Introdução à Clusterização

Olá, seja bem-vindo!

Ao longo desse curso, você descobriu que o objetivo da técnica de aprendizado supervisionado é encontrar padrões, utilizando dados de entrada, associados aos dados de saída correspondentes, diferente do aprendizado não supervisionado, posto que, nesta técnica, não há essas associações entre os dados de entrada e os dados de saída, está lembrado? Então, se liga!

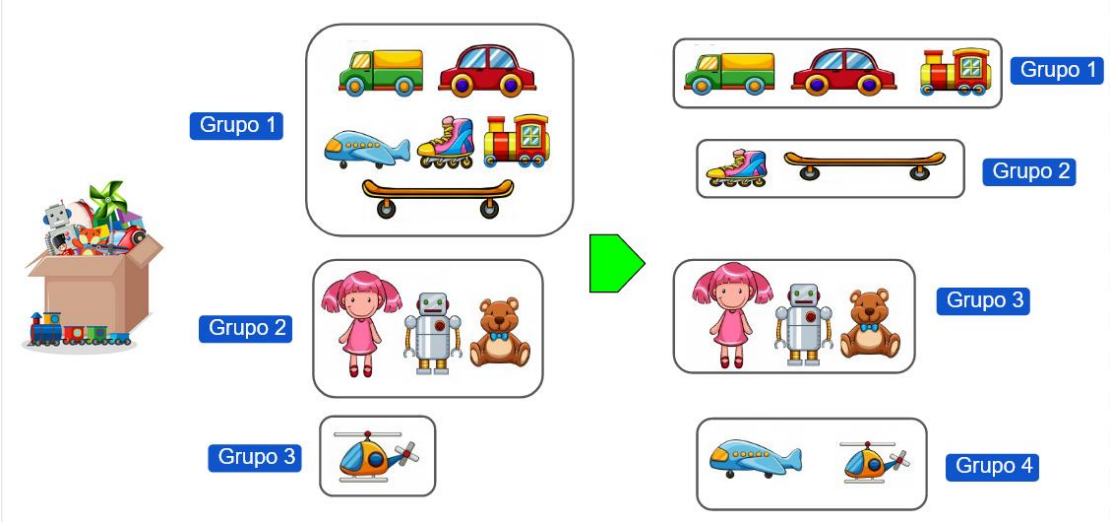
Nesta aula, você vai aprender outra técnica de Machine Learning, que faz parte do aprendizado não supervisionado, a **"clusterização"**, também conhecida como "análise de agrupamentos". Ficou curioso? Ótimo! Vamos lá!

A clusterização é uma técnica não supervisionada, cujo objetivo é agrupar, automaticamente, um número **k** de diferentes grupos em um determinado conjunto de dados. Esses grupos são chamados de "**clusters"**, sendo cada um formado por uma parte dos dados que são mais parecidos.

Esta técnica é comumente aplicada na Estatística, Engenharia, Medicina e Biologia, além de ser bastante utilizada em sistemas de recomendação de lojas e serviços online, uma vez que consegue agrupar os diferentes perfis de clientes e interesses dos usuários. Para entender melhor a aplicação dessa técnica, acompanhe o seguinte exemplo:

Uma caixa com vários brinquedos misturados que, em seguida, são divididos em grupos: o Grupo 1 possui um carrinho, um avião, um par de patins, um trem, um skate e um caminhão; o Grupo 2 tem uma boneca, um robô e um urso; e o Grupo 3 tem um helicóptero. Assim, esses grupos passam por outra divisão para serem divididos em quatro. Portanto, agora, o Grupo 1 fica com um carinho, um trem e um caminhão; o Grupo 2 com o par de patins e um skate; o Grupo 3 continua a ter os

mesmos itens, uma boneca, um robô e um urso; e, por fim, o Grupo 4 tem um avião e um helicóptero.



Imagine que uma criança possui essa caixa de brinquedos e quer organizá-los de acordo com o tipo de cada um. Sem saber, previamente, quais são os brinquedos que estão dentro da caixa, ela vai retirando um por um e separando os itens da caixa que possuem alguma similaridade. Desse modo, os objetos que possuem rodinhas são agrupados inicialmente; depois, os que possuem pernas; e assim por diante.

