**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO ĐỒ ÁN 2**

Gaussian Mixture Model

Expectation maximization (EM) clustering

**GVHD: Thầy Trần Nhật Quang**

**SVTH: Nguyễn Khôi -- 16110125**

**Trần Quang Anh -- 16110009**

Mục lục

[TÌM HIỂU VỀ ĐẠO VĂN 3](#_Toc9535466)

[I. Gaussian Mixture Model 4](#_Toc9535467)

[1) Gaussian là gì ? 4](#_Toc9535468)

[2) Mô hình hỗn hợp Gausse (Gaussian Mixture Model) 4](#_Toc9535469)

[II. Maximum Likelihood trên một GMM 8](#_Toc9535470)

[III. GMM Parameter Estimation thông qua EM 10](#_Toc9535471)

[IV. Đánh giá thuật toán 15](#_Toc9535472)

[a) Thư viện sklearn 15](#_Toc9535473)

[b) Code ví dụ 16](#_Toc9535474)

[DANH MỤC CÁC TRANG WEB THAM KHẢO 18](#_Toc9535475)

# TÌM HIỂU VỀ ĐẠO VĂN

Theo Merriam-Webster Online Dictionary, đạo văn nghĩa là:

* Ăn cắp và hình thành những ý tưởng hay ngôn từ mới khởi nguồn từ ý tưởng của ai đó
* Sử dụng sản phẩm của một ai đó mà không công bố nguồn
* Giới thiệu một ý tưởng hay sản phẩm mới được chuyển hóa từ một nguồn đã có từ trước

Ở Mỹ, câu hỏi được đưa ra là “Từ ngữ và các ý tưởng có thực sự bị ăn cắp?” và Luật pháp Mỹ trả lời là có. Việc diễn đạt một ý tưởng độc đáo nào đó cũng được xem là một thành quả của trí tuệ và được bảo vệ theo Luật tác giả tương đương như việc bảo vệ một phát minh hoàn toàn mới.

Có rất nhiều cách để đạo văn. Việc “hô hoán” thành quả của một ai đó là của mình là trường hợp đầu tiên, chỉ riêng việc **sao chép** từ ngữ hay ý tưởng của một ai đó mà không ghi rõ nguồn cũng có thể được xem là đạo văn. Một khi đã có ý định sử dụng trích dẫn từ thành quả sáng tạo và lao động của người khác, bạn phải ghi rõ nguồn và tên trích dẫn từng đoạn một. Tuy nhiên thậm chí trích dẫn cụ thể nhưng lại sao chép quá nhiều cũng sẽ là một bằng chứng của đạo văn.

Hầu như tất cả các trường hợp “bị mang tiếng” đạo văn đều có thể tránh được, ít nhất là nếu bạn ghi rõ nguồn.

**Lời cam kết:** “Bọn em xin cam đoan đồ án này do chính chúng tôi thực hiện. Bọn em không sao chép, sử dụng bất kỳ tài liệu, mã nguồn của người khác mà không ghi rõ nguồn gốc. Bọn em xin chịu hoàn toàn trách nhiệm nếu vi phạm.”

# I. Gaussian Mixture Model

## 1) Gaussian là gì ?

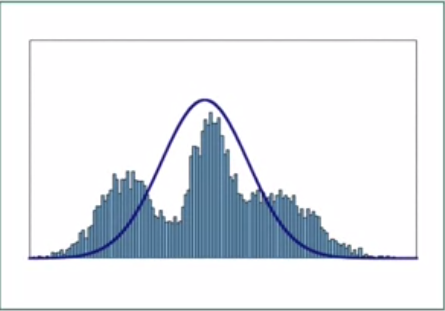
**Phân phối chuẩn**, còn gọi là **phân phối Gauss** hay (Hình chuông Gauss), là một phân phối xác suất cực kì quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số *vị trí* (giá trị kì vọng μ) và *tỉ lệ* (phương sai σ2). (theo Wikipedia)



