Kümeleme (Clustering)

Kümeleme, denetimsiz öğrenmenin bir yöntemidir ve birçok alanda kullanılan istatistiksel veri analizi için yaygın bir tekniktir. Denetimsiz öğrenme, veri kümesi ile çıktıların olmadığı bir öğrenme metodudur. Veri kümesindeki verileri yorumlayarak ortak noktaları bulmak ve bunları kümeleştirme işlemi yapılarak anlamlı bir veri elde edebilmektir. Sistem, öğreten olmadan öğrenmeye çalışır. Ham verileri organize verilere dönüştüren bir makine öğrenimi türüdür.

Denetimsiz öğrenmede sadece veriler vardır onlar hakkında bilgi verilmez. Bu verilerden sonuçlar çıkarılmaya çalışılır. En baştan veriler hakkında herhangi bir bilgi verilmediği için çıkartılan sonuçların kesinlikle doğru olduğu söylenemez. Veriyi değişkenler arasındaki ilişkilere dayalı olarak kümeleyerek çeşitli modeller, yapılar oluşturabiliriz.

Örneğin, bir alışveriş sitesinde alınan bir ürünün yanında kullanıcıların alabileceği diğer ürünlerin tavsiye olarak belirlenmesi. Ya da bir hizmet satın aldığında, o hizmetle etkileşimli diğer hizmetlerin müşterinin ilgi alanına girmesi.

Kümeleme işlevinden kimlik belirlenebilir mi? Özellikle alış veriş yaparken aldıklarımız, aynı anda kimliğimiz olabilir mi?

Kümeleme Sınıflandırma Denetimsiz öğrenmede kümeleme söz konusudur. Kümeleme algoritmaları basitçe, veri kümesindeki elemanları kendi aralarında gruplamaya çalışır. Burada kaç grup olacağı veya en uygun küme sayısını algoritmanın kendisi belirler. Verilerin yakınlık, uzaklık, benzerlik gibi ölçütlere göre analiz edilerek sınıflara ayrılmasına kümeleme denir. Örnek; alışveriş yapılan bir marktte kasiyerin bir robot olduğun düşünün ve tüm ürünler birbirine karışmış olsun. Elma bulup, onu tanıyıp diğer elemanları yığın içerisinde topladğını düşünün. Seçme işlemi devam ederken yetenek kazanarak performansını atırabilir; hatalı seçtiği ürünler var ise ayıklayabilir. Böylece ürünlerin birbirlerine benzeme yakınlığı uzaklaşarak, ayrım yapma yeteneği artırılmış olur. Böylece sınılandırma da yapılmış olur.

Elmalar da kendi aralarında kümeleme yapılabilir mi? Aynı tipte verilerin değişik segmentlere bölünmüş halidir. Elinde örnekler var ama hangi veya kaç sınıfa ait olduğunu bilmiyor. Sınıfları(kümeleri) kendisi inşaa ediyor. Örneğin elimizde sadece domatesler varsa, ve bunları kalitelerine göre ayırılırsa bu kümeleme işlemidir.

Kümeleme Algoritması Türleri:

● Bağlantı Modelleri: Bu modeller veri alanındaki veri noktalarının birbirlerine daha uzaktaki veri noktalarından daha fazla benzerlik sergilediği düşüncesine dayanmaktadır.

● Merkez Modelleri:Bunlar, bir veri noktasının kümelerin merkezine yakın olmasıyla benzerlik kavramının türetildiği yinelemeli kümeleme algoritmalarıdır. K-Means kümeleme algoritması, bu kategoriye giren popüler bir algoritmadır. Bu modellerde, sonunda gerekli olan kümelerin önceden belirlenmesinden önce, veri kümesiyle ilgili önceden bilgi sahibi olmayı önemsemektedir. Bu modeller, yerel optimumu bulmak için yinelemeli olarak çalışır.

