***Gaussian Mixture Models (GMMs) Clustering***

**Laporan Tugas**

**Kelas Penambangan Data**

**Taufik Fathurahman**

**NIM: 1301160790**

**Refka Muhammad Furqan**

**NIM: 1301162748**

**Haddad Alwi Yafie**

**NIM: 1301162756**



**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung 2019**

# Latar Belakang *Gaussian Mixture Models*

*K-means clustering* model merupakan model yang sangat sederhana dan relatif mudah untuk dipahami, namun dengan kesederhanaan dan kemudahannya itu memiliki kekurangan ketika diterapkan terhadap situasi dunia nyata karena sifat dari *k-means* yang non-probabilistik. Dari kekurangan yang muncul di *k-means* menjadi latar belakang munculnya metode *gaussian mixture models (GMMs).*

Untuk membuktikan kelemahan dari *k-means*, kami membuat simulasi terhadap *sample data* sebanyak 400 sampel. Akan ada dua jenis data yang akan diujikan terhadap *k-means*, yaitu data yang sederhana, dan data yang lebih kompleks untuk dikelompokkan.

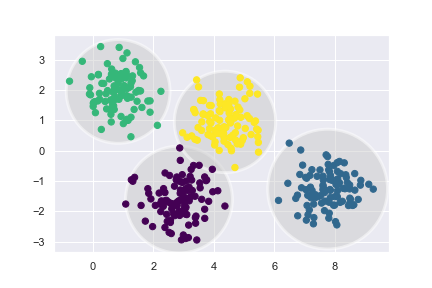


Figure 1 k-means untuk data sederhana

Dari Figure 1, didapatkan hasil yang baik dari *k-means* ketika melakukan *clustering* terhadap data yang sederhana. Lalu bagaimana bila data yang harus dikelompokkan lebih kompleks, seperti persebaran data yang tidak berbentuk lingkaran seperti diatas? Mari kita perhatikan Figure 2 dibawah ini.

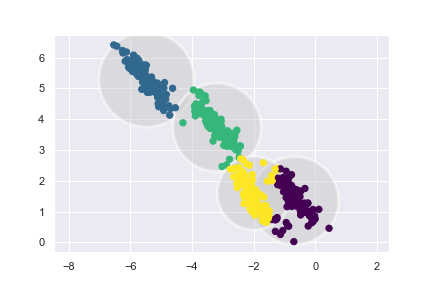


Figure 2 Kmeans untuk data yang kompleks

Dari Figure 2 didapatkan hasil bahwa *k-means* ini lebih tidak fleksibel terhadap sebaran data yang kecenderungannya tidak berbentuk lingkaran. Dari kelemahan ini lah yang menjadi motivasi untuk munculnya *gaussian mixture models (GMMs)*.

# Distribusi Gaussian

Formula untuk distribusi gaussian disebut juga dengan *probability density function (PDF).* *PDF* merupakan fungsi dari variable acak kontinu, yang integral dalam suatu interval memberikan probabilitas bahwa nilai variable berada dalam interval yang sama. Distribusi gaus dan formula *PDF* adalah digambarkan pada Figure 3 dan 4 berikut ini.