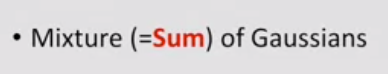
## 2) Mô hình hỗn hợp Gausse (Gaussian Mixture Model)

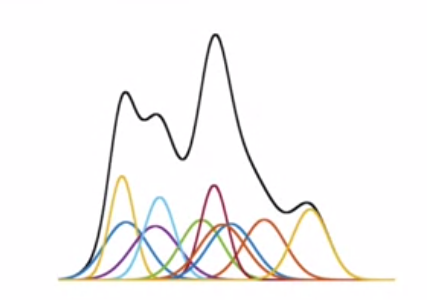
Trước đây ta đã từng học qua các thuật toán phân loại dữ liệu như logistic, phân cụm dữ liệu như k-mean, KNN,… Thì Gaussian Mixture Model cũng là một thuật toán phổ biến và cực kì quan trọng trong bài toán phân cụm dữ liệu.

Trong thực tế, thường chúng ta sẽ gặp những mô hình dữ liệu như thế này



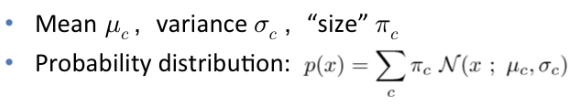
Việc chỉ sử dụng Gaussian để fit dữ liệu như thế này hoàn toàn không phù hợp, do mô hình dữ liệu kiểu này có nhiều trạng thái và nó hoàn toàn thiếu đi sự đối xứng. Thay vào đó ta sử dụng Gaussian Mixture Model (GMM). GMM nói cho đơn giản, nó là tổng của tất cả các Single Gaussian



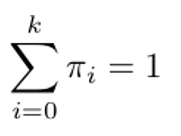


Đường màu đen chính là tổng của các Gaussian màu được tạo ra ngẫu nhiên.

Nếu ta gọi N là một Gaussian thành phần có giá trị kì vọng là μc và phương sai là σc thì lúc này công thức của GMM có dạng:

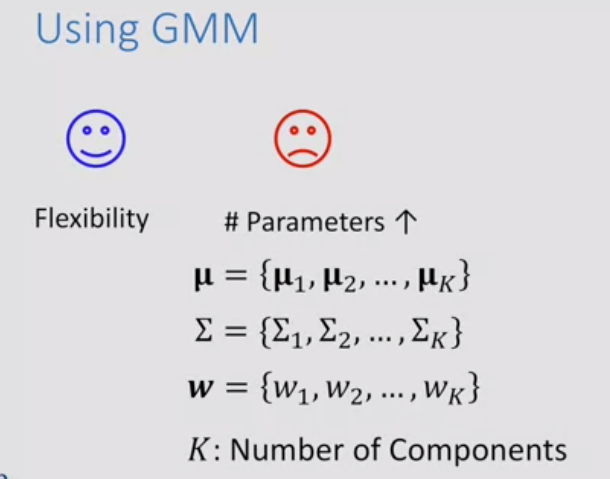


Trong đó, tổng trọng số của các Gaussian mà trọng số được xác định bởi mỗi phân bố là:



Tổng sigma của c sẽ đi từ 1 đến k với k là tổng số các Gaussian.

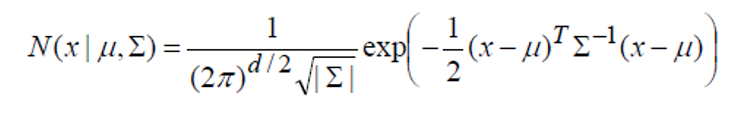
Việc sử dụng GMM phức tạp hơn do nóp có nhiều tham số hơn và có thêm một tham số mới gọi là trọng số và nó biến một Gaussian thành một tham số.



