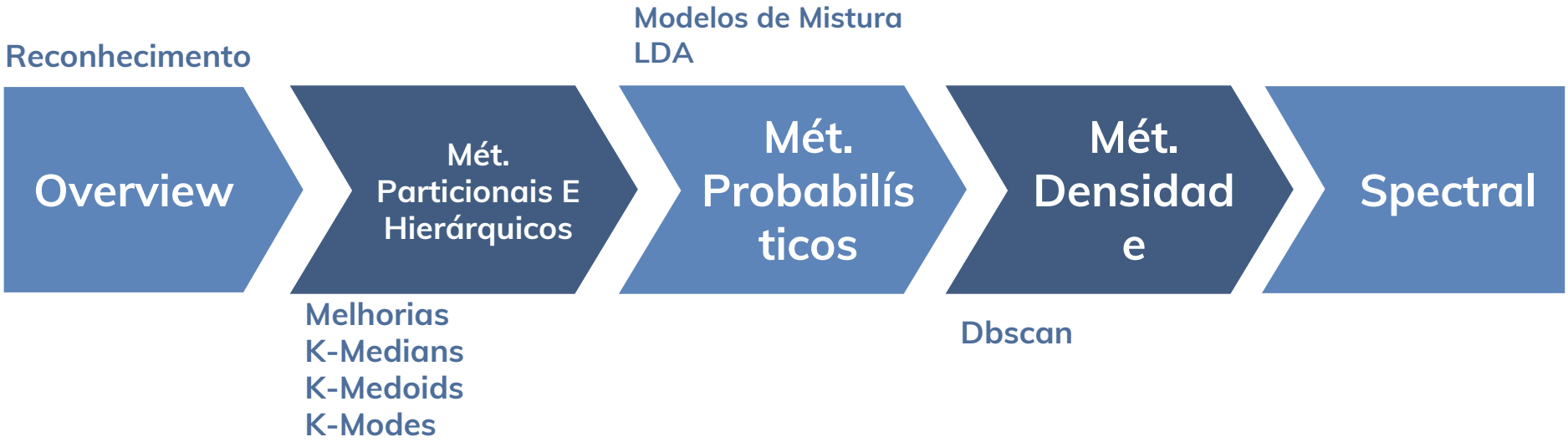
Cluster

Larissa Sayuri Futino Castro dos Santos

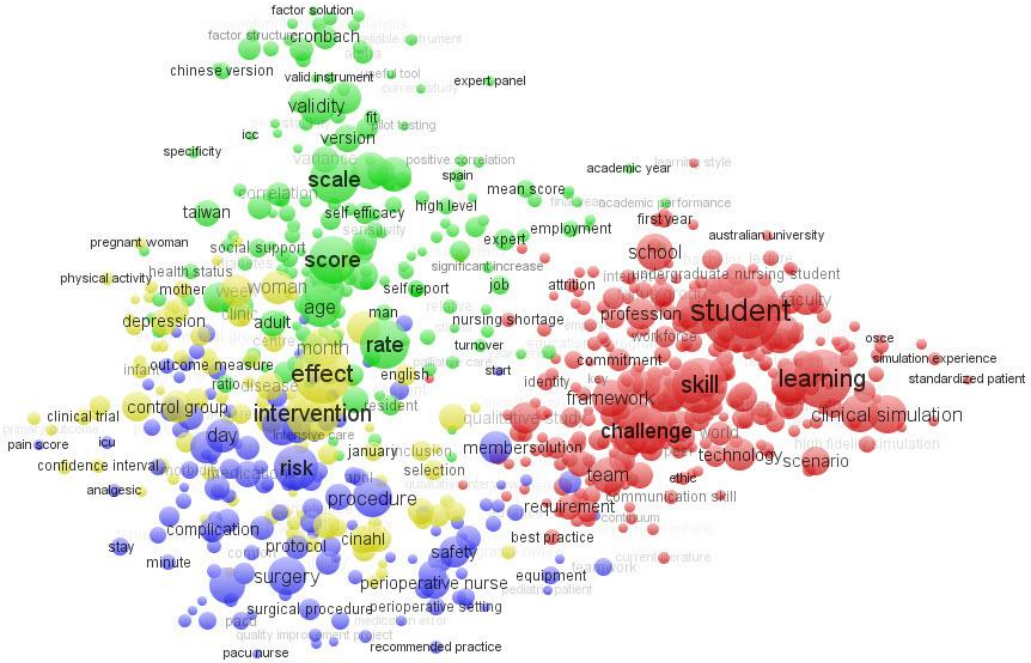
Agenda

O que faremos



Cluster

Motivação nunca é demais



A análise de cluster divide dados em grupos (clusters) que tenham significado ou sejam úteis (ou ambos).

Cluster

Motivação nunca é demais

Passo intermediário para demais algoritmos

Segmentação de Clientes

Noção de padrões de seq genéticas

Determ. Padrões de mudança

Resumo de Dados

Filtros Colaborativos

Determ. músicas similares

Determ. de comunidades em uma rede Social

Mét. Particionais e Hierárquicos

Simplicidade e Facilidade de Implementação

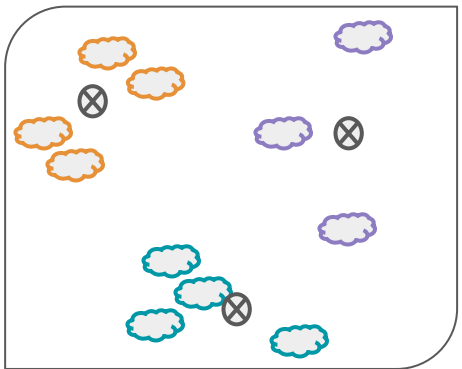
Particional		Hierárquico
Otimiza uma função objetivo específica e iterativamente melhora a qualidade das partições.	**Como funciona**	Desenvolve uma estrutura de árvore binária chamada dendograma.
Problema de otimização minimizando SSE (Soma de Quadrados do Erro).	**Definição**	
O método necessita do número de *clusters* da base, previamente.	**Definição do Número de Clusters**	A escolha do número de clusters dá-se particionando a árvore em diferentes níveis para obter soluções distintas para o mesmo banco sem a necessidade da repetição do emprego do algoritmo.
Escolha dos centróides iniciais Parâmetro do número de *clusters* predefinido Comportamento não determinístico	**Aspectos dificultadores**	Sensibilidade a ruídos e *outliers* Exige a especificação de vários parâmetros, incluindo uma métrica de distância
Formato esférico	***Shape***	

