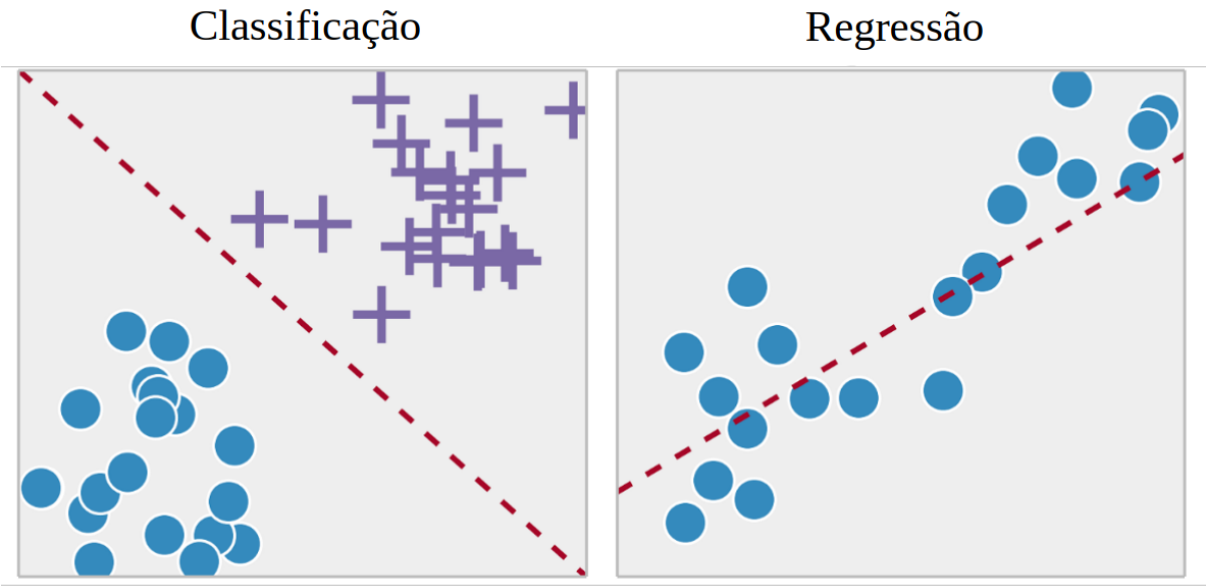
Aula 09 - Algoritmos de Regressão e de Agrupamento (clusters)

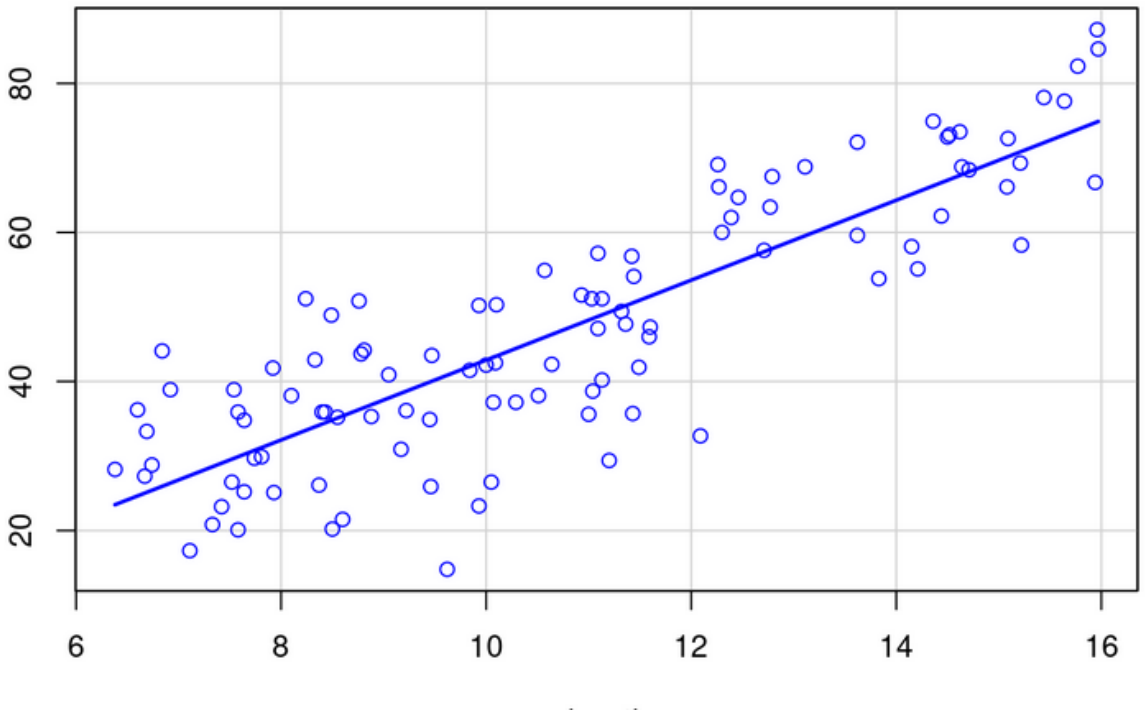
Olá, na aula anterior vimos exemplos de algoritmos de classificação e como funcionam. Nesta aula vamos ver e entender melhor o funcionamento do algoritmo de regressão. A gente aprendeu que Regressão é um tipo de tarefa supervisionada que tem o objetivo de classificar dados cujos os valores são diferentes de sim e não. Temos por exemplo as imagens abaixo:



Na imagem da esquerda temos uma reta vermelha que divide os dados entre a cruz e a bola. Ou seja, existem rótulos dos quais é para fazer a previsão. Os dados são categóricos, ou seja, se é cruz ou bola. Já na regressão os dados são numéricos. Na imagem direita temos uma linha que é tracejada entre os dados que dessa vez não são categóricos e sim numéricos. A linha representa a modelagem da relação entre essas variáveis numéricas. O que dividimos em (variável dependente **y** e variável explanatória **x**).

Mas o que isso quer dizer?
Vamos imaginar alguns exemplos:
Temperatura, pressão do ar e umidade - variável explanatória (x).
Velocidade do vento - variável dependente (y).

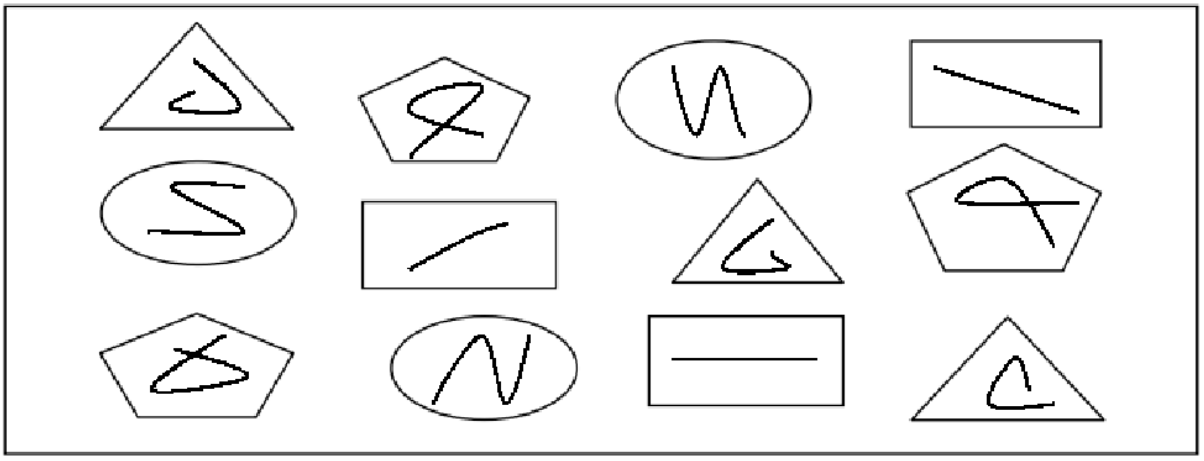
Isto é, de acordo com os nossos dados, temperatura, pressão do ar e umidade que são dados numéricos. Tentamos prever a velocidade do vento com base nestes dados.
Isto é o que chamamos de regressão linear.