**Multivariate Gaussian Mode:** Thông thường dữ liệu đầu vào của chúng ta luôn là dữ liệu nhiều chiều, x được biểu thị là một vecto gồm tất cả các dòng mà mỗi dòng là tất cả các thuộc tính của một giá trị đầu vào.



Và công thức của Gaussian lúc này được định nghĩa như sau:



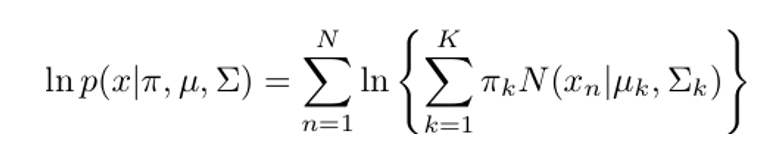
# II. Maximum Likelihood trên một GMM

**Ước lượng hợp lý cực đại** (trong tiếng Anh thường được nhắc đến với tên **MLE**, viết tắt cho **Maximum Likelihood Estimation**) là một phương pháp trong thống kê dùng để ước lượng giá trị tham số của một mô hình xác suất dựa trên những dữ liệu quan sát được. Phương pháp này ước lượng các tham số nói trên bởi những giá trị làm cực đại hóa likelihood function. Các ước lượng thu được cũng được viết tắt là MLE (**Maximum Likelihood Estimates**).

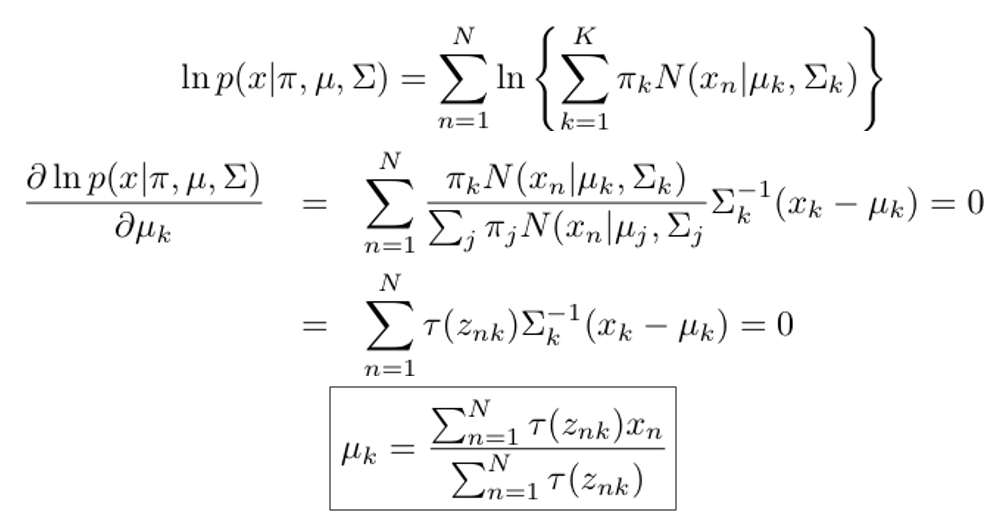
MLE được sử dụng chung với các phân tích thống kê khác. Lấy ví dụ khi chúng ta muốn ước lượng chiều cao nói chung của chim cánh cụt cái trưởng thành, nhưng lại không thể nào đo được chiều cao của tất cả chim cánh cụt trong một quần thể (do ràng buộc về thời gian hoặc chi phí). Bằng việc giả sử chiều cao trong quần thể được phân pjhối chuẩn với các tham số giá trị kì vọng và phương sai chưa biết, chúng ta chỉ cần khảo sát chiều cao của một vài cá thể mẫu trong quần thể và dùng MLE để ước lượng các tham số này. Khi nhìn vào các chiều cao mẫu đã thu thập, có thể hình dung là, phương pháp MLE sẽ tìm ra cách giải thích hợp lý nhất cho những chiều cao nhận được đó.