Mét. Particionais

Métodos para inicializar os centróides

- **MacQueen** — 1967
- **Hartigan e Wang** — 1979
- **Milligan** — 1981
- **Bradley e Fayyad** — 1998
- **Kmeans++** — 2007

K centróides iniciais amostrados aleatoriamente

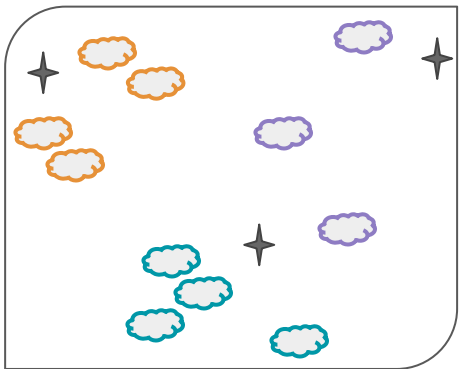


Mét. Particionais

Métodos para inicializar os centróides

- **MacQueen** — 1967
- **Hartigan e Wang** — 1979
- **Milligan** — 1981
- **Bradley e Fayyad** — 1998
- **Kmeans++** — 2007

K centróides iniciais amostrados aleatoriamente



Mét. Particionais

Métodos para inicializar os centróides

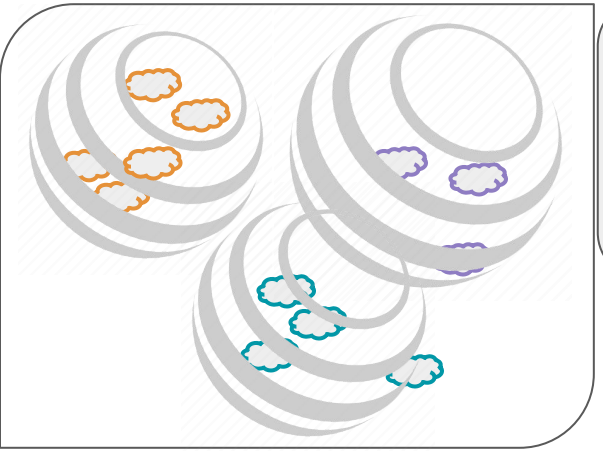
MacQueen — 1967

Hartigan e Wang — 1979

Milligan — 1981

Bradley e Fayyad — 1998

Kmeans++ — 2007



Nearest neighbor density

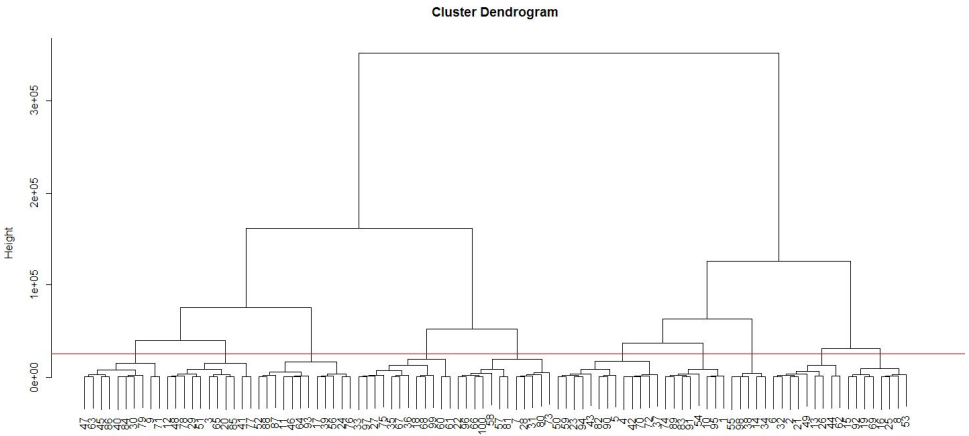
Pontos bem separados e que tem muitos pontos na sua redondeza multidimensional são bons candidatos

Mét. Particionais

Métodos para inicializar os centróides

- **MacQueen** — 1967
- **Hartigan e Wang** — 1979
- **Milligan** — 1981
- **Bradley e Fayyad** — 1998
- **Kmeans++** — 2007