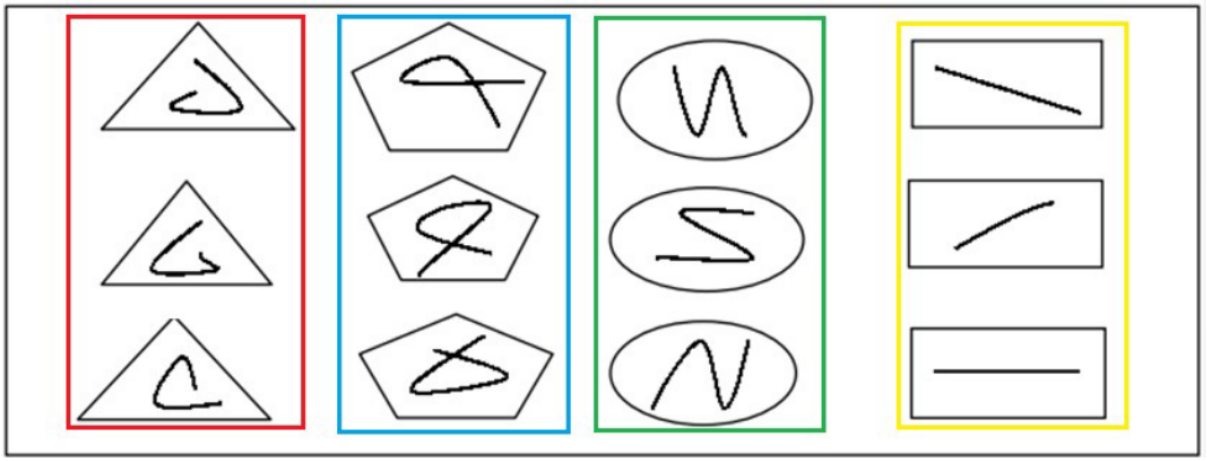
Supondo que os pontos representam os dados de temperatura, pressão do ar e umidade, a reta azul representa uma modelagem que tenta prever um possível novo valor para a velocidade do ar. O treinamento deste algoritmo é encontrar os melhores parâmetros para que esta linha se encaixe e acompanhe os dados. Então após treinar com os dados reais e encontrar a melhor localização para a linha. O algoritmo pode tentar prever qual será a velocidade do vento para uma determinada temperatura, pressão do ar e umidade que antes não estava catalogada. Então esta é a função de um algoritmo de regressão linear.

Algoritmos de Agrupamento (clusters)

Lembrando que na aprendizagem não-supervisionada o programa aprende a executar uma tarefa baseando-se em dados não rotulados, ou seja, ele encontra padrões e relações entre os dados e o próprio algoritmo rotula estes dados com base nestes padrões. Nesta aula vamos entender como funciona o agrupamento, um algoritmo de aprendizagem não-supervisionada. A aprendizagem não-supervisionada é muito utilizada em aplicação de mineração de dados (*data-mining*) onde a grande quantidade de dados não é rotulada previamente, ou seja, os dados não são conhecidos antecipadamente.
O principal aspecto do agrupamento é descobrir a organização dos padrões que existem entre os dados encontrados e assim é possível então obter as diferenças ou similaridades entre os dados e tirar conclusões úteis a respeito destes dados. Temos como exemplo a seguinte imagem:



Na imagem temos um conjunto de objetos que é passado como parâmetro para o algoritmo. E estes dados não estão previamente rotulados como na aprendizagem de máquina supervisionada. O objetivo do algoritmo será encontrar similaridades ou diferenças entre estes objetos e dividi-los em agrupamentos, também chamados de ***clusters***.