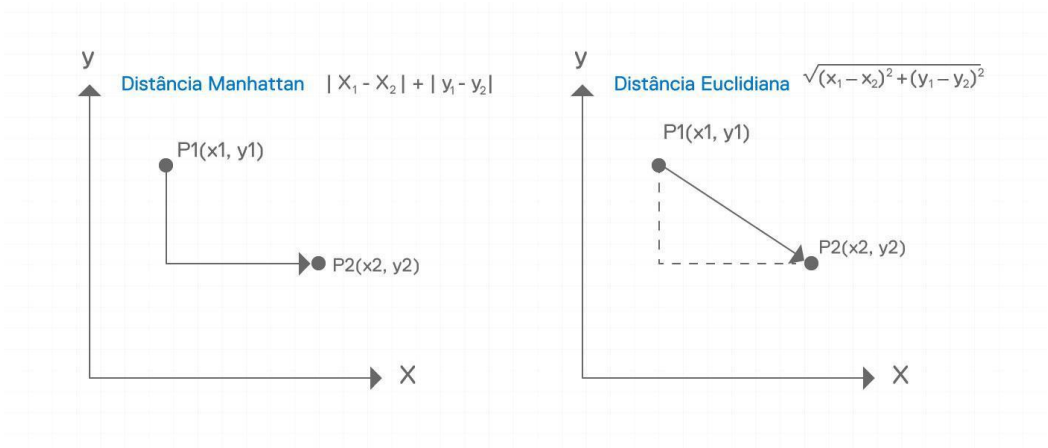
Os primeiros grupos formados, mesmo contendo elementos semelhantes, podem não seguir a melhor forma de organização, pois, analisados minuciosamente, os itens ainda poderiam ser desmembrados para novos grupos. Por exemplo, o grupo dos brinquedos que possuem rodas podem conter carrinho, trem, avião, patins, skate, entre outros. Porém, os itens patins e skate podem fazer parte de outro grupo, já que estes representam brinquedos utilizados com os pés.

Essa forma intuitiva de organização mostra que são necessários alguns passos até a formação dos grupos de itens com características equivalentes. Então, para estabelecer a semelhança ou a diferença entre os itens de um conjunto de dados, você pode usar uma função de distância, que pode ser definida de acordo com o contexto do problema. As medidas de distâncias mais comuns utilizadas em clustering são:

A primeira distância, conhecida como **“Distância de Manhattan”**, é a mais simples e tem esse nome, pois simula o percurso que um táxi faria para ir de um ponto a outro, passando pelos quarteirões na cidade de Manhattan, e considera a soma dos módulos das diferenças entre os pontos.

A segunda distância, chamada de **“Distância Euclidiana”**, utiliza a raiz quadrada da soma das diferenças, ao quadrado, entre os pontos. A terceira distância, a **“Distância de Minkowski”**, é uma generalização das duas distâncias anteriores, ou seja, das distâncias Manhattan e Euclidiana. Por fim, a **“Distância de Mahalanobis”** é uma distância que considera as correlações existentes no conjunto de dados.

Desse modo, o cálculo das distâncias Manhattan e Euclidiana pode ser compreendido nas demonstrações a seguir: À esquerda, a Distância Manhattan e sua fórmula matemática $I X_1 – X_2 I + I y_1 – y_2 I$. Logo abaixo, há uma representação dela formada pelos eixos y (vertical) e x (horizontal); e, acima, pelo ponto P1 (x1, y1), na vertical, que está ligado por um segmento de reta ao ponto P2 (x2, y2) na horizontal. À direita, a Distância Euclidiana e sua fórmula matemática $\sqrt{((x_1 – x_2)^2 + (y_1 – y_2)^2)}$. Logo abaixo, há uma representação dela, formada pelos eixos y (vertical) e x (horizontal); e, acima, pelo ponto P1 (x1, y1), que está ligado, por um segmento de reta na diagonal, ao ponto P2 (x2, y2). Desse modo, uma linha pontilhada, entre os pontos, forma um triângulo.

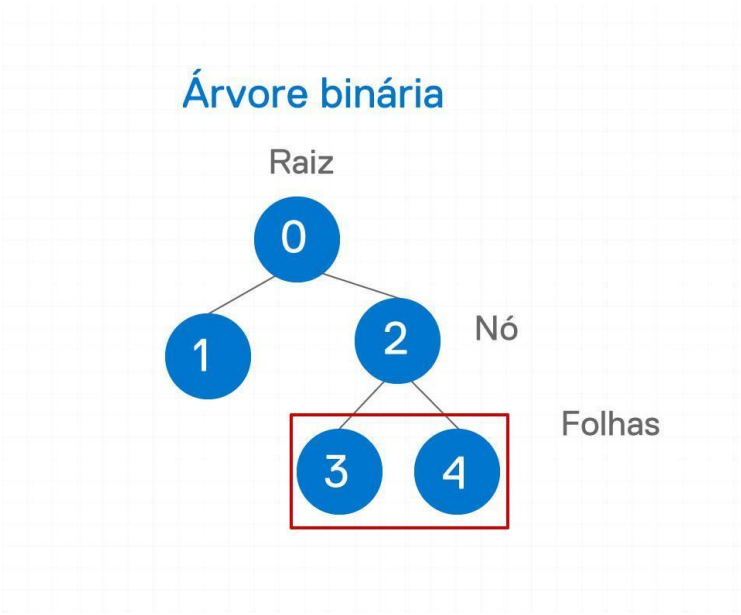


Agora que você já sabe como as distâncias entre os dados podem ser calculadas, aprenda como a clusterização pode ser dividida.

A análise de agrupamento pode ser dividida em métodos hierárquicos e não hierárquicos. No método hierárquico, os algoritmos de clusterização organizam os dados em uma estrutura de árvore binária ou em um **dendograma**. Portanto, uma árvore binária é uma estrutura de dados composta por nós.

