BERNARD BARUCH COLLEGE

MTH 9898 BIG DATA IN FINANCE (SPRING 2021)

Final Project

The first part of the project is an illustration of how a Gaussian Mixture Model (GMM) can fail when used as a clustering and a density estimation algorithm. In the second part we will generalize the GMM algorithm to handle the two moon example.

The output of the project should be a pdf file with an accurate description of the adopted procedure in part 1 and 2 as well comments to the figures. No jupyter notebook snipped should be included in the pdf file. The code you use should be submitted in a separate jupyter notebook file and should be readable and with comments.

Part one.

- *GMM as clustering algorithm*: Figure 1 shows how different clustering algorithms perform on the two-moon example two-moon example and other examples included in the scikit library.

  We will focus our attention on the first and last column of the figure obtained with k-means and GMM respectively. Intuitively on can think that the data are divided into two clusters each one represented by a moon. Can you explain why the two algorithms fails to detect the two clusters correctly?

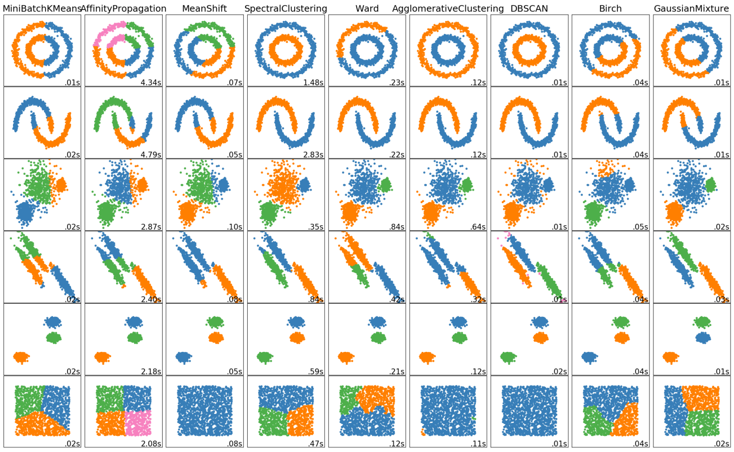

Figure 1. Comparison of cluster techniques for the two-moon example and other distributions. Points are colored according to the cluster they belong to.

- *GMM as density estimation algorithm*: Code GMM from scratch implementing the EM algorithm. Use your code to estimate the distribution of weekly return of the S&P 500 index. Show that the obtained distribution is characterized by fat tails (a log plot is sufficient). Fit a Gaussian mixture model and show that the model is unable to reproduce correctly the tails of the histogram.

Part two. This paper describes a very sophisticated way to formulate a GMM based algorithm to cluster correctly the two-moon example. You don't need to understand every detail of the paper as the mathematics is pretty heavy. However, the authors provide the intuition behind their algorithm together with a detailed pseudo-code describing how to modify the EM algorithm we saw in class to maximize a regularized log-likelihood they define in the text. The idea for you is to summarize the important part of the paper, code their algorithm and verify on the two moon example that it works.