Algoritmo de agrupamento Expectation-Maximization

Seminário de Fundamento de Sistemas Inteligentes

Antônio Carlos Junior 16/0112745
Mayara C. Marinho 18/0025210

Visão geral

- **Expectation Maximization (EM)**: Método iterativo para estimar parâmetros em modelos estatísticos. Foi pensado para ser utilizado em dados multidimensionais.
- **Gaussian Mixture Model (GMM)**: Modelo que assume que todos os pontos são gerados como uma mistura de um número finito de distribuições Gaussianas.
- Utilizado na prática como algoritmo de aprendizado não supervisionado de clusterização como o K-Means.
- Hard-Cluster (K-Means) x Soft Cluster (EM)

GMM

1 Dimensão

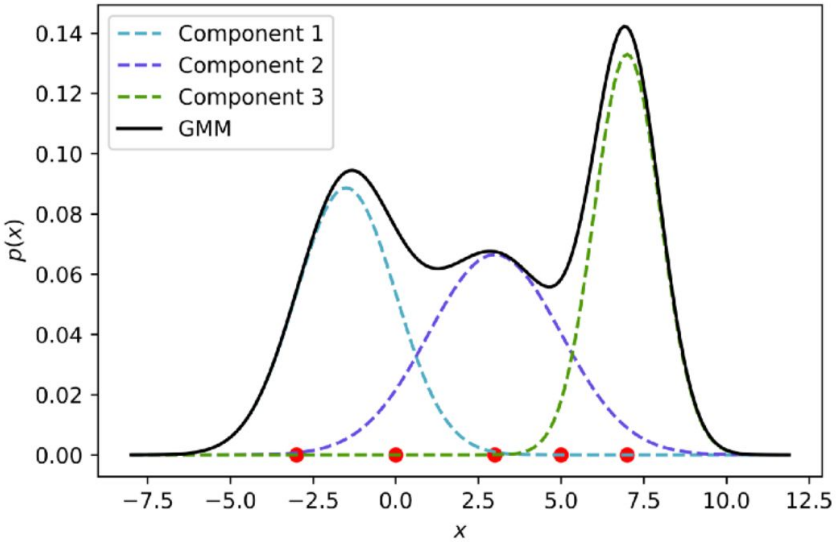
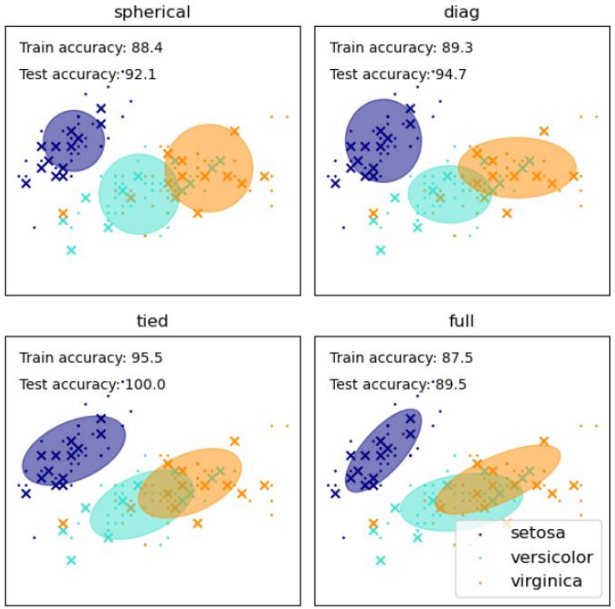


Gráfico de um MMG com 3 componentes

Fonte: Medium.

Múltiplas Dimensões



Fonte: Scikit Learn.