Theo quan điểm của suy luận Bayes, MLE là một trường hợp đặc biệt của **Maximum A Posteriori estimation (MAP)**, phương pháp đưa ra giả thiết về phân phối đều của các tham số. Trong suy diễn tần số, MLE lại là một trong số rất nhiều các phương pháp ước lượng tham số mà không cần dự đoán trước về phân phối. Việc dự đoán trước này được tránh bằng cách không khẳng định về xác suất của các tham số mà chỉ khẳng định về xác suất của các ước lượng, do các ước lượng đã được định nghĩa đầy đủ với các dữ liệu quan sát được và mô hình xác suất. (theo wikipedia)

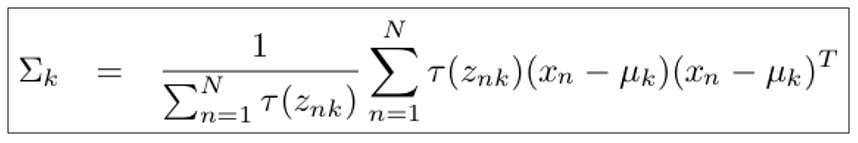
Như thường lệ, ta định nghĩa Likelihood function:



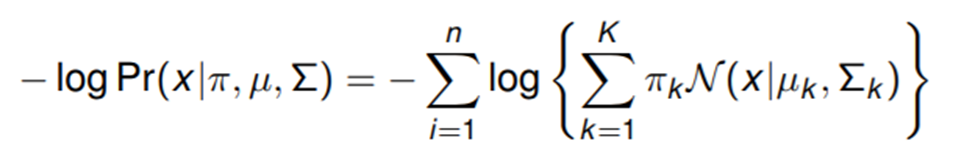
Lấy đạo hàm theo giá trị trung bình µ và giải đạo hàm đó bằng 0 ta được:



Tương tự với phương sai, ta tính ra được:



Đổi dấu 2 vế Likelihood function ta sẽ có được hàm Loss function J của GMM:

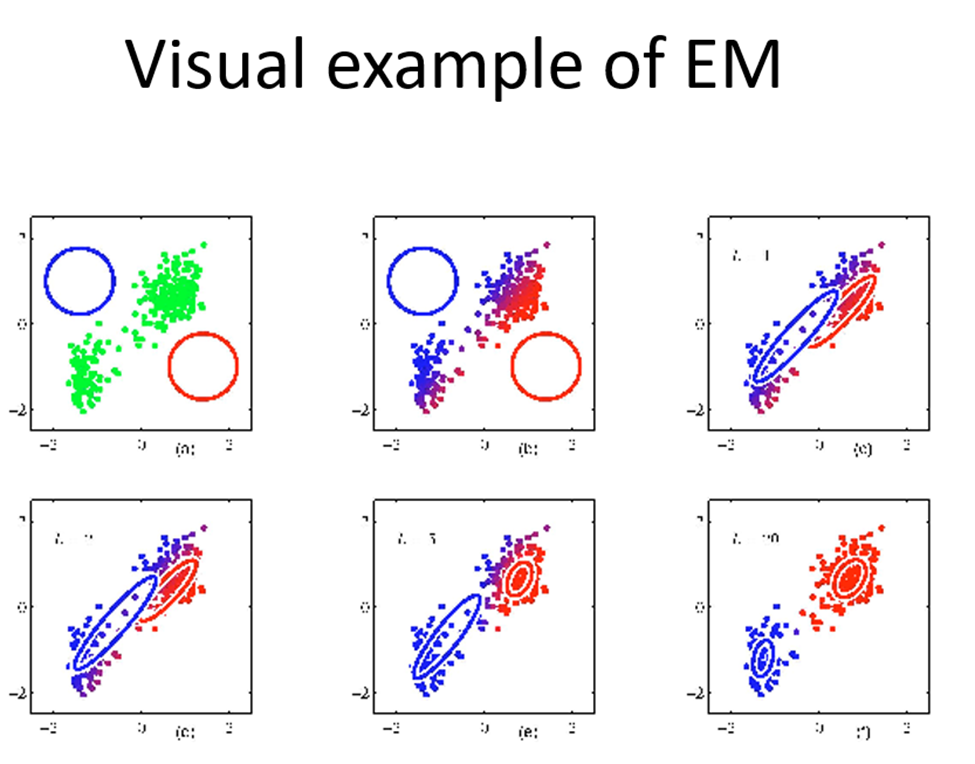


Hàm này rất khó để tính toán. Tuy nhiên, nếu như Linear Regression có Gradient Descent thì GMM cũng có một phương pháp để tối ưu hóa bằng thuật toán EM (Estimation Maximization).

# III. GMM Parameter Estimation thông qua EM

Trong khai thác dữ liệu, phương pháp tối đa hóa kì vọng (EM) là thuật toán gom nhóm (clustering) dữ liệu (như k-means) được dùng trong tác vụ khám phá tri thức (knowledge discovery).

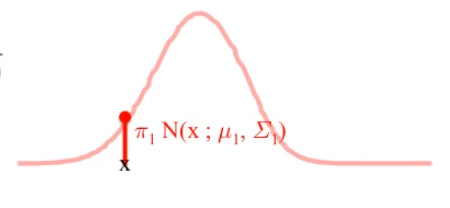
Trong thống kê, thuật toán EM sử dụng lặp và tối ưu hóa khả năng (likelihood) nhìn thấy dữ liệu quan sát (seeing observed data) thông qua việc ước lượng tham số (parameters estimation) cho mô hình thống kê (statistical model) cho các biến không quan sát được (unobserved variables).

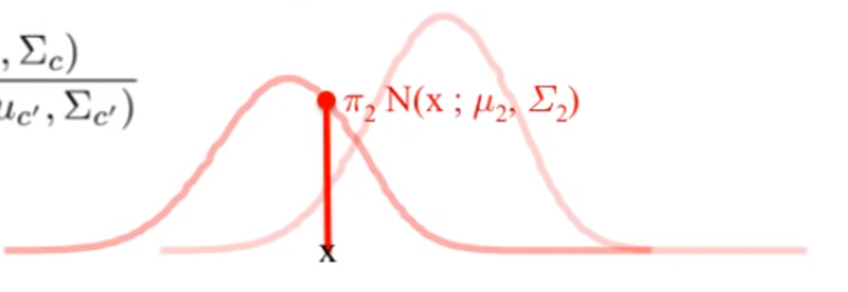


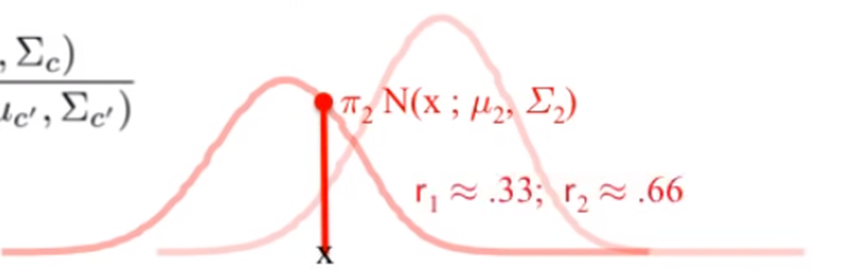
Thuật toán EM bao gồm hai bước, bước E hoặc bước Kỳ vọng và bước M hoặc bước Tối đa hóa.