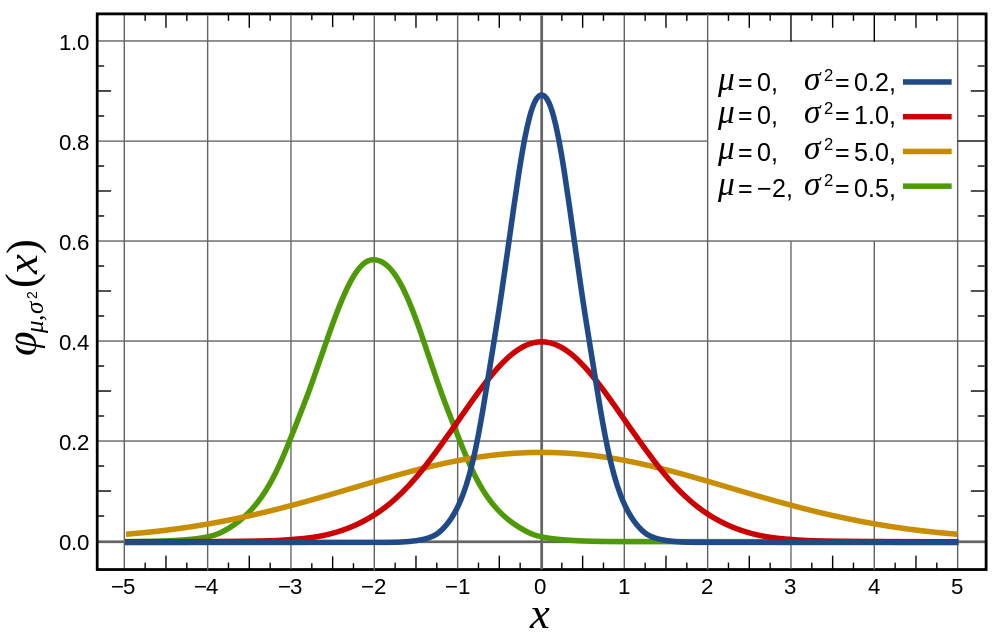


Figure 3 Distribusi gaussian

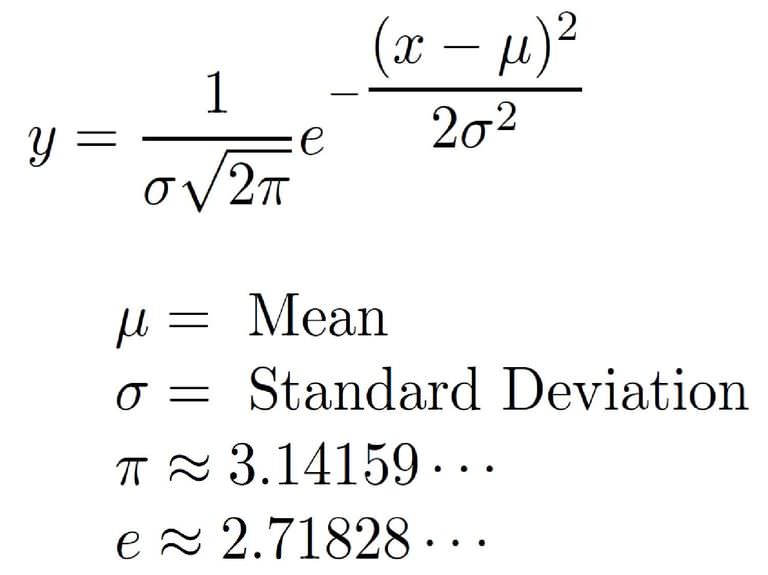


Figure 4 Probability density function

# Kovarian

Kovarian adalah ukuran bagaimana perubahan dalam satu variabel dikaitkan dengan perubahan dalam variabel kedua. Secara khusus, kovarians mengukur sejauh mana dua variabel terkait secara linear. Namun, ini juga sering digunakan secara informal sebagai ukuran umum tentang bagaimana dua variabel terkait secara monoton.

Terdapat tiga jenis kovarian yang dapat digunakan dalam *GMMs,* yaitu *diag*, *spherical,* dan *full.* *Diag* berarti bahwa ukuran cluster di sepanjang masing-masing dimensi dapat diatur secara independen, dengan elips yang dihasilkan dibatasi untuk sejajar dengan sumbu. *Spherical* merupakan model yang sedikit lebih cepat, yaitu dengan membatasi bentuk *cluster* sehingga semua dimensi sama. Dan yang terakhir adalah *full,* merupakan model yang memungkinkan setiap *cluster* dimodelkan sebagai elips dengan orientasi sebebas bebasnya. Gambaran dari tiga jenis kovarian adalah sebagaimana pada Figur 5 berikut ini.

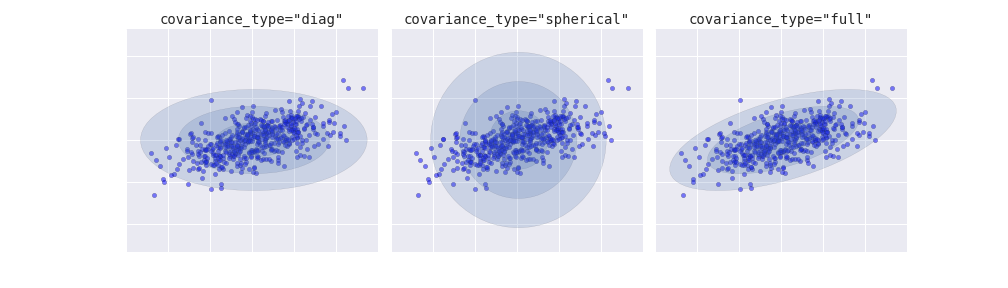


Figure 5 Covariance type

# Langkah-Langkah *Gaussian Mixture Models*

Metode *GMMs* berupaya menemukan campuran distribusi probabilitas gaussian multi-dimensi yang paling baik memodelkan setiap *dataset* *input*. Dalam melakukan *clustering*, metode ini memiliki kemiripan dengan *k-means* yaitu dengan menggunakan pendekatan *expectation-maximation*, yang secara sederhana adalah seperti berikut ini:

1. Pilih tebakan awal untuk lokasi dan bentuk.
2. Melakukan langkah berikut ini sampai konvergen:
   1. *Expectation-step/E-step*: Untuk setiap data points yang ada, temukan bobot probabilitas untuk setiap *cluster* yang ada.
   2. *Maximation-step/M-step*: untuk setiap data points yang ada, perbaharui lokasi, nomalisasi, dan dengan menggunakan bobot ubah bentuk dari setiap *cluster* yang ada.

