Gözetimsiz Öğrenme

- K-Means
- Hiyerarşik Kümeleme Analizi
- Temel Bileşen Analizi

Gözetimsiz öğrenme, makine öğrenmesi modellerinde bağımlı değişkenin elimizde olmadığı durumlarda gözlem birimlerinin bir şekilde nitelendirilme çabası olarak ifade edilebilir.

Uygulama Örnekleri

→Belge Sınıflandırması

Belgeleri etiketlere, konulara ve belgenin içeriğine göre birden fazla kategoride kümeleme

→Suç Yerlerinin Belirlenmesi

Bir şehirdeki belirli bölgelerde mevcut olan suçlarla ilgili veriler, suç kategorisi, suç alanı ve ikisi arasındaki ilişki, bir şehirdeki ya da bölgedeki suça eğilimli alanlara ilişkin kaliteli bilgiler verebilir.

→Müşteri Segmentasyonu

Kümeleme, pazarlamacıların müşteri tabanını geliştirmelerine, hedef alanlarda çalışmasına ve müşterileri satın alma geçmişine, ilgi alanlarına veya etkinlik izlemeye göre segmentlere ayırmasına yardımcı olur.

Uygulama Örnekleri

→Oyuncu Analizi

Oyuncu istatistiklerini analiz etmek, spor dünyasının her zaman kritik bir unsuru olmuştur ve artan rekabetle birlikte, makine öğrenmenin burada oynayacağı kritik bir rol vardır.

→Dolandırıcılık Tespiti

Makine öğrenimi sahtekarlık tespitinde önemli bir rol oynar ve otomobil, sağlık ve sigorta sahtekarlığı tespitinde sayısız uygulamaya sahiptir. Sahte iddialarla ilgili geçmiş verileri kullanarak, yeni iddiaları, sahte kalıpları belirten kümelere yakınlığına dayanarak izole etmek mümkündür.

→Çağrı Kaydı Detay Analizi

Bir çağrı detay kaydı (CDR), telekom şirketleri tarafından bir müşterinin araması, SMS ve internet etkinliği sırasında elde edilen bilgilerdir. Bu bilgiler, müşteri demografisiyle birlikte kullanıldığında, müşterinin ihtiyaçları hakkında daha fazla bilgi sağlar. Kümeleme algoritmaları kullanarak müşteri faaliyetlerini 24 saat boyunca kümelendirebiliriz.

Uygulama Örnekleri

→BT Uyarılarının Otomatik Kümelenmesi

Ağ, depolama veya veritabanı gibi büyük kurumsal BT altyapı teknolojisi bileşenleri büyük hacimli uyarı mesajları üretir. Uyarı mesajları potansiyel olarak operasyonel sorunlara işaret ettiğinden, sonraki işlemler için önceliklendirme için manuel olarak taranmaları gerekir. Verilerin kümelenmesi, uyarı kategorileri hakkında bilgi verebilir ve ortalama onarım süresi ve arıza tahminlerinde yardımcı olabilir.

Burada yazmayan daha birçok ugulamada da kümeleme algoritması kullanılabilir.