**E-Step:**

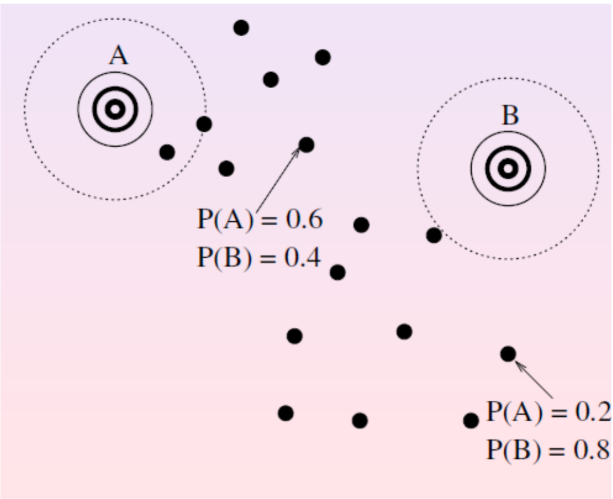
Với mỗi điểm dữ liệu xi, tính toán xác suát mà nó thuộc về mỗi cụm



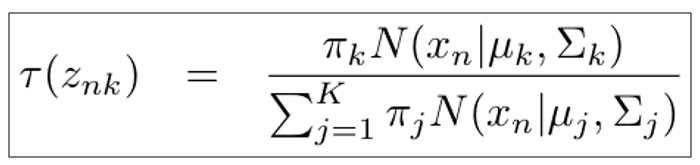




Hoặc để dễ hình dung hơn ta nhìn vào dữ liệu 2d dưới đây:

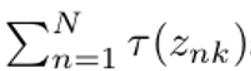


Công thức tính xác suất xn thuộc về cụm c là:



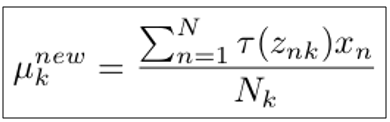
**M-Step:** Với mỗi cluster (Gaussian) z = c = k, thay đổi các tham số để tối đa hóa khả năng của dữ liệu, tức là đưa các Gaussian ngày càng gần các khu vực có nhiều điểm dữ liệu thuộc của mình nhất.

Cập nhật tổng xác suất của tất cả các I được phân phối cho cụm k:

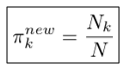
Nk = 

Trong, N là tổng số dữ liệu x, hàm số trong tổng sigma là xác xuất điểm dữ liệu n thuộc Gaussian k.

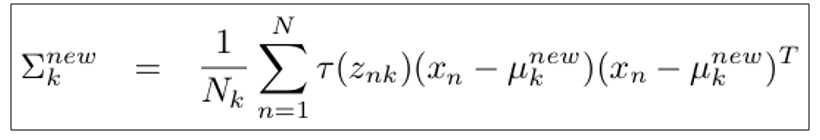
Dưới đây là công thức cập nhật lại tham số giá trị trung bình của Gaussian k.

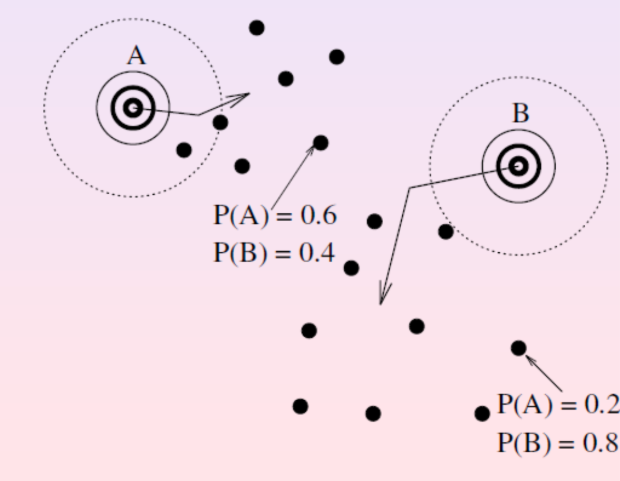


Sau đó cập nhật lại trọng số của các Gaussian:



Cập nhật lại phương sai dựa trên các tham số mới:





Qúa trình lặp lại này sẽ diễn ra cho đến khi:

+ Đến khi likelihood hội tụ.

+ Hội tụ đến tối ưu của hàm log likelihood.