● Dağılım Modelleri:Bu kümeleme modelleri, kümedeki tüm veri noktalarının aynı dağılıma (örneğin: Normal, Gauss) ait olma ihtimali üzerine kuruludur. Bu modeller çoğunlukla aşırı uyum gösterir. Bu modellerin popüler bir örneği, çok değişkenli normal dağılımları kullanan beklenti maksimizasyon algoritmasıdır.

● Yoğunluk Modelleri: Bu modeller, veri alanındaki veri noktalarının yoğunluğunun yoğun olduğu alanlar için veri alanını arar. Farklı yoğunluk bölgelerini izole eder ve bu bölgelerdeki veri noktalarını aynı kümeye atar. Yoğunluk modellerinin popüler örnekleri DBSCAN ve OPTICS'dir.

1 Parçalama Tabanlı

Her bir veri nesnesi tam olarak bir alt kümede olduğu gibi, veri nesneleri kümesinin basitçe örtüşmeyen alt kümelere (kümeler) bölünmesidir.

2 Hiyerarşik Tabanlı

Hiyerarşik kümelenme, adından da anlaşılacağı gibi, kümelerin hiyerarşisini oluşturan bir algoritmadır. Bu algoritma, kendi kümelerinin bir kümesine atanan tüm veri noktalarıyla başlar. Daha sonra, en yakın iki küme aynı kümeye birleştirilir. Sonunda, bu algoritma sadece tek bir küme kaldığında sona erer.

3 Yoğunluk Tabanlı

Yoğunluğa dayalı kümelemede, kümeleri, veri kümesinin geri kalanından daha yüksek yoğunluklu alanlar olarak tanımlanır. Bu seyrek alanlardaki nesneler - kümeleri ayırmak için gerekli olan - genellikle gürültü ve sınır noktaları olarak kabul edilir. En popüler yoğunluk tabanlı kümeleme metodu DBSCAN'dır. Birçok yeni yöntemin aksine, "yoğunluk-erişilebilirlik" adı verilen iyi tanımlanmış bir küme modeli özellikleri. Bağlantı tabanlı kümelenmeye benzer şekilde, belirli

mesafe eşikleri içindeki bağlantı noktalarına dayanır. Bununla birlikte, bu yarıçaptaki minimum sayıda nesne olarak tanımlanan orijinal varyantta yalnızca bir yoğunluk ölçütünü karşılayan noktaları birleştirir. Bir küme tüm yoğunluk bağlı nesnelere (birçok başka yöntemin aksine rasgele bir şekle sahip bir küme oluşturabilir) ve bu nesnelerin menzilindeki tüm nesneleri içerir.

4 Izgara Tabanlı

Çoğu kümeleme algoritmasının hesaplama karmaşıklığı, en azından veri kümesinin boyutuna doğrusal olarak orantılıdır. Izgara tabanlı kümelenmenin en büyük avantajı, özellikle çok büyük veri kümelerini kümelemek için hesaplama karmaşıklığının önemli ölçüde azaltılmasıdır. Izgara tabanlı kümeleme yaklaşımı, veri kümeleriyle değil, veri noktalarını çevreleyen değer alanıyla ilgili olduğu için geleneksel kümeleme algoritmalarından farklıdır.

Genel olarak, tipik bir grid tabanlı kümeleme algoritması aşağıdaki beş temel adımdan oluşur (Grabusts ve Borisov, 2002):

1. Izgara yapısının oluşturulması, yani, veri boşluğunun sınırlı sayıda hücreye ayrılması.
2. Her hücre için hücre yoğunluğunun hesaplanması.
3. Hücrelerin yoğunluklarına göre sınıflandırılması.
4. Küme merkezlerini belirleme.

5 Model Tabanlı

Modele dayalı kümeleme, verilerin bir model tarafından oluşturulduğunu ve orijinal modeli verilerinden kurtarmaya çalıştığını varsayar. Verilerden elde ettiğimiz model daha sonra kümeleri ve nesnelerin kümelere atanmasını tanımlar.