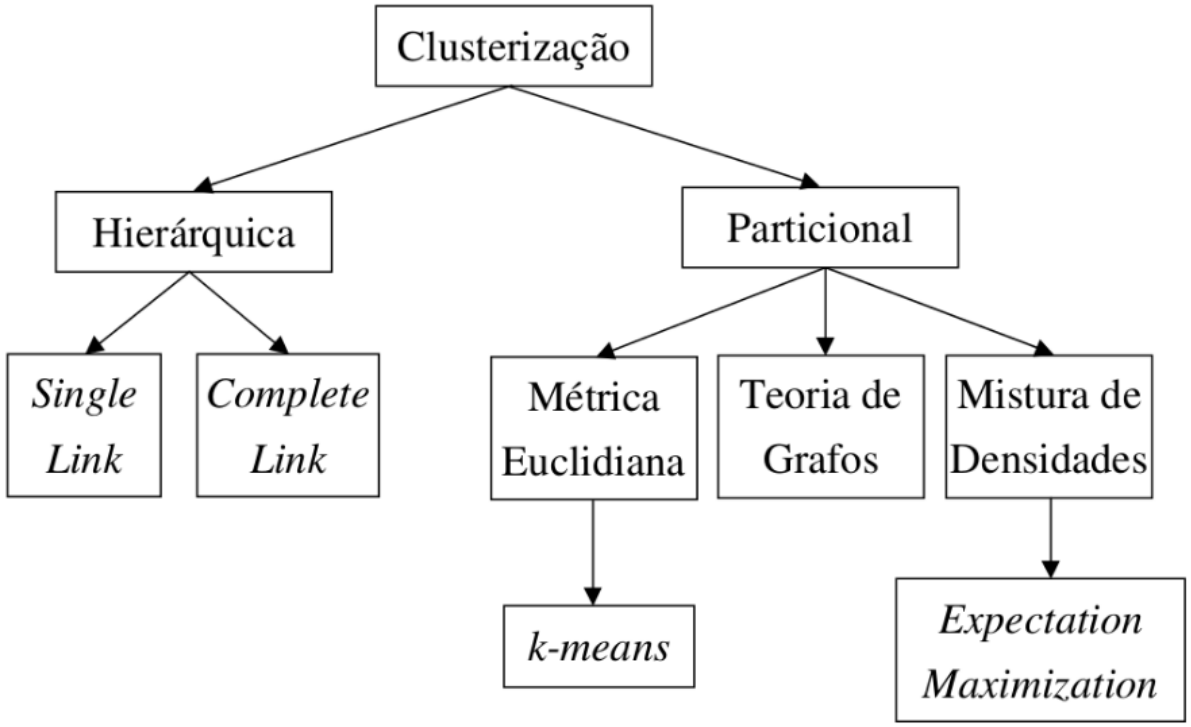
Esta imagem representa os dados, isto é, as figuras que foram a base de dados do algoritmo após ele rotular com base nas semelhanças que elas apresentam. Notem que foram criados três grupos diferentes.

Mas onde é aplicado o **agrupamento**?
É muito usado, por exemplo, em sites de recomendação de compras. Onde os clientes são separados em grupos de acordo com o tipo de compra que fazem, e assim, uma possível recomendação de compra para clientes que façam parte de um determinado grupo, pode ser mais efetivo.

Em sistemas de agrupamento de filmes/séries com base no seu gênero, por exemplo. Ou até mesmo com base no perfil de pessoas que assistem esse conteúdo. Criando assim um sistema de recomendação mais propenso a recomendar um filme ou série que um usuário que tenha perfil semelhante venha a assistir. Por exemplo, estes sistemas são muito usados nas plataformas de streaming de vídeo como prime vídeo, netflix e até mesmo o youtube.

Também pode ser usado para agrupar pacientes de um hospital baseado nos sintomas que apresentam ou características semelhantes.

Dentro do agrupamento (***clusters***) vamos ter algumas maneiras de fazer este agrupamento. A imagem abaixo nos mostra as várias maneiras de como o algoritmo vai dividir os dados para poder realizar o agrupamento.



O agrupamento é primeiro dividido em **Hierárquico** e **Particional**. Uma das principais características do Particional é que ele é um algoritmo rápido e que vai direto ao ponto. Já o Hierárquico é ainda subdividido em duas categorias: aglomerativo e decisivos. No aglomerativo é produzida uma sequência de agrupamentos decrescente de clusters (Isto é agrupamentos) em contrapartida no decisivo é produzida uma sequência crescente de clusters. Porém não vamos entrar em tantos detalhes sobre cada tipo de algoritmo. O que vamos entrar em detalhes aqui é no que é um dos mais famosos, o **K-means**.

