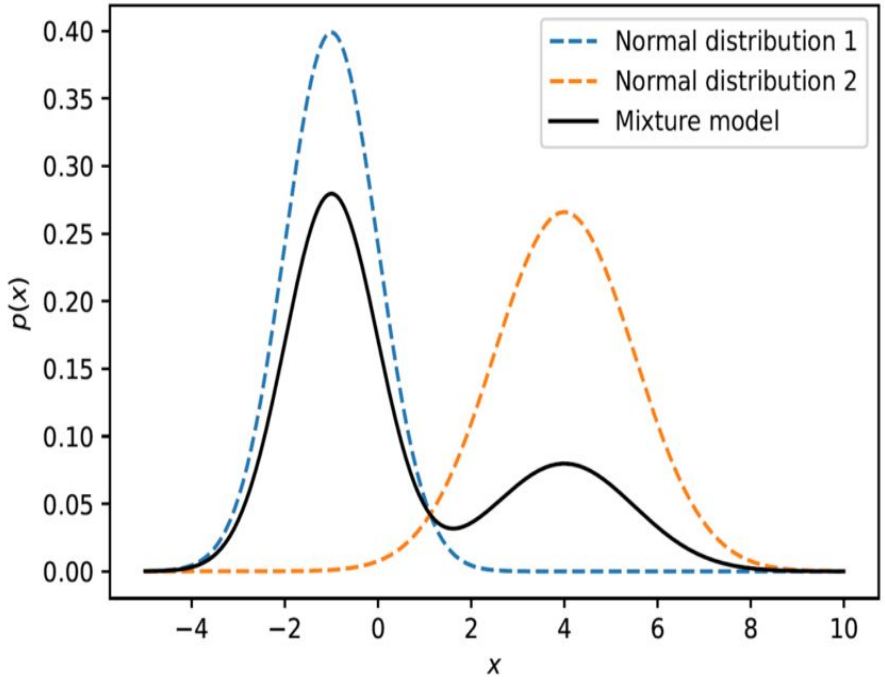
Gaussian Mixture Model

Apresentado por:
Bruno Leonardo, Clebson Santos, Victor guedes, e Wallan Melo

1 - Explicação do Funcionamento do Algoritmo

O Modelo de Misturas Gaussianas, ou GMMs, é uma técnica de aprendizado não supervisionado, que representa os dados como uma mistura de várias distribuições gaussianas, e é vista como uma generalização do k-means. Cada cluster é uma Gaussiana com média, variância, e peso.



Diferença entre GMM e K-Means:

Modelo de Mistura Gaussiana	K-Means
Mais versátil, mas também mais complicado de treinar.	Não serve para muitos propósitos, mas é simples o suficiente para treinar.
Alto requisito de tempo de execução	Treinar mais rápido e com menor tempo de corrida.
Parte-se do pressuposto de que cada ponto de dados se origina de uma combinação de distribuições gaussianas.	Não faz suposições. Simplesmente divide os dados em clusters.
Leva em consideração a variância	Não aborda a variância de forma alguma.
Mais eficazes, pois conseguem lidar com valores ausentes.	Não é possível lidar com dados faltantes, portanto, serão necessários recursos para limpar ou complementar os dados.
O formato dos aglomerados é flexível e pode ser alterado.	Limitado a aglomerados esféricos
Mais preciso para conjuntos de dados pequenos e clusters que não são distintos.	Mais preciso quando o conjunto de dados é grande e possui agrupamentos distintos.

Funcionamento

No GMM cada componente é uma distribuição Gaussiana, um modelo de mistura gaussiana é um modelo de mistura comum, onde a densidade de probabilidade é dada por uma mistura de distribuições gaussianas:

Onde:

p (x) é a função de densidade ou massa geral do modelo de mistura..

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$$

Um modelo de mistura gaussiana

K é o número de distribuições de componentes na mistura.

w□ é o peso de mistura do k -ésimo componente, com 0 ≤ w□ ≤ 1 e a so) a probabilidade a priori do componente k .

x é o vetor de dimensão d.

μ□ é o centro do agrupamento, em uma distribuição unidimensional, este será um vetor com somente um único valor, mas em uma distribuição n-dimensional, será um vetor com n valores.

Σ□ esta é a dispersão/formato da própria Gaussiana. Em uma distribuição unidimensional, será um único valor, mas em uma distribuição n-dimensional, será uma matriz nxn.