Tenta usar resultados de *cluster* hierárquico aglomerativo



Mét. Particionais

Métodos para inicializar os centróides

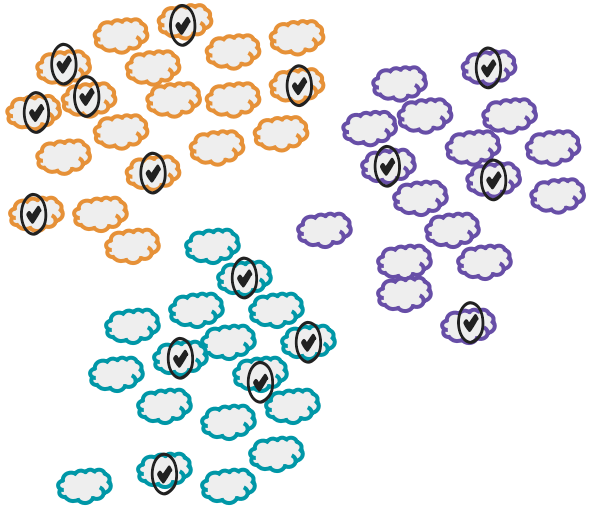
MacQueen — 1967

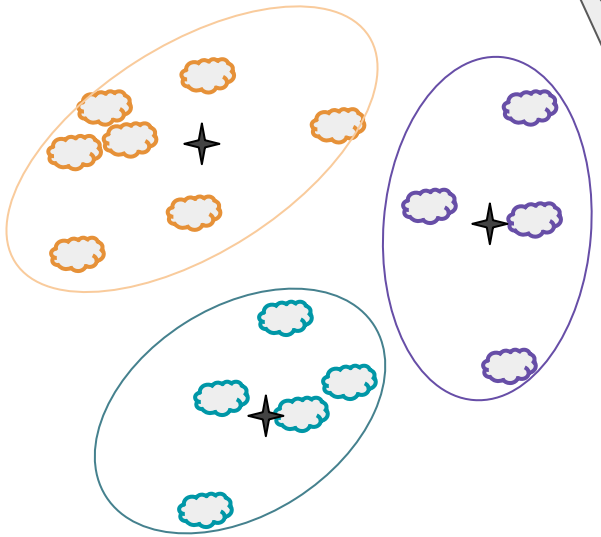
Hartigan e Wang — 1979

Milligan — 1981

Bradley e Fayyad — 1998

Kmeans++ — 2007

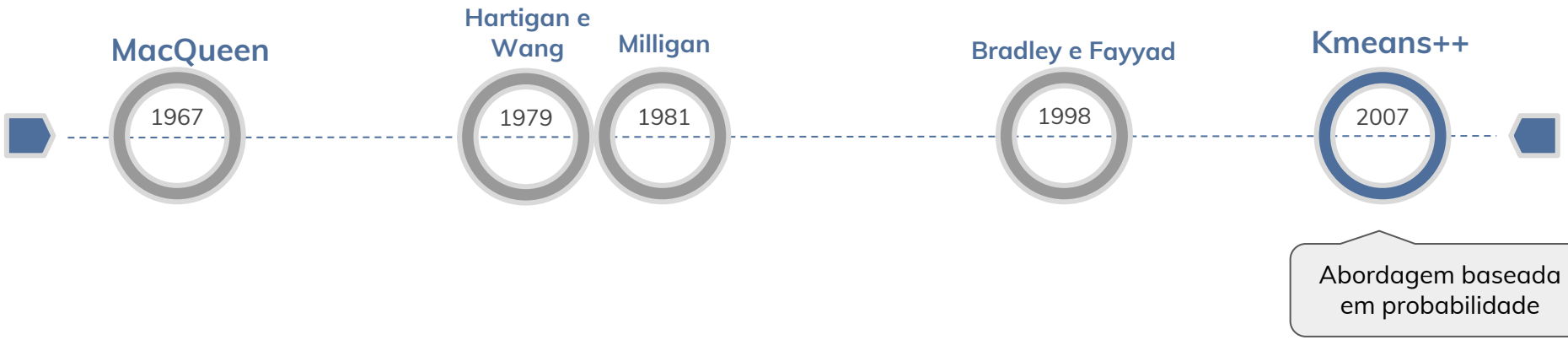




Toma amostras dos dados e faz os *clusters* com elas.
Usa os centróides dos *clusters* obtidos como centróides iniciais no *cluster* com todas as observações.