K-Means

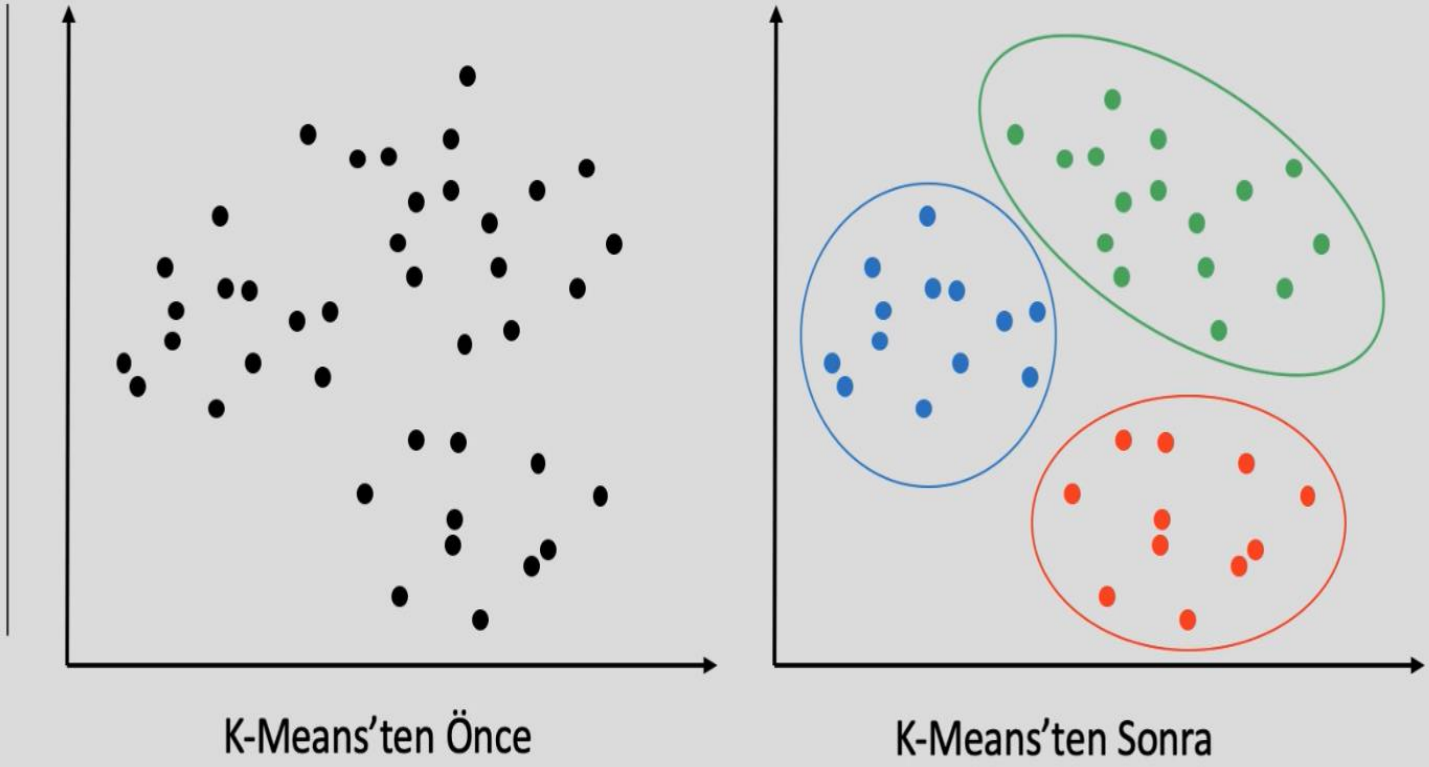
Kümelemenin Amacı

Amaç gözlemleri birbirlerine olan benzerliklerine göre kümelere ayırmaktır.

Oluşturulmaya çalışılan kümelerin kendi içinde homojen, birbirlerine göre ise heterojen olması istenir. *Kümeler içi benzerlikler yüksek kümeler arası benzerlikler az olması istenir.*

K-Means, Hiyerarşik olmayan bir kümeleme yöntemidir.

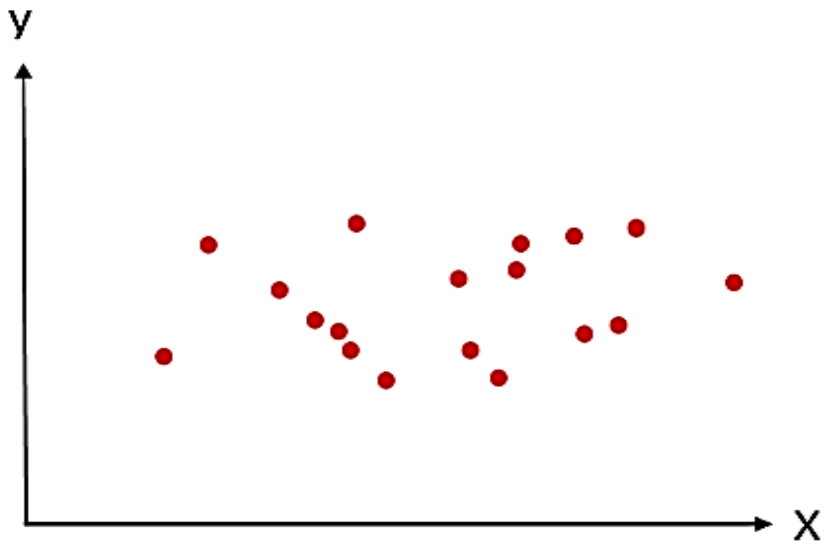
K-Means



Örneğin elimizde müşterilerimiz var ve bu müşterilerin bazı açılardan bize sağladığı katkılar var. Ve mesela biz iş gücümüzü bu müşterilerimizi ikiye bölüp farklı farklı kullanmak istiyoruz. Müşteri sayısı ve değişken çoksa göz yordamıyla müşteriyi ikiye ayıramayız. Bu durumda makine modelleme yöntemleri kullanarak birbirine benzer müşterileri iki gruba otomatik ayırabiliriz. Bu durumda kümeleme yöntemlerinden faydalanabiliriz.

