Merhaba arkadaşlar, sizlere clustering türlerinden bahsedeceğim.

Clustering, hard clustering ve soft clustering olmak üzere ikiye ayrılır. İkisi de veri noktalarını gruplara ayırmak için kullanılan farklı yöntemlerdir. Önce hard clustering’e sonra da soft clustering’e göz atalım.

**Hard Clustering:**

Hard clustering, veri noktalarını ayrı gruplara (kümelere) ayıran bir kümeleleme yöntemidir. Burada, her veri noktası yalnızca bir kümeyle ilişkilendirilir. Yani, bir veri noktası tam olarak bir kümenin içinde bulunur. En popüler hard clustering algoritması olan K-Means'i ele alalım.

Örneğin, bir mağaza müşterilerinin demografik verilerine sahipsiniz ve bu müşterileri belirli özelliklere göre gruplara ayırmak istiyorsunuz. K-Means algoritması, belirli bir sayıda (K) küme oluşturur ve her müşteriyi bu kümelerden birine atar. Her müşteri yalnızca bir kümeyle ilişkilendirilir.

Örneğin, demografik özelliklerine dayalı olarak müşterileri "genç", "orta yaşlı" ve "yaşlı" olarak üç farklı kümeye ayırmak isteyebilirsiniz. K-Means, her müşteriyi bu üç kümeye atar.

**Soft Clustering:**

Soft clustering, hard clustering'in aksine, veri noktalarını kesin bir şekilde bir kümeye atamak yerine, her bir veri noktasının her kümeyle ilişkisinin olasılığını belirler. Bu nedenle, bir veri noktası birden fazla kümeye ait olabilir. Bir örnek olarak, Fuzzy C-Means (FCM) algoritmasını ele alalım.

Örneğin, bir fotoğraf koleksiyonunu gruplara ayırmak istiyorsunuz ve bazı fotoğrafların birden fazla konuya ait olabileceğini biliyorsunuz. FCM, her fotoğrafın her kümeye ait olma olasılığını belirler. Örneğin, bir deniz manzarası fotoğrafı hem "doğa manzaraları" hem de "deniz manzaraları" kategorilerine ait olabilir. FCM, bu fotoğrafı her iki kategoriye de kısmi olarak atar.

Özetle, hard clustering her veri noktasını sadece bir kümeyle ilişkilendirirken, soft clustering her veri noktasının her kümeyle ilişkisinin olasılığını belirler. Bu nedenle, soft clustering daha esnek bir gruplama yöntemidir.

Şimdi de clustering method ve algoritmalarına göz atalım.

**Centroid Based Clustering:**

Centroid-based clustering, yani merkez nokta tabanlı kümeleme, veri noktalarını bir merkez nokta etrafında gruplamak için kullanılan bir yöntemdir. Bu yöntemde, her bir kümenin bir merkezi vardır buna merkez nokta veya centroid denir, ve bu merkez noktaları çevresindeki veri noktalarını gruplamak için kullanılır. En yaygın centroid tabanlı kümeleme algoritmalarından biri K-Means'tir.

**K-Means Algoritması:**

K-Means, veri noktalarını K sayısında küme veya grup halinde gruplamak için kullanılan bir centroid tabanlı kümeleme algoritmasıdır. İşleyiş şu adımları izler:

**Küme Merkezlerinin Başlangıç Konumlarının Belirlenmesi:** İlk olarak, K adet rastgele küme merkezi (centroid) seçilir.

**Veri Noktalarının Kümelere Atanması:** Her veri noktası, en yakın olan küme merkezine atanır. Bu atama, öklid mesafesi gibi bir uzaklık ölçüsü kullanılarak yapılır.

**Küme Merkezlerinin Güncellenmesi:** Her bir kümenin merkezi, o kümeye ait tüm veri noktalarının ortalaması alınarak güncellenir.

**Yeniden Atama ve Merkez Güncelleme:** 2. ve 3. adımlar, küme atamaları ve merkez güncellemeleri artık değişmeyene kadar tekrarlanır.

**Algoritmanın Sonlanması:** Küme merkezleri ve veri noktalarının atamaları artık değişmezse veya belirli bir iterasyon sayısına ulaşılırsa algoritma sonlanır.

Örneğin, bir şehirdeki konut fiyatlarını analiz etmek istediğimizi varsayalım. Bu örnekte, evlerin metrekare büyüklüğü ve oda sayısı gibi özelliklerini kullanarak evleri belirli fiyat segmentlerine ayırmak istiyoruz. K-Means algoritmasını kullanarak bu verileri belirli sayıda küme veya segmente ayırabiliriz. Örneğin, 3 farklı fiyat segmenti belirleyebiliriz: düşük fiyatlı, orta fiyatlı ve yüksek fiyatlı evler. Algoritma tamamlandığında, her bir evin hangi fiyat segmentine ait olduğu ve her bir segmentin temsil edici özelliklerinin neler olduğu hakkında bilgi edinebiliriz.

Bu şekilde, K-Means algoritması kullanılarak konut fiyatlarını belirli fiyat segmentlerine ayırarak veri analizi yapabiliriz. Bu analiz, emlak piyasasındaki trendleri anlamak, fiyat farklılıklarını belirlemek ve fiyat tahminleri yapmak gibi birçok uygulama için değerli olabilir.

Şimdi de Density-based Clustering’den bahsetmek üzere Zeynep arkadaşıma sözü devrediyorum.