EM equations for maximising the likelihood function under such a model and include your solution in "answers.pdf".

(20 marks)