# IV. Đánh giá thuật toán

So với thuật toán k-mean, GMM linh hoạt hơn rất nhiều, nếu như k-mean 1 điểm dữ liệu chỉ được gán cho một cụm thì với GMM, dữ liệu có thể thay đổi vào các cụm tùy theo tính toán xác suất nó nằm trong một cụm thay đổi qua mỗi vòng lặp.

Ưu điểm:

+ Điểm then chốt của EM đó là sự dễ hiểu và cài đặt dễ dàng. Thêm vào đó, không chỉ tối ưu hóa được các tham số của mô hình, nó còn có thể dự đoán cho các dữ liệu bị thiếu xuyên suốt quá trình lặp.

Phương pháp này hữu ích cho tác vụ gom nhóm và hình thành mô hình qua các tham số. Khi biết được các nhóm và tham số của mô hình, ta có thể suy luận ra điểm dữ liệu mới thuộc về nhóm nào.

Hạn chế:

+ Thứ nhất, EM chạy nhanh ở các vòng lặp ban đầu nhưng chậm hơn ở các vòng lặp sau.

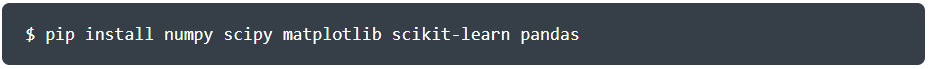
+ Thứ hai, EM không phải lúc nào cũng tìm được tham số tối ưu và bị mắc kẹt ở điểm tối ưu cục bộ (local optima) thay vì toàn cục (global optima).

V. Ví dụ nhỏ về GMM bằng python.

## a) Thư viện sklearn

Thư viện Scikit-learn (viết tắt là sklearn) là một thư viện mã nguồn mở trong ngành Machine Learning, được sử dụng rộng rãi trong cộng đồng Python, được thiết kế trên nền NumPy và SciPy. Sklearn chứa hầu hết các thuật toán Machine learning hiện đại nhất, vì tính chất các thuật toán đã được viết sẵn, chúng ta chỉ cần đưa dữ liệu vào là nó có thể tự tính toán rồi đưa ra kết quả. Điểm mạnh của thư viện này là do nó được sử dụng nhiều trong giáo dục và công nghiệp cho nên nó luôn được cập nhật.

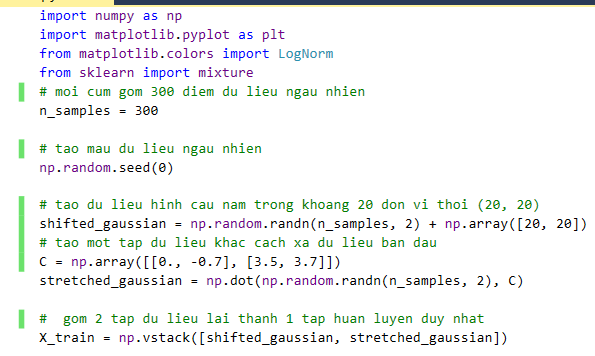
Để cài đặt thư viện ta có thể sử dụng dòng lệnh để cài đặt nó vào trong Python cùng với một số thư viện thông dụng như numpy, pandas,…



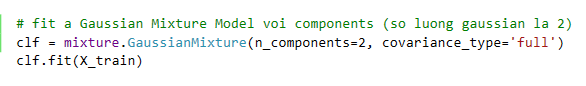
Lý do mà chúng ta nên dùng thư viện sklearn là vì nó cho phép ta sử dụng ngay các thuật toán một cách nhanh chóng và hiệu quả, sklearn là một trong những sự lựa chọn hàng đầu của các Dev. Đứng sau Scikit-learn là các viện nghiên cứu hàng đầu thế giới, gồm có Inria, Télécom Paristech, Paris-Saclay (Pháp), NYU Moore-Sloan Data Science Environment và Columbia University.