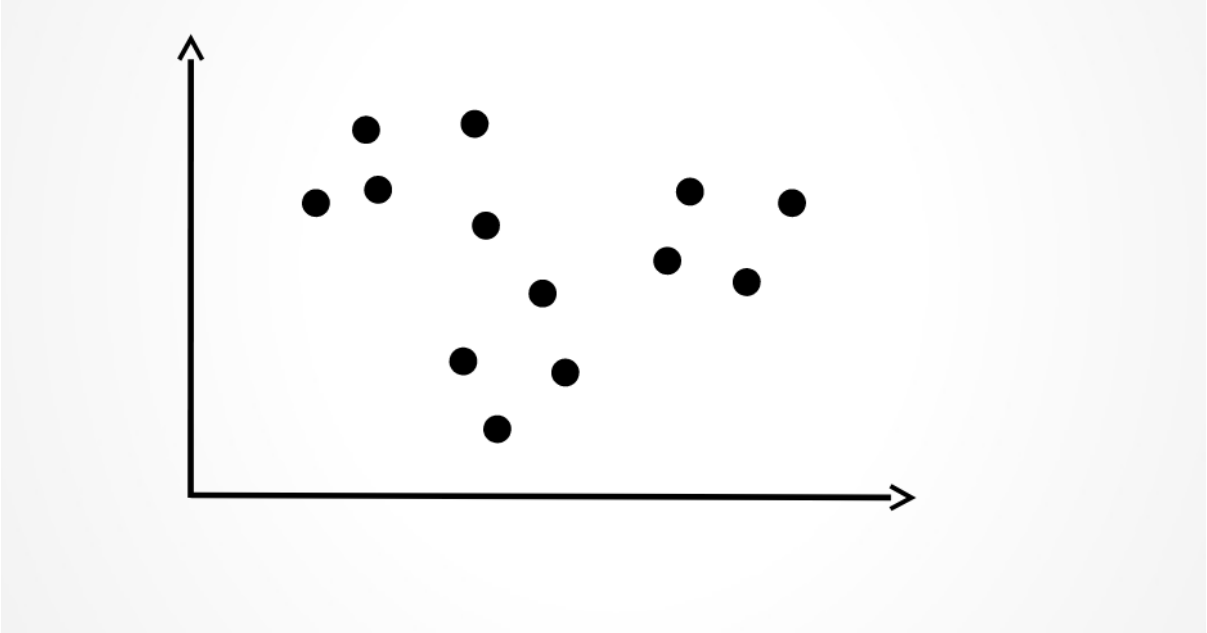
O **K-means** é do tipo particional baseado na métrica euclidiana seu objetivo é criar um modelo para avaliar e agrupar os dados de acordo com as suas características. O k-means cria os *clusters* baseado nas proximidades dos dados. Os clusters, como já dito, são agrupamentos ou grande conjunto de dados. É a técnica mais simples de aprendizado não supervisionado.

Mas como funciona o K-means?

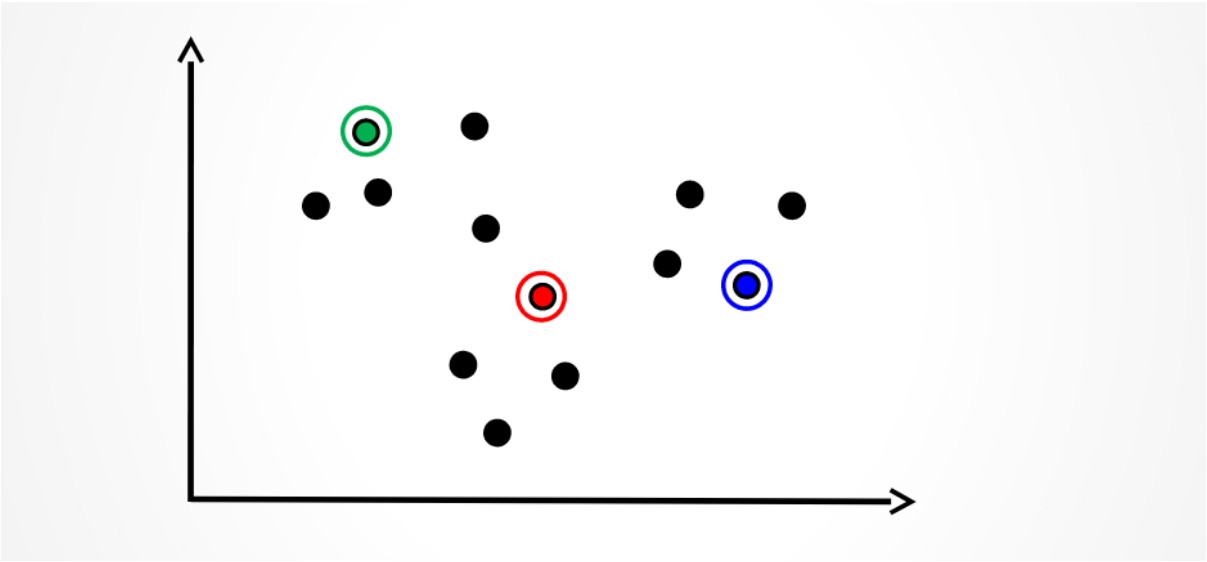
- baseia-se em um método de cálculo de distância euclidiana entre os pontos (isto é, os dados).
- O número de grupos (***clusters)*** é dado pela letra k. E cada k chamamos de centróides e estes centróides são definidos aleatoriamente.