K-Means Adımları

- **Adım1**: Küme sayısı belirlenir. **(k)**
- **Adım 2**: Rastgele k merkez seçilir.
- **Adım 3**: Her gözlem için k merkezlere uzaklıklar hesaplanır ve gözlemler
  kendisine en yakın k merkezine atanır.
- **Adım 4**: Her gözlem en yakın olduğu merkeze yani kümeye atanır.
- **Adım 5**: Atama işlemlerinden sonra oluşan kümeler için tekrar merkez hesaplamaları yapılır.
- **Adım 6**: Bu işlem belirlenen bir iterasyon adedince tekrar edilir ve küme içi hata kareler toplamlarının toplamının (total within-cluster variation) minimum olduğu durumdaki gözlemlerin kümelenme yapısı nihai kümelenme olarak seçilir.

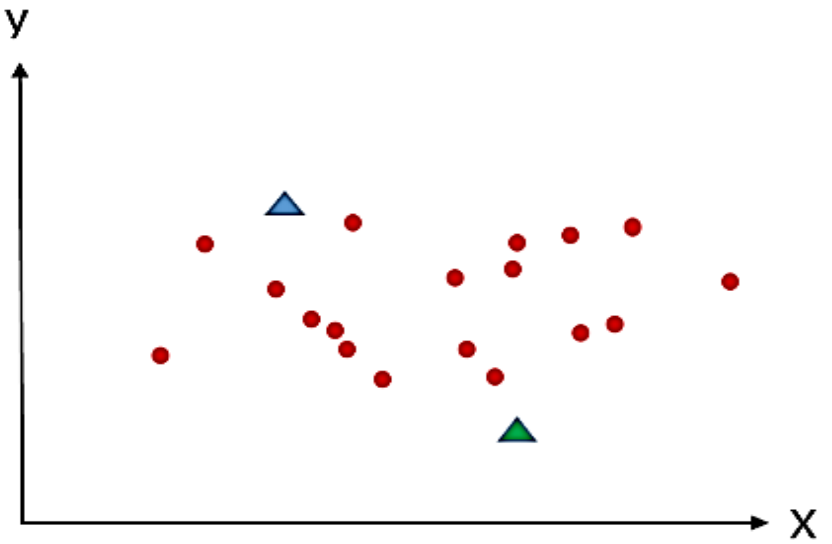


Elimizde yukarıda görülen veri seti olsun. X bağımsız değişken y ise bağımlı değişken olsun. Biz yukarıdaki veri setini kümelere ayıralım ve belirlediğimiz adımları uygulayalım:

1. Küme sayısını belirle:

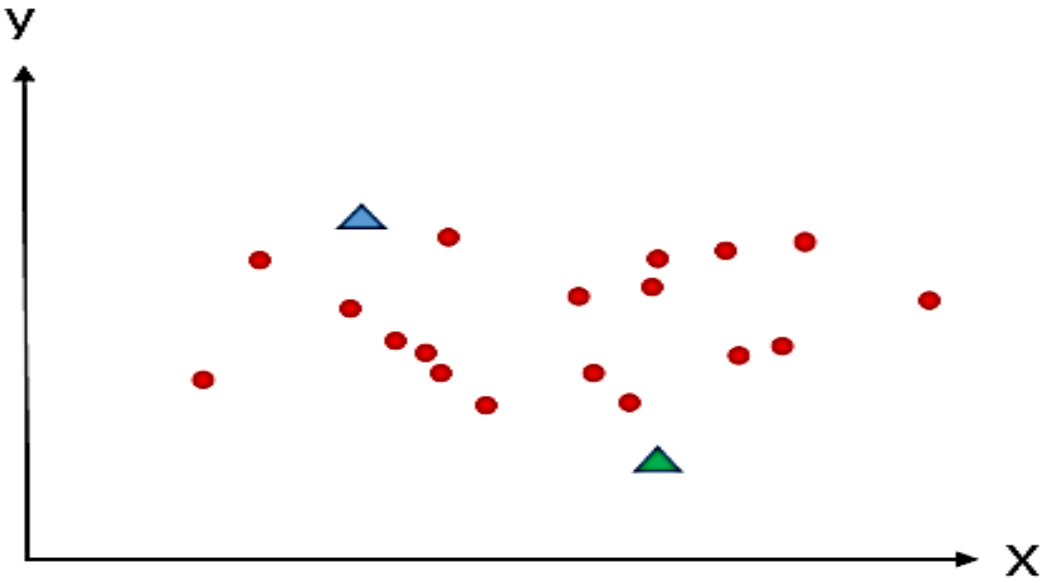
Ben küme sayısını 2 olarak belirliyorum.