Mét. Particionais

Métodos para inicializar os centróides



Mét. Particionais e Hierárquicos

Determinando número de Clusters

IsoData 1965

Duda e Hart 1973

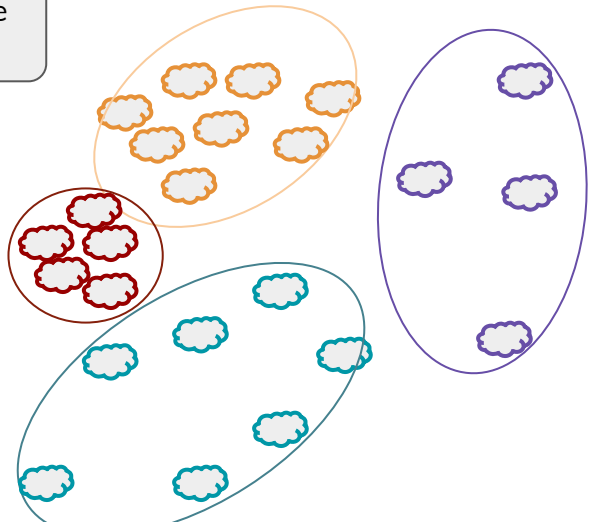
Calinski-Harabasz 1974

BIC 1977

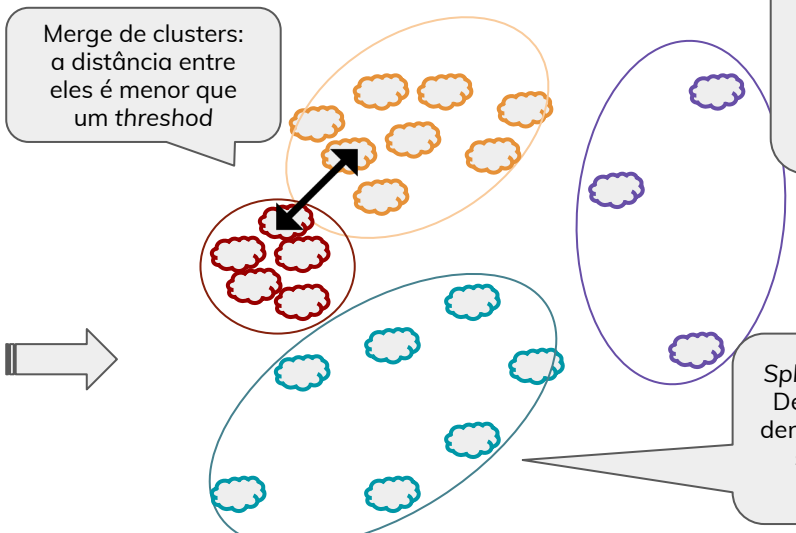
Bradley e Fayyad 1990

Gap statistic AIC 2001

Método do centróide mais próximo



Merge de clusters: a distância entre eles é menor que um *threshod*



Merge de clusters: Cluster com menos que um dado número de pontos

Split de clusters: Desvio padrão dentro do *cluster* supera um *threshod*

Mét. Particionais e Hierárquicos

Determinando número de Clusters

IsoData 1965

Duda e Hart 1973

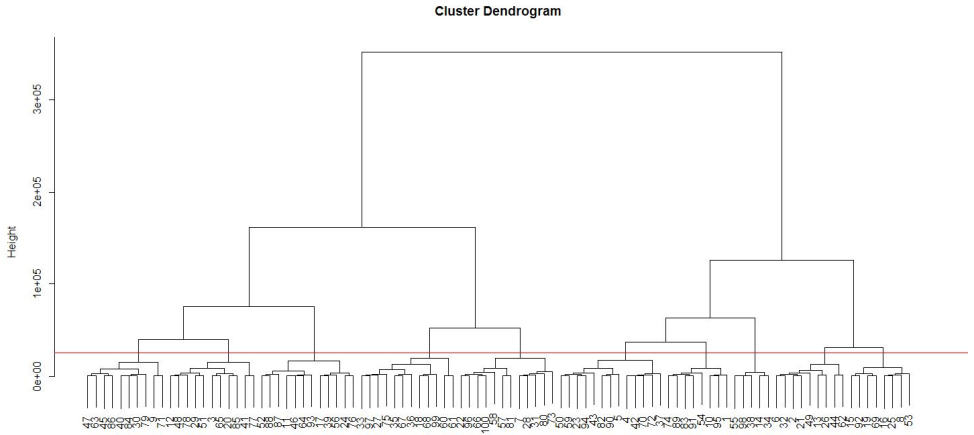
Calinski-Harabasz 1974

BIC 1977

Bradley e Fayyad 1990

Gap statistic AIC 2001

Como escolher o melhor ponto de corte do dendograma



Mét. Particionais e Hierárquicos

Determinando número de Clusters

IsoData — 1965

Duda e Hart — 1973

Calinski-Harabasz — 1974

BIC — 1977

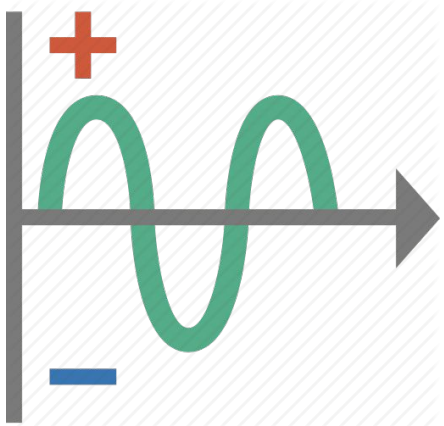
Bradley e Fayyad — 1990

Gap statistic AIC — 2001

K que maximiza uma função que relaciona Soma Quadrados dentro e entre *clusters*

K que minimiza BIC

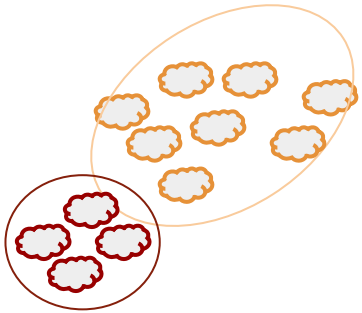
K que minimiza AIC

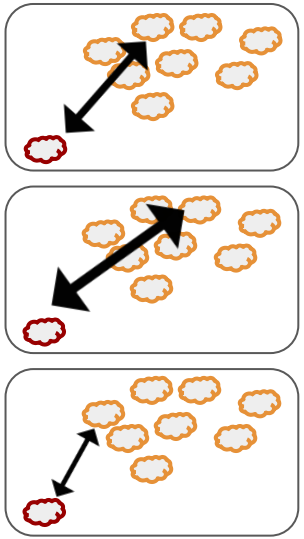


Mét. Particionais e Hierárquicos

Determinando número de Clusters

- **IsoData** — 1965
- **Duda e Hart** — 1973
- **Calinski-Harabasz** — 1974
- **BIC** — 1977
- **Bradley e Fayyad** — 1990
- **Gap statistic AIC** — 2001





Sillhouette coefficient

Considera distâncias intra e entre *clusters*.

Calcula sillhouette de todos os pontos

Sillhouette médio é uma medida da qualidade de um cluster

Variações do K-Means

Escolher ≠ centróides representativos

K-medoids - próprios pontos
Mais robusto a outliers
Alto custo computacional
PAM: minimizar trocas
CLARA: considera muitas amostras

K-medians - mediana do cluster
Mais robusto a outliers

K-modes
Permite aplicações com dados categóricos

Escolher melhor estimativas dos centróides

Intelligent K-means

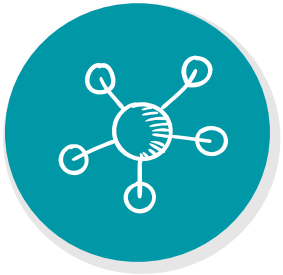
Genetic K-means

Aplicar técnica de transformação de atributos

Weighted K-means

Kernel k-means

Variações Hierárquicas



CURE

Representar um *cluster* a partir de pontos dispersos representativos

Formatos arbitrários
Robusto a *outliers*

Chameleon

Usa métodos para particionar grafos. As partições são sementes para a formação dos *clusters*

Formatos arbitrários

SOM

Usa Redes Neurais artificiais.
Se assemelha ao algoritmo Hierárquico alterando apenas a etapa de atualização.