Mas como é feito o cálculo dessa distância?

- Primeiro deve ser selecionado os k centróides iniciais.

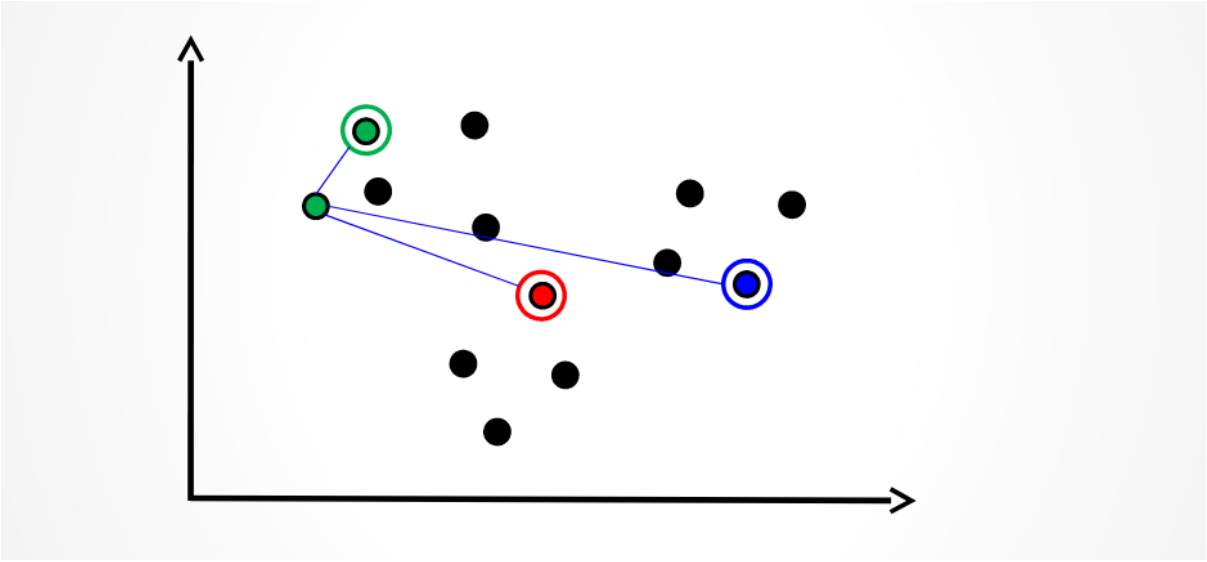


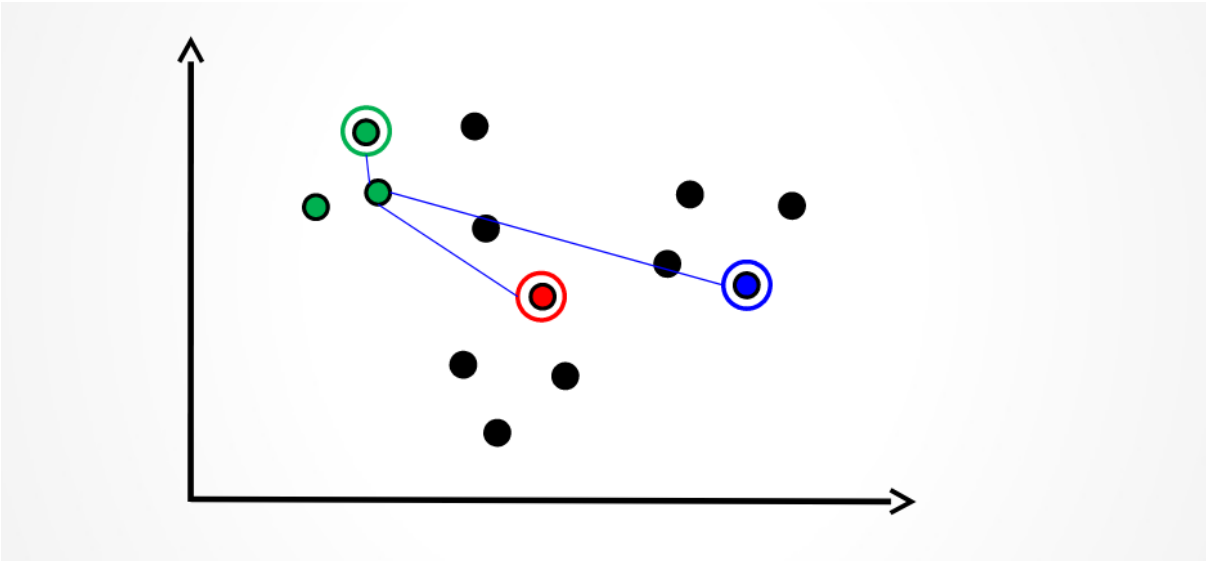
nessa base de dados, por exemplo, posicionamos k centróides.



