ALGORITMO K-MEANS

Bianca Lara Gomes
biancalaragomes@hotmail.com

Introdução

O K-means é um algoritmo do tipo não supervisionado, ou seja, que não trabalha com dados rotulados. O objetivo desse algoritmo é encontrar similaridades entre os dados e agrupá-los conforme o número de cluster passado pelo argumento K. O algoritmo de forma iterativa atribui os pontos de dados ao grupo que representa a menor distância, ou seja, ao grupo de dados que seja similar.

1. DESCRIÇÃO DO ALGORITMO

O processo executado pelo K-means é composto por quadro etapas.

Inicialização: na fase de inicialização simplesmente o algoritmo gera de forma aleatória k centroids, onde o número de centroids é representado ao parâmetro K. Estes centroids são pontos de dados que serão utilizados, como pontos centrais dos clusters. Podemos pensar como referências que serão utilizadas para calcular a distância entre os dados e gerar os clusters.

Atribuição ao cluster: nesta etapa é calculado a distância entre todos os pontos de dados e cada um dos centroids. Cada registro será atribuído ao centroid ou cluster que tem a menor distância. O cálculo de distância é utilizado através da distância euclidiana. Finalizando esta etapa com os dados divididos conforme o número de centroids estipulado pelo argumento k.

Movimentação de centroids: Uma vez que os pontos de dados foram atribuídos aos clusters conforme sua distância, o próximo passo é recalcular o valor dos centróides. Nesta etapa é calculada a média dos valores dos pontos de dados de cada cluster e o valor médio será o novo centróide. O termo movimentação se refere a alteração da localização do centróide em um plano se pensarmos em um gráfico.

Otimização do K-médias: Na fase final da execução do K-means as fases Atribuição ao Cluster e Movimentação de Centroids são repetidas até o cluster se tornar estático ou algum critério de parada tenha sido atingido. O cluster se torna estático quando nenhum dos pontos de dados alteram de cluster. Um critério de parada pode ser o número de iterações máximas que o algoritmo irá fazer durante a fase de otimização. Por fim o K-means chega ao fim da sua execução dividindo os dados no número de clusters especificado pelo argumento k.

Pseudocódigo do algoritmo

Para i = 1 até k **faça**
Mk <- localização randômica do k-ésimo cluster
Fim para
Repita
Para n=1 até n **faça**
Zn <- min |mk-Xn|
Fim para
Para i = 1 até k **faça**
Xk <- { Xn : Zn = k }
Mk <- {mean (xk)
Fim para
Até M parar de alterar
Retorna z

2. APLICAÇÕES

- Segmentação de dados de acordo com categorias;

3. CONCLUSÃO

Entender como funciona o agrupamento de dados é um requisito fundamental quando se trabalha com Data Science. Sabemos também a importância dessa técnica,

pois, é base para diversos métodos e algoritmos. Compreender o processo em detalhes nos permite atingir melhores resultados. Vimos algumas aplicações e cenários onde podemos aplicar a técnica de agrupamento e seus benefícios. Por fim, entendemos o funcionamento do algoritmo K-means através do detalhamento de cada fase e seus conceitos.

4. REFERÊNCIAS

ENTENDA O ALGORITMO K-MEANS E SAIBA COMO APLICAR ESSA TÉCNICA. Disponível em: <http://minerandodados.com.br/index.php/2017/12/12/entenda-o-algoritmo-k-means/#tipos_cluster_grupos>.Acesso em: 05 maio 2018.