Possibilita visualização

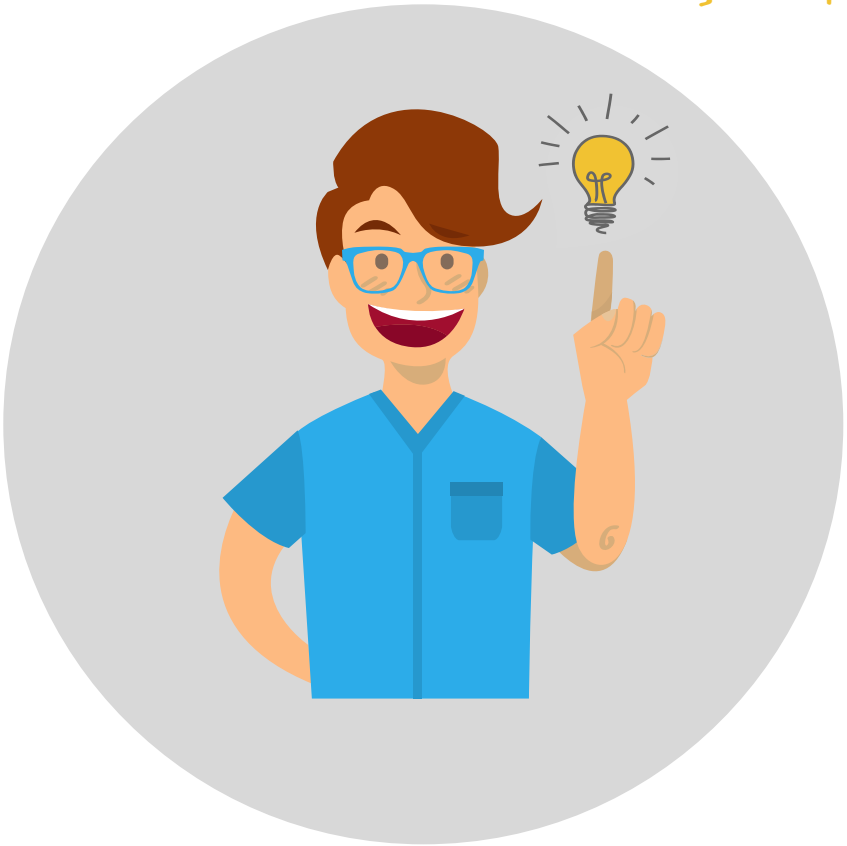
COBWEB

Usa Naive-Bayes e Árvore de Decisão


Análise de Cluster hoje
o que vai além dos métodos hierárquicos e particionais

Mét. Probabilísticos

Problema de estimação de parâmetros



Cada *cluster* é representado por uma distribuição de probabilidade paramétrica

Mét. Probabilísticos - GMM

Modelo de Mistura Normal

Mod. Normal

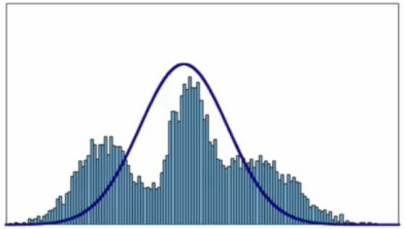
Conhecido/Familiar
Disponível em softwares

Vida real

Densidade de formato não usual

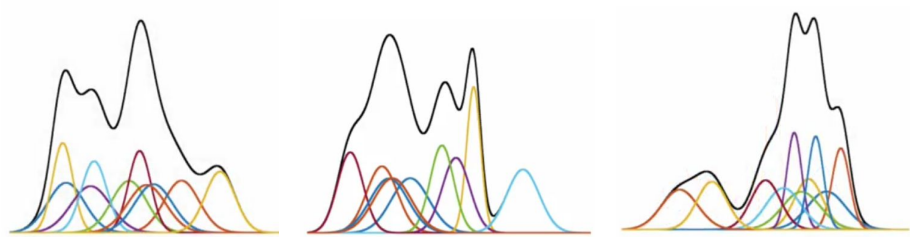
Solução

Modelo de Mistura



Normal mostra-se uma densidade limitada:

- Simétrica
- Unimodal



> “Se a gente escolhe os elementos Gaussianos adequados a gente pode expressar toda/qualquer distribuição não usual.”

Mét. Probabilísticos - GMM

Modelo de Mistura Normal

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k g_k(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

w_k: mixing coefficient (weight)

g_k: Gaussian with $\boldsymbol{\mu}_k$ and Σ_k

$$w_k > 0, \sum_{k=1}^{K} w_k = 1$$

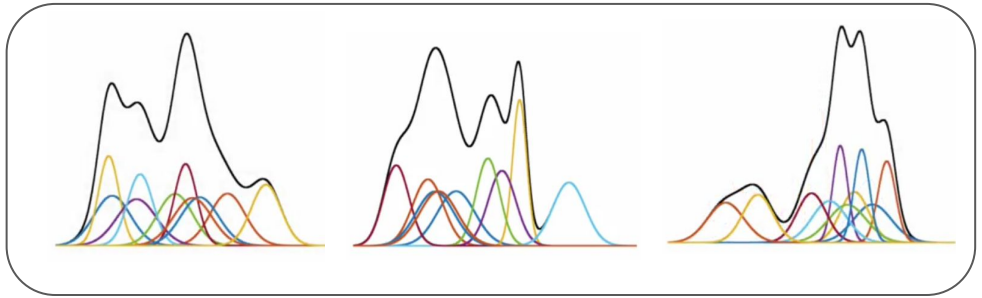
Garantir que o GMM é uma densidade de Probabilidade que integra 1

Mét. Probabilísticos - GMM

Modelo de Mistura Normal

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k g_k(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

> “Se é possível tomar k arbitrariamente grande e variâncias arbitrariamente pequenas pode-se expressar qualquer formato de distribuição, em teoria.”



Mét. Probabilísticos - GMM

Modelo de Mistura Normal

Flexibilidade

Aumento do número de parâmetros a serem estimados

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$$

$$\Sigma = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$$

$$\boldsymbol{w} = \{w_1, w_2, \ldots, w_K\}$$

K: Number of Components

Não há solução analítica/fechada → Computação iterativa sem garantia de solução ótima global

Risco de superajustamento

Necessita-se especificar/decidir um valor para K

Mét. Probabilísticos - GMM

Modelo de Mistura Normal

EM
Expectation Maximization

Estimadores de Máx. Verossim.
("valores dos parâmetros mais prováveis de produzir os dados observados")

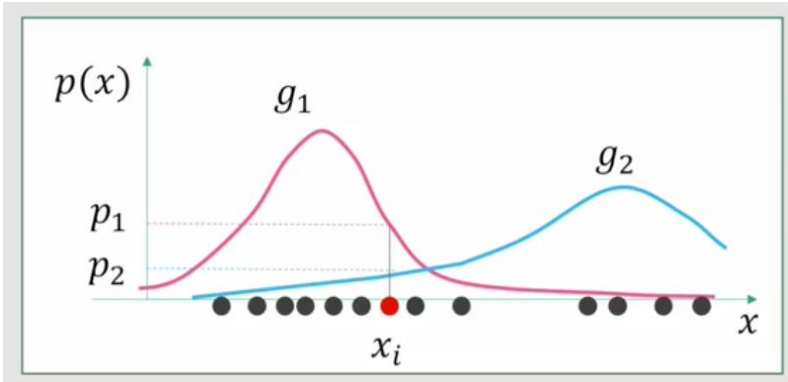
Estimação iterativa

Sem garantia de solução ótima global

V.a. latente (não observamos diretamente)
Representar a componente/*cluster* de onde a observação veio

$$z_k^i = \frac{g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{g_1(\mathbf{x}_i|\boldsymbol{\mu}_1, \Sigma_1) + g_2(\mathbf{x}_i|\boldsymbol{\mu}_2, \Sigma_2)}$$