2. Rastgele iki küme merkezi belirle: Mavi üçgen bir başlangıç küme merkezini, yeşil üçgen ise diğer başlangıç küme merkezi olsun.



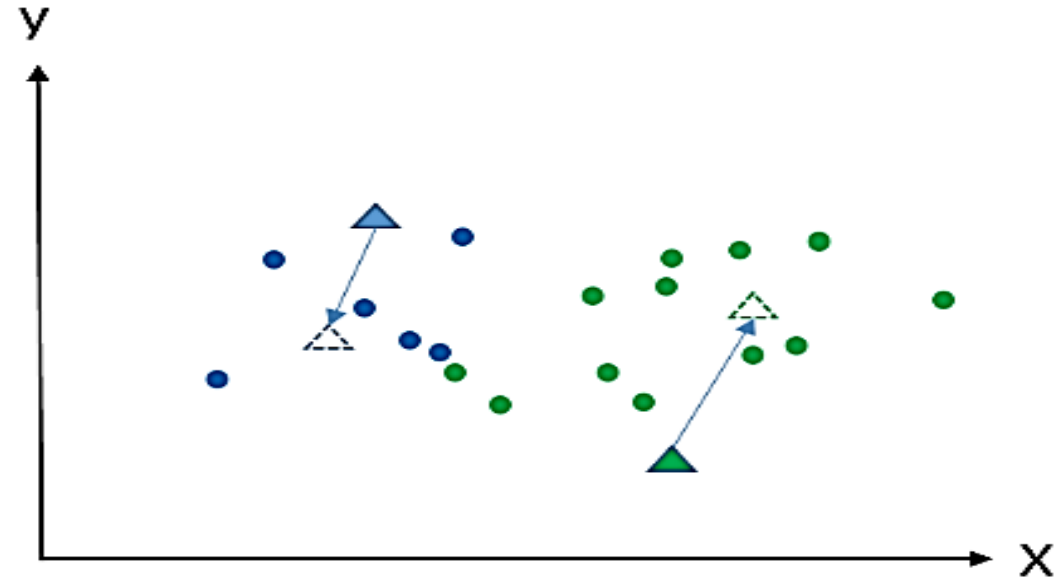
3. Seçilen bu merkezlere en yakın noktaları bulmak:

Kabaca bu merkezlere en yakın noktaları bulalım. Bunun için basit bir geometrik yöntem kullanacağız.