Jangan dilupakan untuk menentukan jenis kovarian untuk yang akan digunakan dalam *GMMs*. Hal ini sangat penting untuk menentukan bagaimana bentuk dari *cluster* yang akan dibentuk selama iteriasi dilakukan untuk mendapatkan hasil yang paling optimal.

Untuk melihat bagaimana *GMMs* bekerja, kami melakukan percobaan terhadap data sample sebanyak 400. Sama halnya dengan percobaan yang dilakukan diatas pada *k-means*, sekarang juga akan dulakukan percobaan terhdap dua jenis data, yaitu data yang sederhana, dan kompleks.

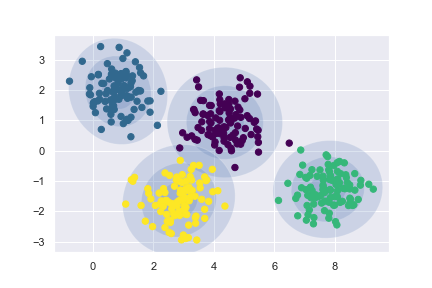
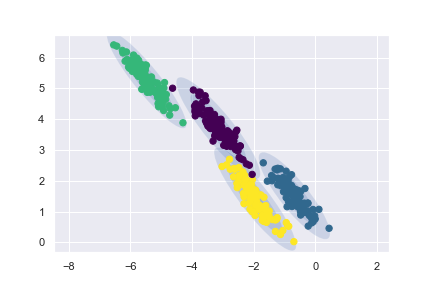
Berdasarkan Figure 6 dan 7 untuk kedua jenis data yang dilakukan proses *clustering*, didapatkan bahwa metode *GMMs* dapat bekerja dengan sangat baik untuk melakukan *clustering*. *GMMs* dapat dengan baik menyesuaikan dengan bentuk seberan data yang cenderung tidak berbentuk lingkaran.

Figure 7 GMMs Pada Data Kompleks

Figure 6 GMMs Pada Data Sederhana

Karena *GMMs* bekerja berdasarkan probabilitas, maka dapat ditampilkan bagaimana probabilitas dari setiap data poit yang ada terhadap setiap *cluster* yang ada. Contoh dari table probabilitas yang dibentuk jika mengambil sample lima data terhadap empat *cluster*, adalah sebagaimana pada Tabel 1 berikut.

Table 1 Probabilitas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Point | Cluster 1 | Cluster 2 | Clsuter 3 | Cluster 4 |
| 1 | 0. | 0.469 | 0. | 0.531 |
| 2 | 0. | 0. | 1. | 0. |
| 3 | 0. | 0. | 1. | 0. |
| 4 | 0. | 0. | 0. | 1. |
| 5 | 0. | 0. | 1. | 0. |

Jika disimpulkan dan digambarkan dengan bagaimana *gaussian* dapat melakukan *clustering* dengan baik, yaitu dapat mengikuti sebaran data yang bentuknya lebih kompleks adalah sebagaimana pada Figure 8 dan 9 berikut ini.

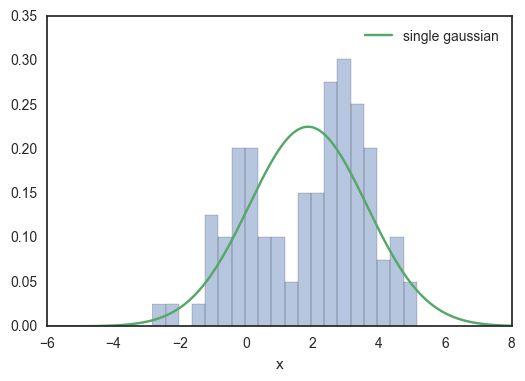
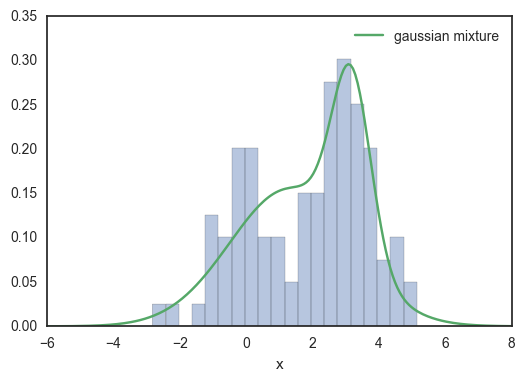


Figure 9 gaussian bentuk akhir

Figure 8 gaussian bentuk awal

Dari Figure 8 dan 9, menggambarkan bagaimana suatu *single gaussian* bekerja. Pada intinya, dengan menggunakan *GMMs*, akan melakukan perbaikan pada bentuk untuk menemukan bentuk yang paling optimal dalam membentuk *cluster.* Apabila *cluster* yang dibentuk ada beberapa, maka jumlah *gaussian* akan menyesuaikan.

**Appendix:** Setiap code untuk membuat figure dan table pada laporan ini dapat diakses secara online pada link:

https://github.com/taufikfathurahman/Clustering/tree/master/Gaussian%20Mixture%20Models