Razão relativa da densidade da Gaussiana 01 no ponto x_i



$$z_1^i = \frac{p_1}{p_1 + p_2}$$

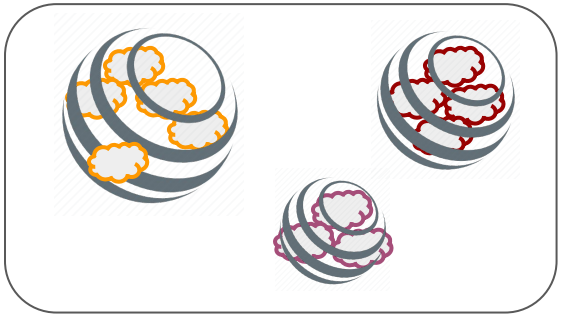
$$z_2^i = \frac{p_2}{p_1 + p_2}$$

Mét. Densidade

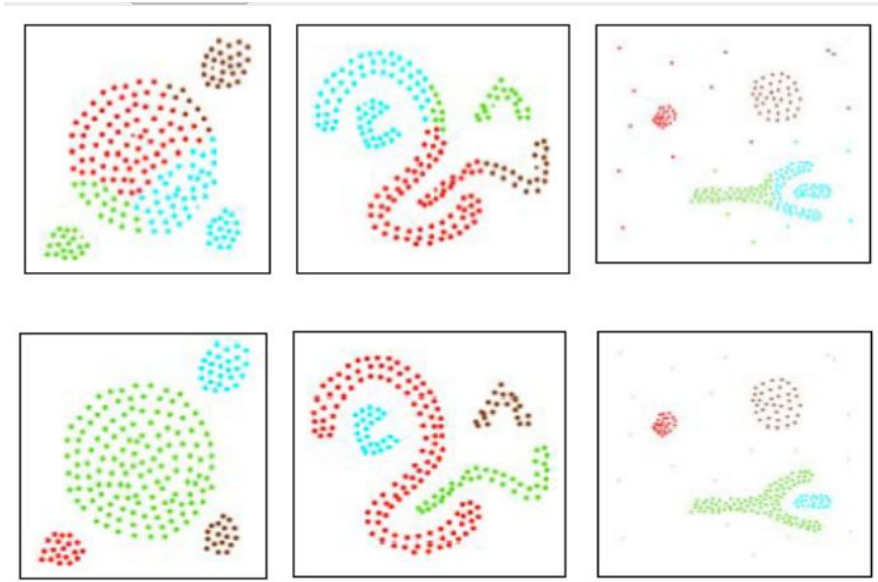
Clusters de formatos arbitrários

Motivação

Usualmente, os métodos para detecção de *clusters* captam os agrupamentos de *shape* esférico



Necessita-se de um método para detecção de *clusters* de formato arbitrário



CLARANS

DbScan

Mét. Densidade - DbScan

Clusters de formatos arbitrários

Método não-paramétrico para detecção de *clusters*

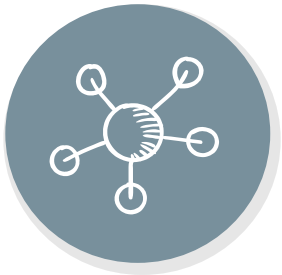
Não exige definição do número de *clusters*

Sem pressupostos sobre a distribuição dos *clusters*

Sensível a parâmetros de entrada

Podem existir observações sem *cluster*

Questões/aspectos do delineamento



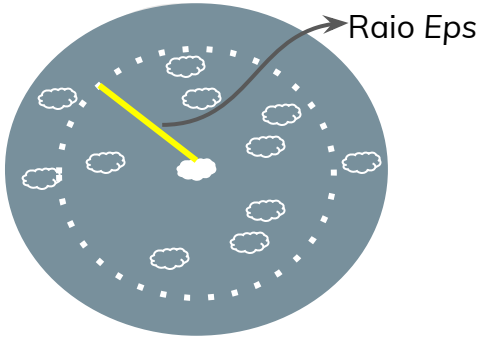
Como estimar a densidade?

Como definir conectividade?

Mét. Densidade - DbScan

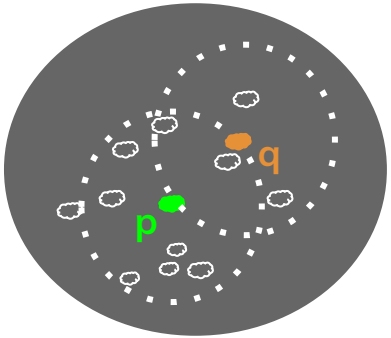
Clusters de formatos arbitrários

CORE point



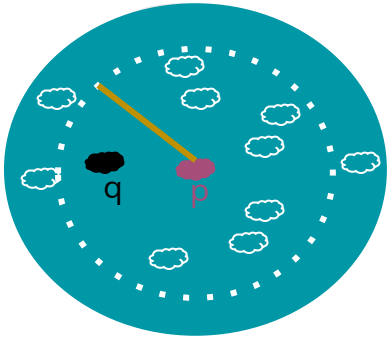
A vizinhança de raio *Eps* contém pelo menos *MinPts* (5, por exemplo)

Border point



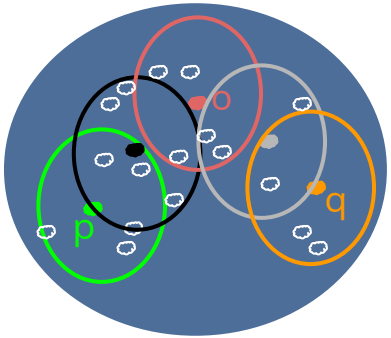
Ponto q está dentro da vizinhança de raio *Eps* de p e não é core point