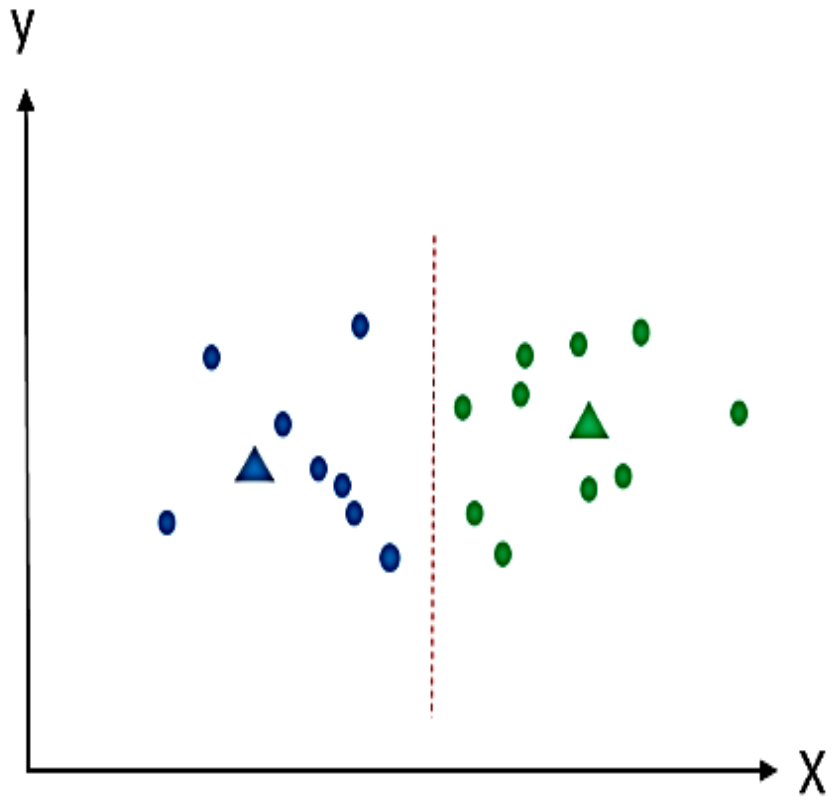


4. Yeni küme merkezlerini hesaplamak

Küme merkezlerini tekrar hesaplayıp kaydırıyoruz.



5. Yeni küme merkezlerine göre noktaları tekrar kümelemek:



Yukarıda iki nokta yeşil kümede iken mavi kümeye geçti. Bu adımdan sonra yine yeni küme merkezleri hesaplanacak ve yeni küme merkezlerine göre noktalar yeniden kümelere atanacak ta ki herhangi bir hareket mümkün olmayana kadar. İşte bu nedenle başlangıçta küme merkezlerini rastgele seçmek çok mantıksız görünse bile çeşmeden dökülen suyun öyle veya böyle bir şekilde lavabo giderine ulaştığı gibi başlangıç küme merkezleri ne seçilirse seçilsin kümeler de kendilerine benzeyen noktalarla bir araya gelecektir.

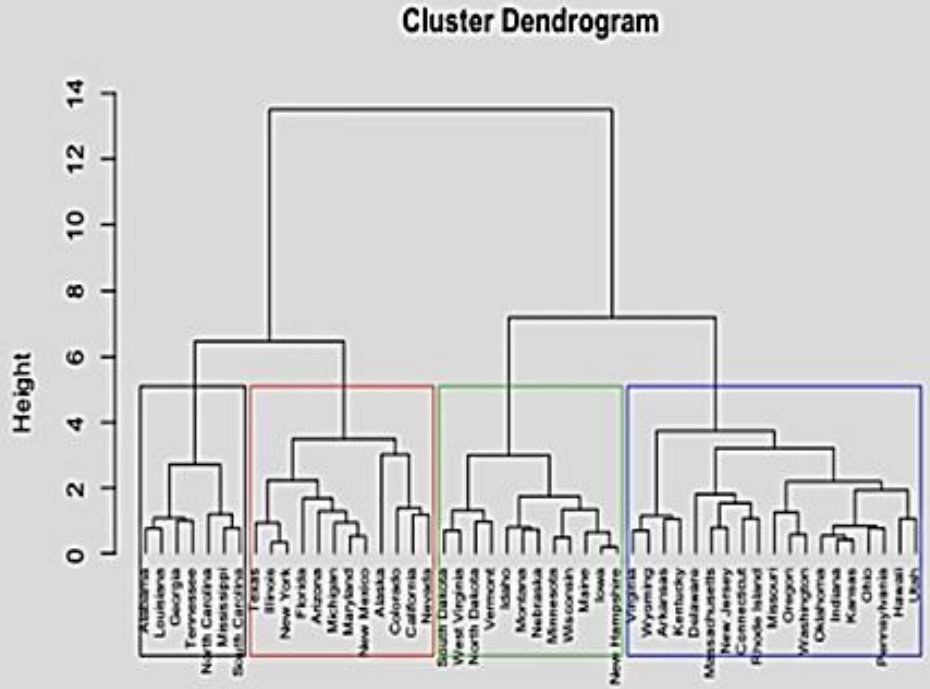
Ancak başlangıç noktalarının seçimi algoritmayı belli bir yönde kümeleme modeli kurmaya dikte edebilir. Bu durum rastgele başlangıç tuzağı (random intitalization trap) olarak adlandırılıyor. k-means++ başlatma metodu (initialization method) ile tuzağın önüne geçilmeye çalışılmış.

Hiyerarşik Kümeleme

Amaç gözlemleri birbirlerine olan benzerliklerine göre alt kümelere ayırmaktır.

Gözlemler daha fazla alt kümeye ayrılmak istenildiğinde hiyerarşik yöntemler gündeme gelmektedir.