Problema

- GMM por si só não tem os parâmetros necessários para definir a distribuição de gaussianas que melhor definem os dados.
- EM surge como uma solução para este problema, por meio da definição de parâmetros co-dependentes das gaussianas.
- Quais são os parâmetros necessários?
  - Pesos de cada gaussiana; ϕ_j.
  - Médias das gaussianas; μ_j
  - Covariância das gaussianas; Σ_j

$$p(x) = \sum_{j=1}^{k} \phi_j \mathcal{N}(x; \mu_j, \Sigma_j)$$

$$\sum_{j=1}^{k} \phi_j = 1$$

Metodologia EM

1. Inicializar aleatoriamente os parâmetros μ's, Σ's, e ϕ's;

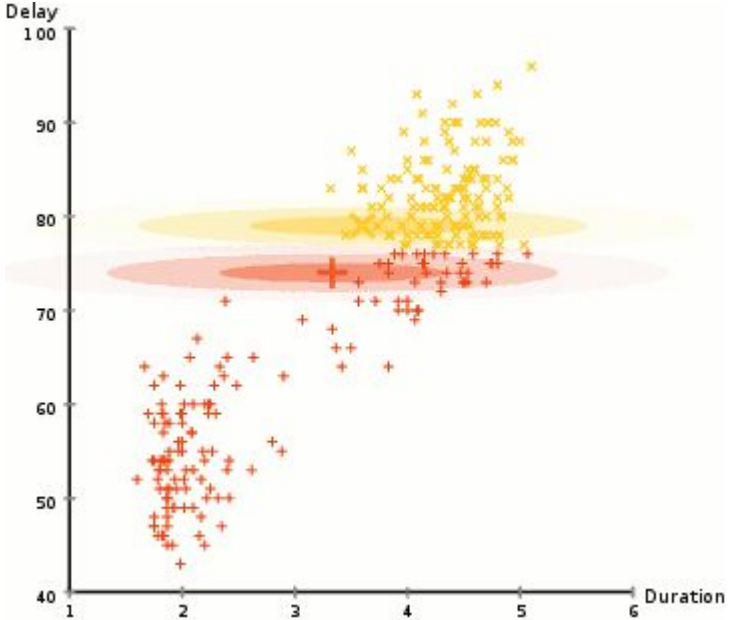
2. Passo E:

   Computar a probabilidade de cada ponto x_i ter sido gerado por uma gaussiana k, usando os parâmetros μ, Σ, e ϕ.

3. Passo M:

   Atualizar os parâmetros μ, Σ, and ϕ, utilizando as probabilidades obtidas no passo E.

4. Repetir os passos 2 e 3 até que não haja mudança significativa na função da esperança do logaritmo da verossimilhança.



Expectation Maximization for Old Faithful Eruption Data (Wikipedia)

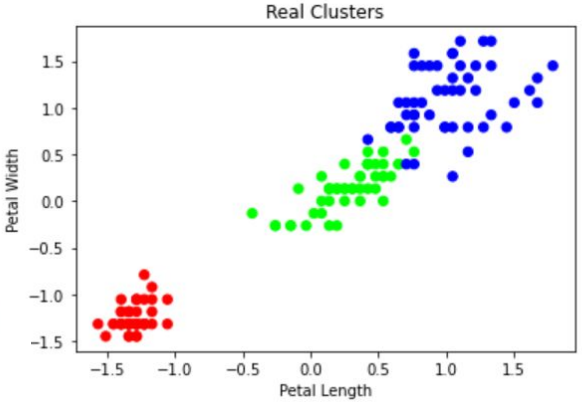
Vantagens

1) Fornece informações de dispersão dos clusters;
2) Processo “generativo”: descreve como os dados foram gerados, por meio de uma distribuição;
3) Pode ser utilizado para preencher dados ausentes em uma amostra;
4) Pode ser utilizado para definir variáveis latentes (definidas por um modelo matemático).

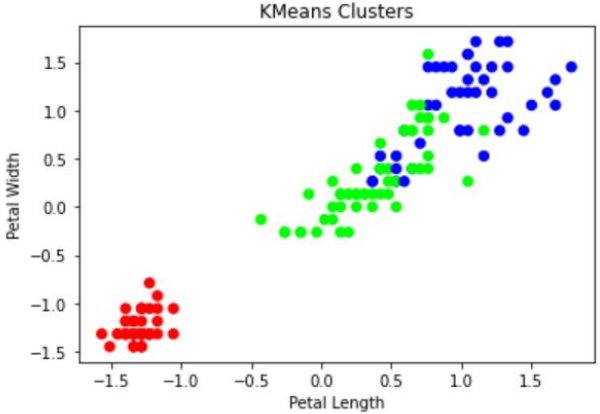
Desvantagens

1) O Algoritmo EM é bastante sensível à inicialização dos parâmetros;
2) A complexidade computacional do EM é maior que a do K-Means;
3) O EM demora para convergir em relação ao K-Means;
4) Requer uma configuração manual do número de clusters.