O primeiro elemento distinto, denominado **raiz**, dá origem a dois outros nós, que funcionam como ponteiros para duas outras estruturas diferentes, denominadas subárvores, esquerda e direita, possuindo até duas folhas. No entanto, um dendograma é um tipo de árvore que subdivide os dados em outros subconjuntos de menor tamanho.

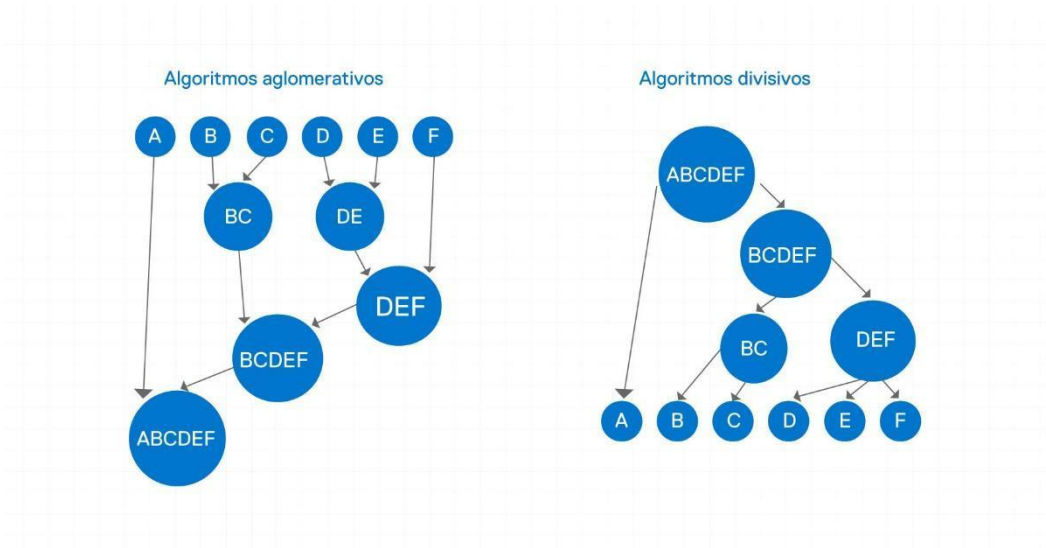
No método hierárquico, cada nó da árvore representa uma informação. Essa hierarquia pode ser observada na seguinte estrutura: uma árvore binária formada por círculos, que representam a raiz, nó e folhas e que são ligados por linhas. Portanto, a árvore tem uma raiz que é 0 de onde partem linhas que vão até os nós 1, à esquerda; e 2, à direita. Do nó 2, saem linhas que vão até as folhas 3 e 4.



Os algoritmos hierárquicos podem ser divididos ainda em aglomerativos e divisivos. Nos algoritmos aglomerativos, os dados são separados em diversos clusters e vão sendo agrupados em clusters maiores, até que se tenha um único cluster. Nos algoritmos divisivos, ocorre o inverso, uma vez que os dados começam juntos, em um único cluster, e vão se desmembrando até que se tenha diversos clusters, onde cada um deles representa um dado. Essas características, que você acabou de conhecer, estão representadas nos seguintes esquemas:

Tanto os algoritmos aglomerativos quanto os algoritmos divisivos são formados por clusters de dados, representados por círculos ligados por setas. À esquerda, nos algoritmos aglomerativos, há os clusters A, B, C, D, E e F. O cluster A se agrupa ao cluster ABCDEF. Os clusters B e C se agrupam em BC e, em seguida, se agrupam no cluster BCDEF. Os clusters D e E se agrupam em DE, e, posteriormente, se juntam ao cluster DEF. O cluster F se agrupa ao cluster DEF, que se agrupa ao cluster BCDEF. Por fim, o cluster BCDEF se agrupa no cluster ABCDEF.

À direita, nos algoritmos divisivos, há o cluster ABCDEF, que se divide no cluster A e no cluster BCDEF. O cluster BCDEF se divide no cluster BC e no DEF. O cluster BC, por sua vez, se divide nos cluster B e C. Por fim, o cluster DEF se divide nos clusters D, E e F.



Os algoritmos não hierárquicos dividem os dados em k clusters, de acordo com as distâncias calculadas entre os dados e os centros do cluster. Dessa forma, cada cluster conterá os dados que possuem a menor distância entre seus elementos e o centro do cluster, ou seja, os dados com características semelhantes, como no exemplo da caixa de brinquedos. Muita informação, não é mesmo? Calma!

Até aqui, você aprendeu como funciona a técnica de clusterização e descobriu que ela é um tipo de aprendizado não supervisionado, pois esse tipo de método é usado quando não se conhece os dados previamente. Além disso, foi conhecido como são calculadas as distâncias entre os elementos dos clusters e como funcionam os algoritmos de cluster hierárquico e não hierárquico.

Por isso, é importante ressaltar que, para um melhor aprendizado do funcionamento e aplicação das técnicas, é essencial você colocar em prática os comandos que foram abordados. Ah, e não se esqueça de aprofundar o conteúdo e resolver os exercícios.

Bons estudos e até mais!