Hiyerarşik Kümeleme



Agglomerative
Birleştirici

Divisive
Bölümleyici

- Gözlemler en alttan birbirlerine benzerliklerine göre kümelenir
- Daha sonra bir benzerlik daha elde edilir, bir benzerlik daha …
- Bu şekilde benzerliklerine göre bir araya getirilirler

- Tüm gözlemler bir aradadır.
- Daha sonra bu tüm gözlemler 2 kümeye ayrılır.
- Daha sonra bu kümeler birbirlerine benzemeyen yine 2 ‘şer alt kümelere ayrılır.
- Bu şekilde aşağı doğru bölünme işlemi gerçekleşir.

Birleştirici Kümeleme

Başlangıçta gözlem sayısı kadar küme vardır.

Adım 1: Veri setinde birbirine en yakın olan iki gözlem bulunur.

Adım 2: Bu iki nokta bir araya getirilerek yeni bir gözlem oluşturulur. Yani artık veri seti ilk birleşimdeki gözlemlerden oluşmaktadır.

Adım 3: Aynı işlem tekrarlanarak yukarı doğru çıkılır. Yani iki kümenin birleşiminden oluşan bu yeni kümeler aynı şekilde birbirlerine benzerliklerine göre tekrar birleştirilir. Bu işlem tüm gözlemler tek bir küme de toplanana kadar bu işlemler tekrar edilir.

Birbirine yakın noktalar uzaklık ölçüleri kullanarak belirlenir. Öklit uzaklığı, manhattan uzaklığı, korelasyon vs

Bölümleyici Kümeleme

Başlangıçta 1 tane küme vardır, o da tüm veri setidir.

Adım 1: Tüm gözlemlerin bir arada olduğu küme iki alt kümeye ayrılır.

Adım 2: Oluşan yeni kümeler birbirlerine benzemeyen alt kümelere bölünür.

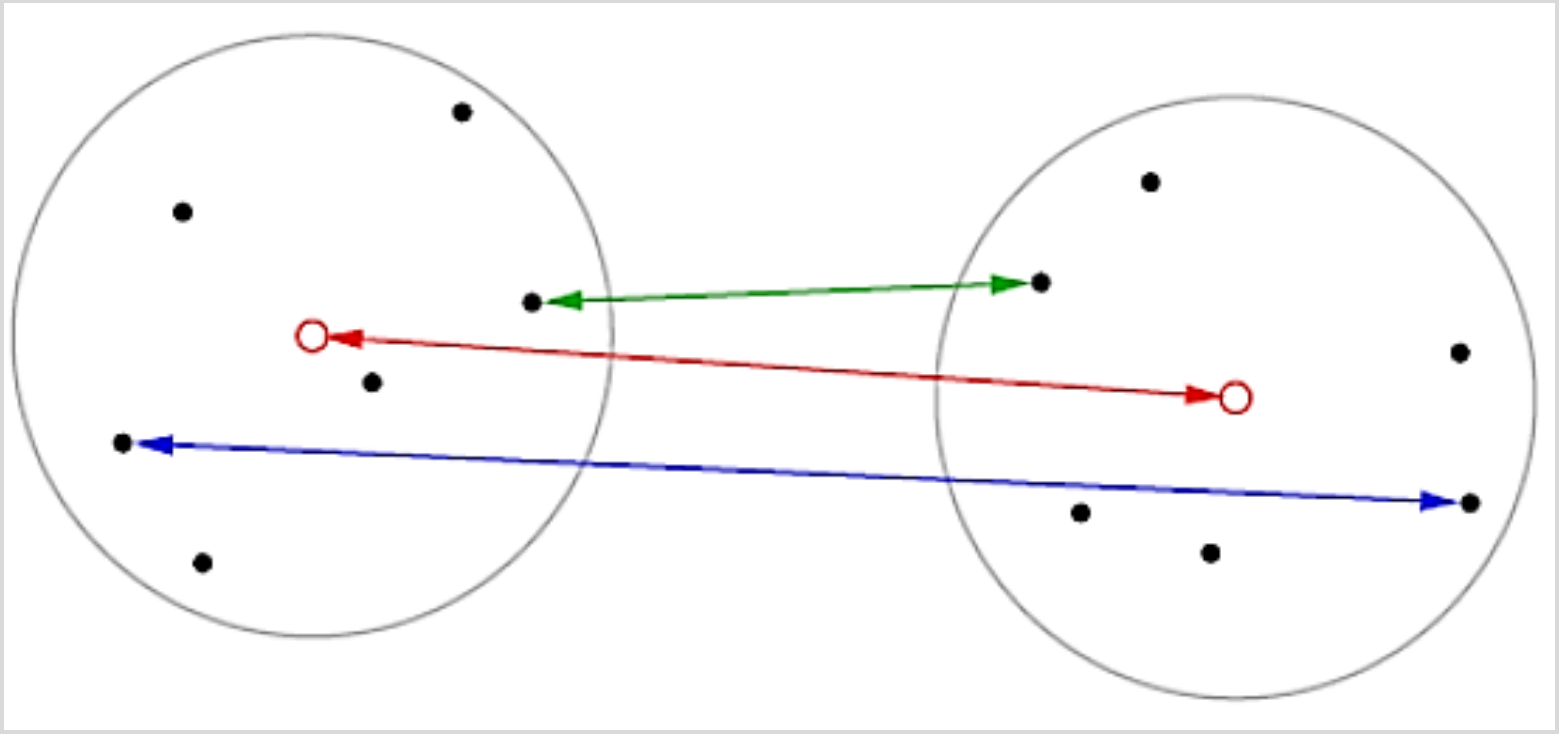
Adım 3: Aynı işlem gözlem sayısı kadar küme elde edilinceye kadar tekrar edilir.