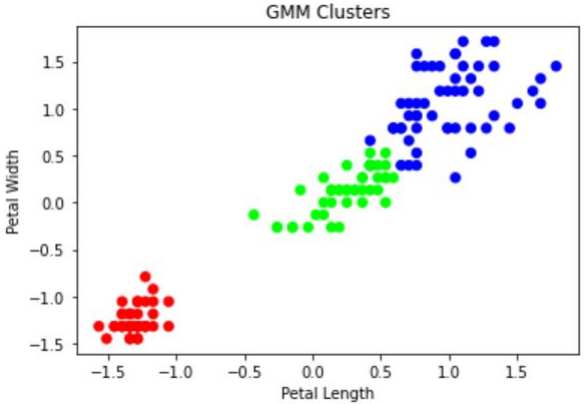
Visualizações



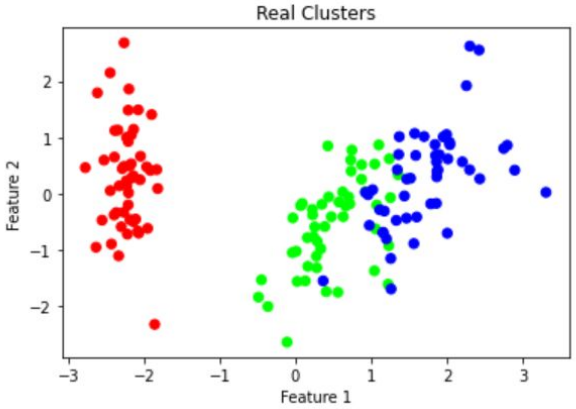

```
Average silhouette_score =  0.45994823920518635
Accuracy =  0.8333333333333334
```



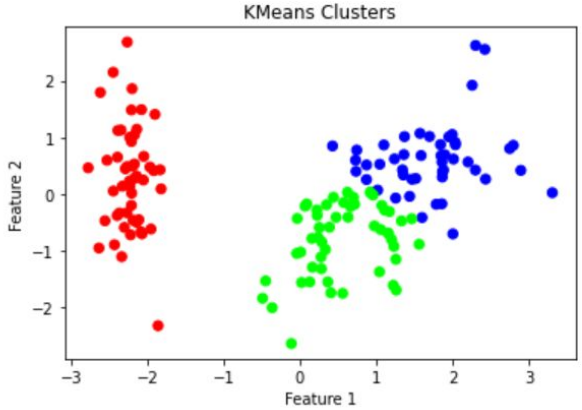

```
Average silhouette_score =  0.37416491866541235
Accuracy using GMM =  0.9666666666666667
```



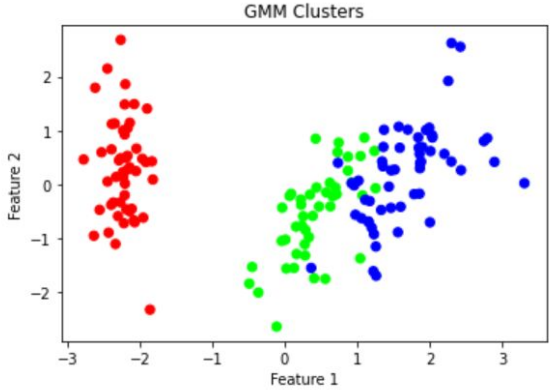
Visualizações PCA




```
Average silhouette_score =  0.45994823920518635
Accuracy =  0.8333333333333334
```




```
Average silhouette_score =  0.37416491866541235
Accuracy using GMM =  0.9666666666666667
```



Principais referências

- https://medium.com/b2w-engineering/o-racioc%C3%ADonio-por-tr%C3%A1s-do-algoritmo-expectation-maximization-91d4a8588778
- https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
- https://pythonmachinelearning.pro/clustering-with-gaussian-mixture-models/
- https://www.cs.toronto.edu/~hinton/absps/emk.pdf
- https://pt.wikipedia.org/wiki/Algoritmo_de_maximiza%C3%A7%C3%A3o_de_expectativa