q é diretamente acessível por densidade



Ponto q está dentro da vizinhança de raio *Eps* de p

p e q são conectados por densidade



Existe um ponto O tal que p e q são alcançáveis por densidade

Mét. Densidade - DbScan

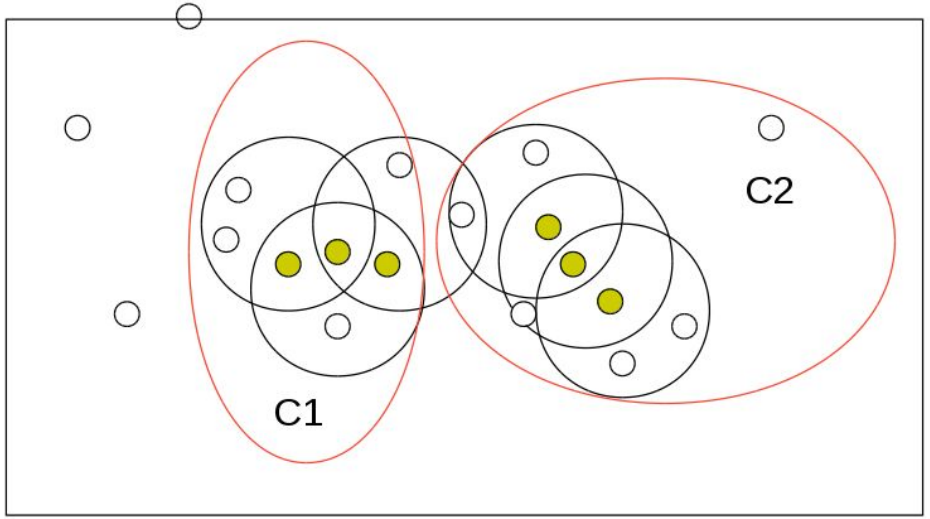
Clusters de formatos arbitrários

Cluster

Conjunto máximo de pontos conectados por densidade

Ruído

Conjunto de pontos que não pertence a algum *cluster*

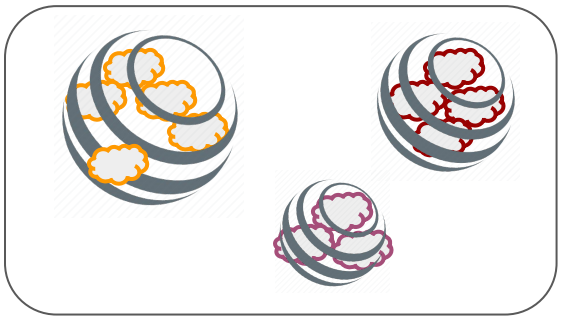


Spectral

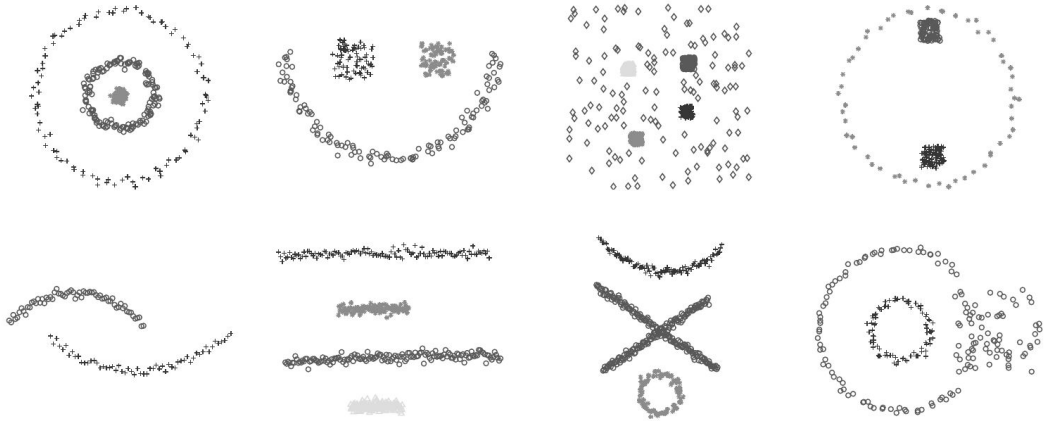
Clusters de formatos arbitrários

Motivação

Usualmente, os métodos para detecção de *clusters* captam os agrupamentos de *shape* esférico



Necessita-se de um método para detecção de *clusters* de formato arbitrário



Spectral

Clusters de formatos arbitrários

Podemos usar os autovetores de uma Matriz de Adjacência para determinar *clusters*

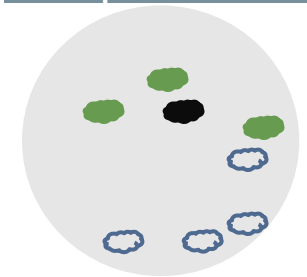
Spectral

Clusters de formatos arbitrários

1° passo: Construir Matriz de Adjacência ou Afinidade

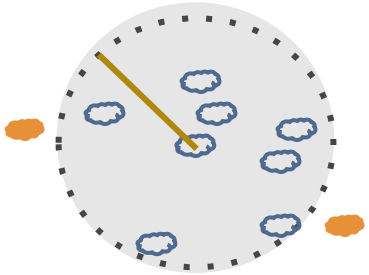
Modela a estrutura de geometria local dos dados

k-vizinhos mais próximos



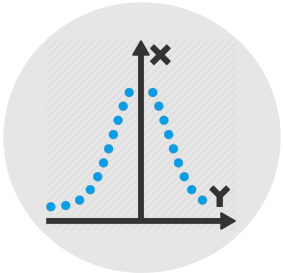
Vi conectado a Vj se:
Vj é um dos k vizinhos de Vi
ou
Vi é um dos k vizinhos de Vj

Vizinhança ε



A distância entre Vi e Vj é menor que ε

Totalmente conectado



$$W_{ij} = e^{-\frac{||x_i - x_j||^2}{\sigma^2}}$$

Spectral

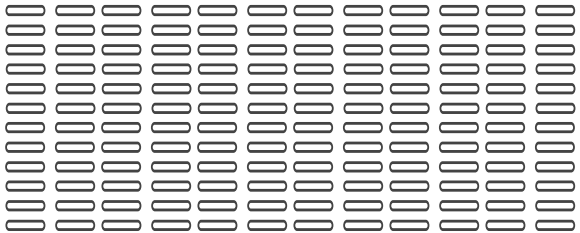
Clusters de formatos arbitrários

2° passo: Computar *graph laplacian matrices*

Passo essencial da técnica

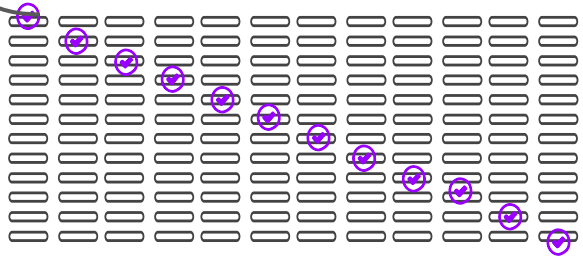
$$L = D - W$$

Matriz de Adjacência



Degree Matrix

Soma dos pesos que incidem no vértice 01



Spectral

Clusters de formatos arbitrários

3° passo: Spectrum Analysis

L pode ser representada por uma estrutura bloco-diagonal

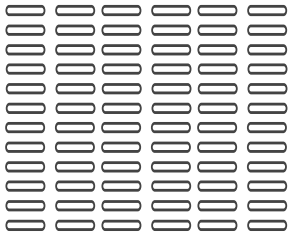
$$L = \begin{bmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{bmatrix}$$