## b) Code ví dụ

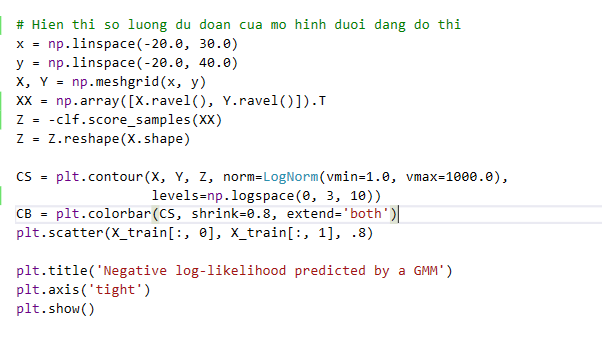
Ta import các thư viện cần thiết và tạo ra các điểm dữ liệu ngẫu nhiên trước



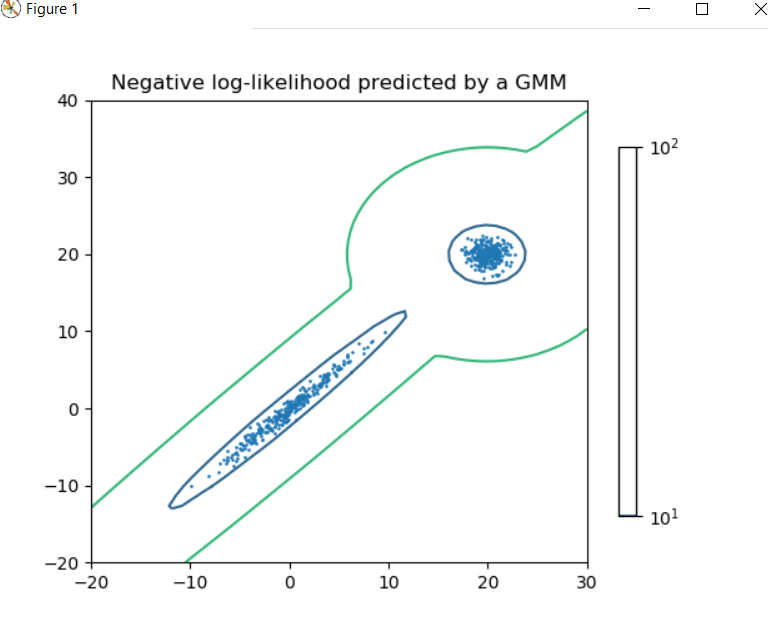
Sau đó ta fit dữ liệu với model GMM, chọn số lượng single Gaussian là 2



Biểu diễn thành dạng đồ thị



Kết quả ta nhận được là:



**DANH MỤC CÁC TRANG WEB THAM KHẢO**

[https://mas-dse.github.io/DSE210/Additional%20Materials/gmm.pdf](https://mas-dse.github.io/DSE210/Additional%20Materials/gmm.pdf?fbclid=IwAR1xWubTd7oHMfq7BOWigBKsqF1jxsC1gyXcve4-EMVUNxsQ_QcgwVsRuvw)

[https://www.youtube.com/watch?v=qMTuMa86NzU&t=3s](https://www.youtube.com/watch?v=qMTuMa86NzU&t=3s&fbclid=IwAR2um6Ol3cbW_6WbS6wW0ZKS2xLgqITe8cIjpE4nHtBrW6WOKUbZq0XlRUk)

[https://www.kaggle.com/charel/learn-by-example-expectation-maximization](https://www.kaggle.com/charel/learn-by-example-expectation-maximization?fbclid=IwAR1bxoOoGi7vwx4RreuqIK-EMzsveqtbzkpu-7D7Oaltxxvc5NwirRy33IQ)

<https://www.coursera.org/lecture/robotics-learning/1-4-2-gmm-parameter-estimation-via-em-79fAL>

<https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_pdf.html>

<https://ongxuanhong.wordpress.com/2015/09/21/expectation-maximization-em-hoi-gi-dap-nay/>