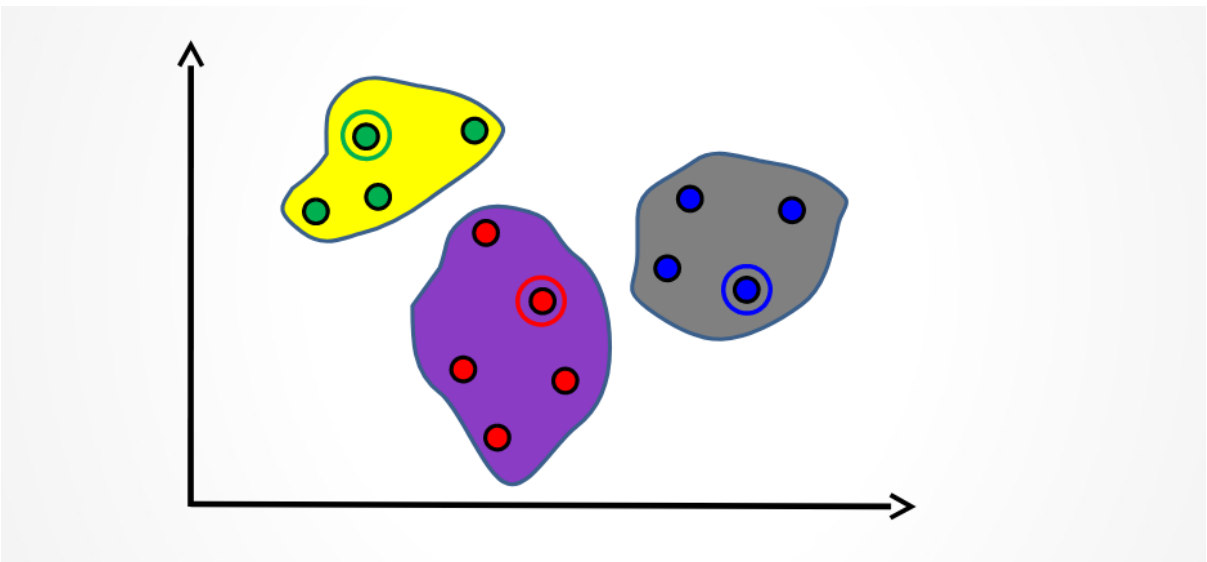
Cada círculo com cor diferente de preto, representa um centróide que foi posicionado aleatoriamente.

- Após selecionados os pontos aleatórios, será calculado a distância de cada dado (no caso os pontos pretos) entre o centróide de menor distância (os pontos coloridos). Assim então cada ponto irá pertencer ao centróide mais próximo.