Autovetores de L (matriz F) podem ser determinados resolvendo:

$$\min_F Tr(F^T L F) \quad \text{s.t. } F^T F = I$$

Toma k autovetores associados aos k menores autovalores de L

Constrói matriz com os k autovetores



Empregar um método usual de agrupamento

Spectral

Clusters de formatos arbitrários

1. Trabalha com a matriz de similaridade dos dados ao invés dos dados de pontos originais.
2. Se aplica a qualquer tipo de dado.
3. É possível trabalhar com dados arbitrários ao invés de pontos de dados representados em um espaço multi dimensional.
4. Não faz pressuposto sobre o formato do cluster.

1. A obtenção da matriz de similaridade é custosa computacionalmente - especialmente para muitas observações.
2. Nem todo tipo de dado tem uma métrica de similaridade/distância bem definida
3. Para dados multidimensionais o uso da matriz de similaridade ao invés das observações originais é redundante.

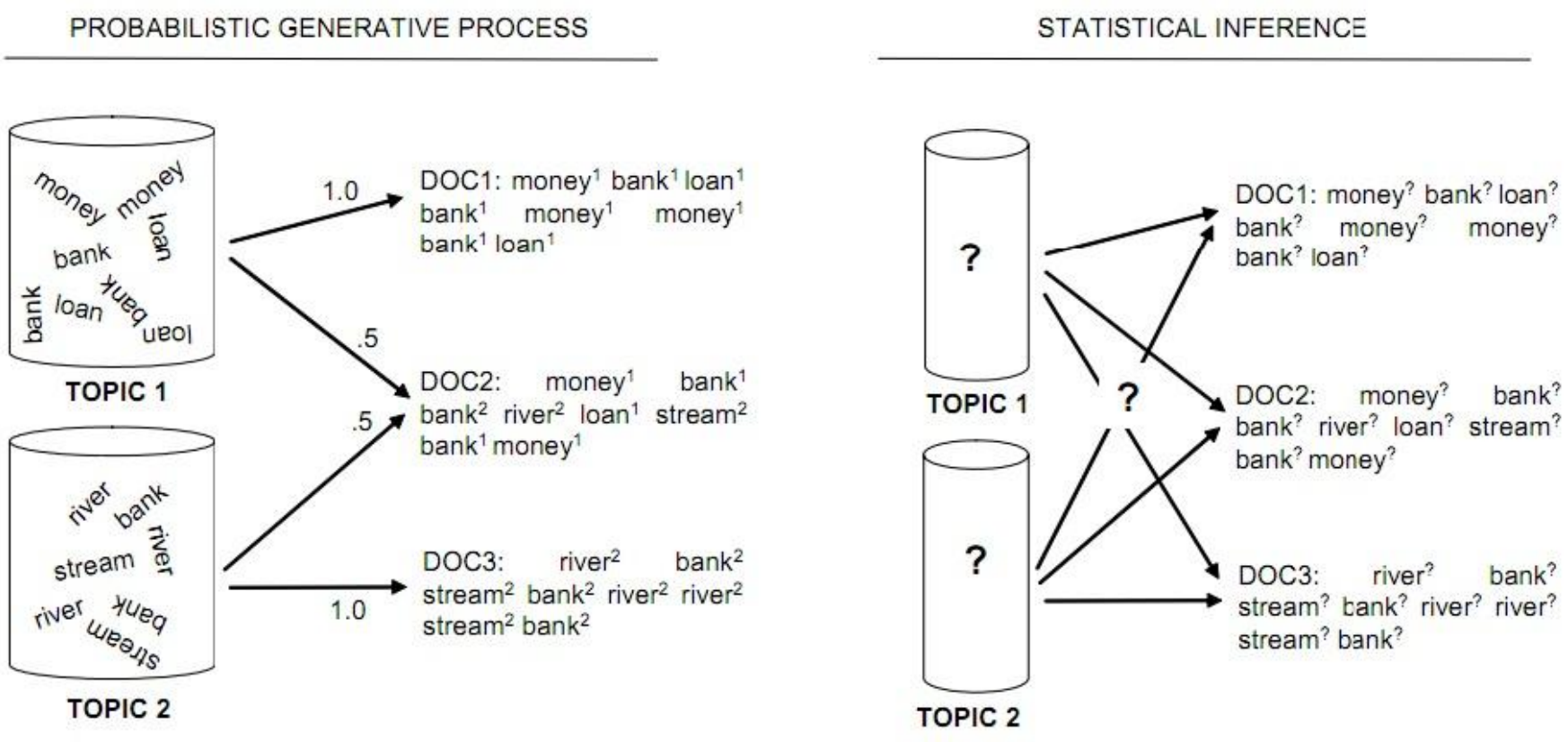
Mét. Probabilísticos - LDA

Modelo de Mistura de Tópicos para Textos

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Mét. Probabilísticos - LDA

Modelo de Mistura de Tópicos para Textos



Mét. Probabilísticos - LDA

Modelo de Mistura de Tópicos para Textos

Problema:

- ➔ Observamos apenas os documentos
- ➔ A quais tópicos cada palavra pertence?
- ➔ Quantos e quais tópicos existem?

Objetivo:

- ➔ Inferir a mistura de tópicos de cada documento
- ➔ inferir quais são os tópicos mais relevantes num conjunto de textos

Dúvidas?

Obrigado