N (x ; μ□ , Σ □) é a função de densidade normal multivariada para o k -ésimo componente

$$\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x} - \mu_k)\right)$$

GAUSSIANA MULTIVARIADA

1 - Nessa primeira parte: $\frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}}$

É uma constante de normalização, com a função de garantir que a área total abaixo da curva seja 1, e vai depender diretamente da dimensão d, pois quanto maior a dimensão maior o termo.

2 - Nessa segunda parte: $\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x} - \mu_k)\right)$

Essa é a parte mais importante do modelo de mistura gaussiana multivariada, isso porque ela determina como a densidade diminui conforme o ponto se afasta do centro do cluster.

$-\frac{1}{2}$ Esse multiplicador serve para ajustar a distância para uma escala apropriada para o modelo.

$(\mathbf{x} - \mu_k)$ É um vetor que aponta do centro do cluster até o ponto x, medindo o deslocamento do ponto.

$\sum_k^{-1}$ É a inversão da matriz de covariância, ela mede a dispersão ao contrário, se um determinado ponto está longe na direção de menor variância, então a sua penalização será maior.

GAUSSIANA UNIVARIADA

No caso de distribuição gaussianas univariadas, a densidade de probabilidade pode ser simplificada para a seguinte fórmula.

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x; \mu_k, \sigma_k)$$

Um modelo de mistura de distribuições gaussianas univariadas

Onde:

μ□ é a média do k -ésimo componente gaussiano.

σ□ é a variância do k -ésimo componente gaussiano.

N (x ; μ□ , σ□) é a função de densidade normal univariada para o k -ésimo componente.

$$\mathcal{N}(x;\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

1 - Nessa Primeira parte:

$\frac{1}{\sqrt{2\pi}\sigma_k}$ É o fator normalização, ele é o termo responsável por garantir que a área sob a curva seja sempre 1.

2 - Nessa Segunda Parte:

$\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$ Esse termo é o responsável pelo formato de sino que o modelo gaussiana tem, gerando uma curva simétrica com o seu ponto máximo sendo a média.

$(x-\mu_k)^2$ Mede o quão longe o ponto x está do centro, mede o deslocamento do ponto x.

$2\sigma_k^2$ Controla o peso da distância, por exemplo, se o cluster é muito variado, ou seja com o **σ** grande, os pontos distantes não sofrem uma penalização tão forte. Já se for pouco variado, ou seja com o **σ** pequeno, até as pequenas distâncias irão diminuir a probabilidade.

Principal Diferença entre o Uni e Multivariado

O GMM univariado modela apenas uma variável por vez, enquanto o GMM multivariado modela várias variáveis simultaneamente, capturando correlações entre elas.

Aplicações do GMM

1. Medicina:
É utilizado para auxiliar na identificação de padrões em conjuntos de dados médicos, e eles podem ser usados para detectar doenças, identificar grupos de pessoas com queixas comuns e até mesmo fazer previsões.

2. Análise do comportamento do consumidor:
Os GMMs podem ser usados em publicidade para realizar as análises de comportamento do usuário e com isso gerar previsões com base em dados anteriores sobre possíveis transações potenciais.

Aplicações do GMM

3. Previsão do preço das ações:
A previsão do preço das ações é mais uma aplicação dos modelos de mistura gaussiana, que podem ser utilizados em séries temporais de preços de ações na economia. Os modelos de mistura gaussiana podem ser empregados para identificar pontos de inflexão em análises de séries temporais e auxiliar na descoberta de momentos cruciais nos preços das ações ou outros movimentos de mercado.

4. Recursos audiovisuais:
Recentemente, os GMMs têm se mostrado eficazes na extração de características de dados de áudio para uso em sistemas de reconhecimento de fala. Eles também são importantes no rastreamento de múltiplos objetos, onde a quantidade de componentes na mistura e os valores de suas médias preveem a localização de um objeto em cada quadro de um vídeo, permitindo o rastreamento de objetos.

Implementação do GMM para o Dataset Íris

Conclusão final

- O **K-Means e o GMM apresentaram alta concordância** na clusterização do conjunto Iris.

- As **principais divergências ocorreram entre versicolor e virginica**, na região de sobreposição dos dados.

- O **K-Means realiza separação rígida**, enquanto o **GMM modela a incerteza de forma probabilística**.

- O **ponto de teste foi classificado pelo GMM com 99,8% de probabilidade** no cluster da *virginica*.

- O **GMM mostrou maior poder interpretativo**, especialmente em regiões com sobreposição entre clusters.

OBRIGADO!