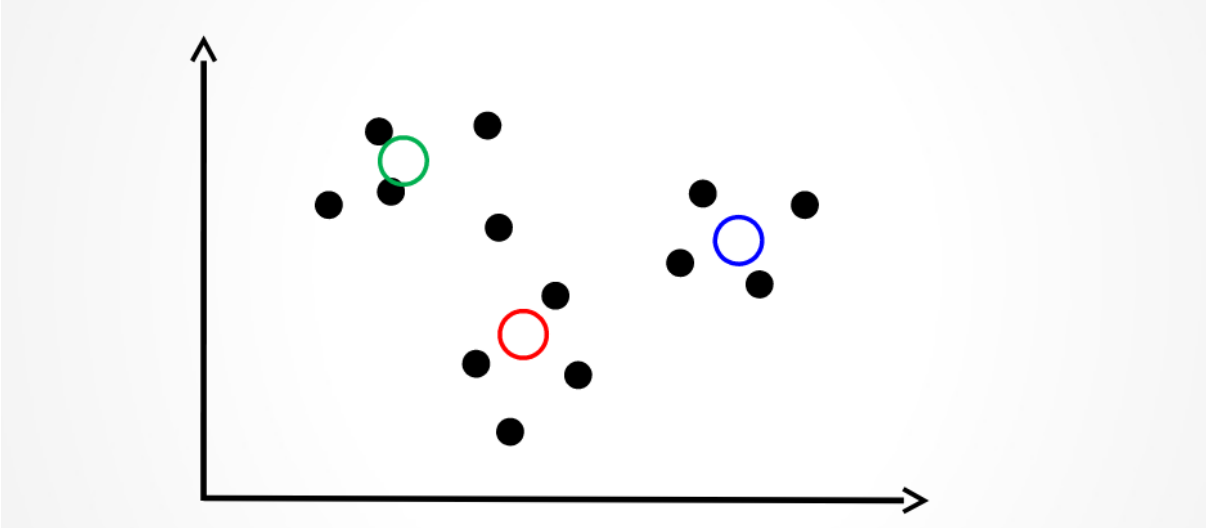




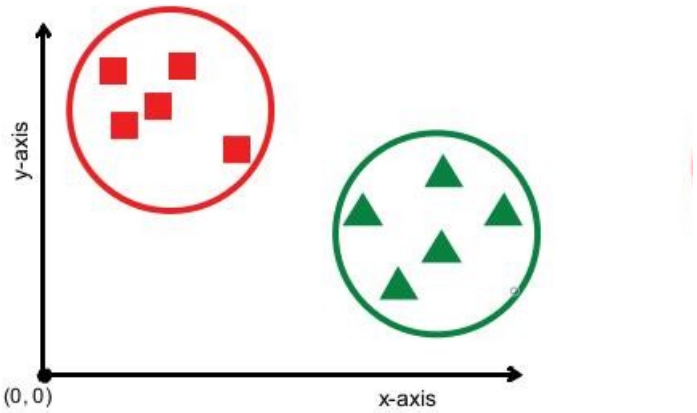
Note que o algoritmo fará isso para todos os pontos, e irá classificar cada ponto preto com base nos seus centróides mais próximos.



Após fazer isso para todos os pontos, os nossos dados irão ficar agrupados. Mas isso ainda não é o final, o algoritmo agora irá reposicionar novamente os centróides com base na média da posição de todos os pontos do (*cluster)* grupo.

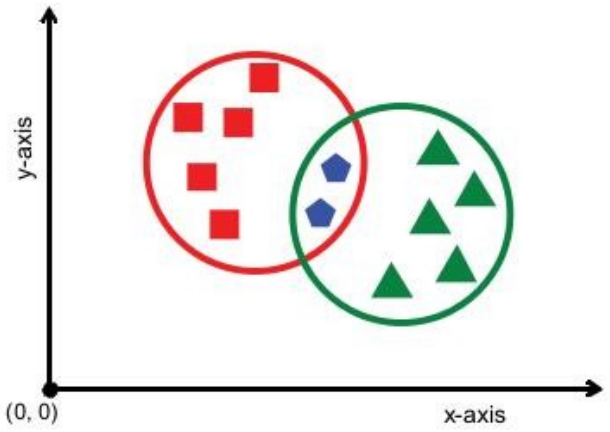


Esse processo é repetido iterativamente até todos os centróides serem redefinidos.
Ainda dentro do agrupamento do tipo **K-means** tipo diferentes de clusters que serão adequados dependendo do tipo de cenário que estamos lidando. Que são o tipo exclusivo e sobreposto.
No primeiro tipo usamos quando nossos dados são mutuamente exclusivos, isto é, pertencem apenas a um tipo de grupo. Por exemplo, um agrupamento de animais entre gatos, cachorros e cavalos. Um cachorro não pode ser definido como um gato ou um cavalo, o mesmo é válido para os demais animais. Abaixo temos uma imagem que representa este tipo de cluster.



Exclusive Clustering

Já no tipo de cluster sobreposto é quando os dados da nossa base podem pertencer a mais de um grupo. Por exemplo, se você tivesse agrupando animais por características como, mamífero, aquático, ou voador. Existem animais que cumprem mais de uma característica, como por exemplo. Golfinhos, que são aquáticos e são mamíferos, ou morcegos que são voadores e mamíferos. Ou seja, os nossos dados podem pertencer a mais de um tipo de cluster.



Overlapping Clustering

Bem, nesta aula foi visto como funcionam alguns tipos de grupamento e mais em foco o funcionamento do algoritmos k-means. Muito obrigado por ver a aula até aqui e até a próxima aula, onde veremos algoritmos de associação.