Hiyerarşik birleştirici kümelemenin her aşamasında, Öklid veya başka bir uzaklık ölçütü kullanarak, birleştirilecek olan birbirine en benzer iki kümenin belirlenmesine farklı yaklaşımlar uygulanılabilir .

- **Tek bağlantı (single)**: ***İki küme (k1 ve k2) arasındaki uzaklık (U),*** k1 kümesi ile k2 kümesinin birbirine en yakın olan iki elemanı (x1 ve x2) arasındaki uzaklık olarak kabul edilir. En kısa mesafe esasına dayandığı için en yakın komşuluk tekniği olarak da bilinmektedir. ***Bu yöntem ile her adımda en küçük mesafeye sahip en yakın üyeler kullanarak iki küme birleştirilir.***

- **Tam bağlantı (complete)**: ***İki küme (k1 ve k2) arasındaki uzaklık (U),*** k1 kümesi ile k2 kümesinin birbirine en uzak olan iki elemanı (x1 ve x2) arasındaki uzaklık olarak kabul edilir. Kümeler arası eleman çiftleri arasındaki maksimum uzaklık dikkate alındığı için en uzak komşuluk tekniği olarak da bilinmektedir (Murtagh ve Contreras, 2017). ***Bu yöntem ile her adımda en büyük mesafeye sahip en uzak üyeler kullanılarak iki küme birleştirilir.***

- **Ortalama bağlantı (average)**: ***Birinci küme (k1) ile ikinci küme (k2) elemanları arasındaki bütün uzaklıklar hesaplanır ve bunların ortalaması iki küme arasında uzaklık (U) olarak kabul edilir.*** Başka bir deyişle, karşılıklı iki küme arasındaki tüm mesafelerin ortalamasıdır (Murtagh ve Contreras, 2017).



- **Merkez bağlantı (centroid)**: ***Bu yöntemde küme arasındaki uzaklık ölçüsü olarak Kareli Euclid (Squared Euclidean) uzaklığı kullanılmaktadır***. Her küme, o andaki kümenin ağırlık noktası ile temsil edilir. ***İki küme birleştiğinde, ağırlık noktalarının birbirlerinden minimal uzaklıkta olması yeterlidir.*** Bu yöntemin en önemli avantajı farklı nitelikteki gözlemlerden çok fazla etkilenmemesidir.

- **Ward yöntemi:** Bir kümenin merkezinde bulunan örneğin, kümenin içinde bulunan örneklerden ortalama uzaklığını dikkate alır. Yani, toplam küme içi varyansı minimize etmeyi hedefler. Bu amaçla, küme içi kareli sapmalardan yararlanarak hata kareler toplamını hesaplar (Murtagh ve Contreras, 2017). ***Bu yöntem ile minimum varyans değerine sahip olan kümeler birleştirilir. Ward yönteminin amacı, kümeleri, bu kümeler içindeki çeşitliliğin çok fazla artmaması için birleştirmektir. Bu, mümkün olduğunca homojen kümeler halinde yapılar oluşturur.***

Hiyerarşik vs Hiyerarşik Olmayan vs Karar Ağaçları

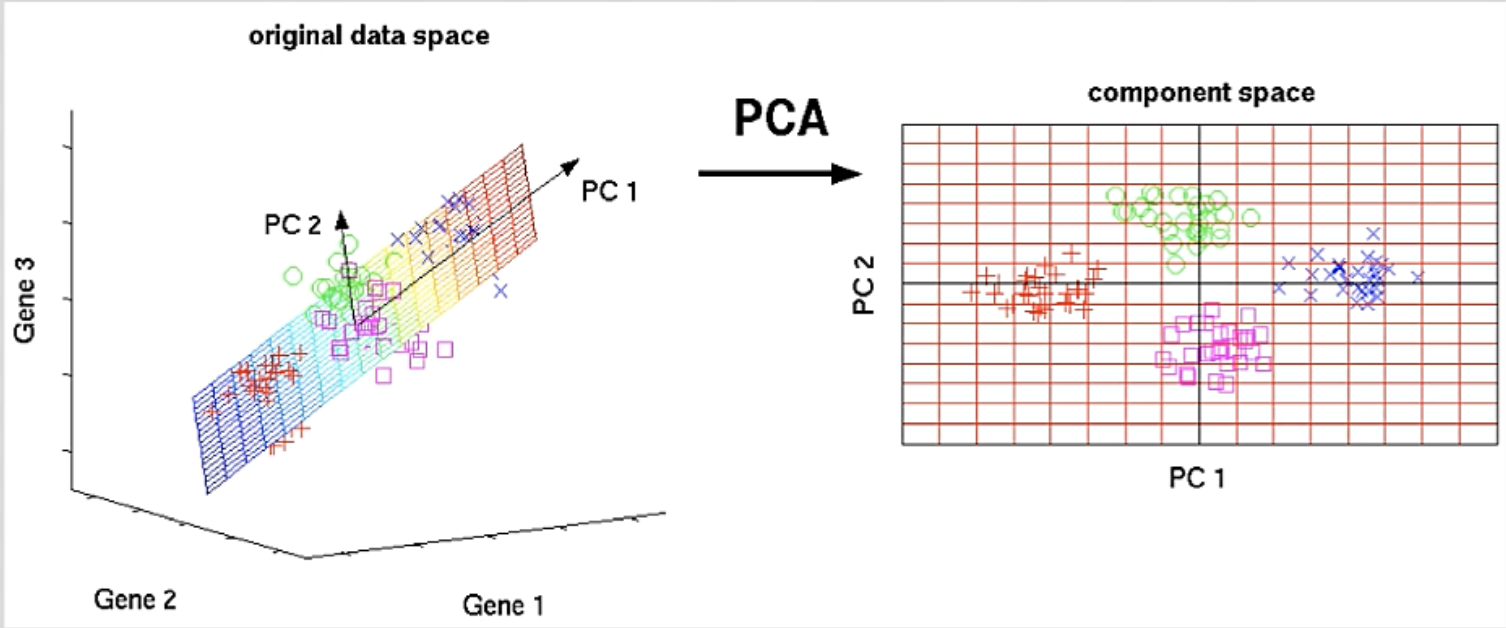
- Hiyerarşik yöntemlerde küme sayısına dendrogram sonuçlarına bakılarak karar verilirken, hiyerarşik olmayan yöntemlerde küme sayısı uygulama yapılmadan önce belirlenir.
- Hiyerarşik kümeleme yöntemlerinde veri seti gözlemler ya da değişkenler bazında kümeleme işlemine sokulabilirken hiyerarşik olmayan yöntemlerde sadece gözlemlerin kümelenmesi mümkündür.
- Karar ağaçlarından farkı; karar ağaçlarında ayırma işlemi hedef değişkene göre yapılırken burada bağımlı değişken olmadığı için gözlemler bağımsız değişkenler üzerinden yapılan uzaklık hesaplarına göre kümelere ayrılır.

Temel Bileşen Analizi (PCA)

Temel fikir, çok değişkenli verinin ana özelliklerini daha az sayıda değişken/bileşen ile temsil etmektir.

Diğer bir ifade ile: küçük miktarda bir bilgi kaybını göze alıp değişken boyutunu azaltmaktır.

Temel Bileşen Analizi (PCA)



- Örnekte 3 değişken 2 değişkene indirgenmiş oldu.
- Bu şekilde değişken boyutunu azaltmak için kullanılabilir.
- Oluşan değişkenler arasında korelasyon yoktur.

PCA'da değişken gruplarının varyanslarını ifade eden öz değerler ile veri setindeki değişkenler gruplandırılır. Gruplar arasında en fazla varyansa sahip gruplar en önemli gruplardır ki bunlar asal bileşenlerdir.