Unsupervised Learning

Bugün Ne Konuşacağız?

- Makine Öğrenmesi'nin temel konularından biri olan Unsupervised Learning'e giriş yapacağız
- Unsupervised Learning ile görselleştirme, kümeleme, özellik çıkarımı, boyut azaltma konularında kullanılan algoritmaları görecek ve bu algoritmaları derinlemesine inceleyeceğiz
- Kümeleme konusunda en çok kullanılan algoritmalardan biri olan K-means algoritmasını göreceğiz
- Boyut azaltma/görselleştirme konusunda en çok kullanılan algoritmalardan biri olan Principal Component Analysis (PCA) Algoritmasını adım adım görecek ve detaylı anlatım sağlayacağız

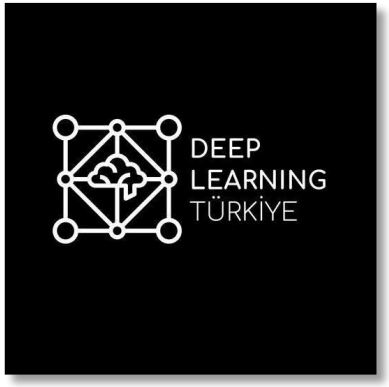
İrem Kömürcü

Github: irem-komurcu

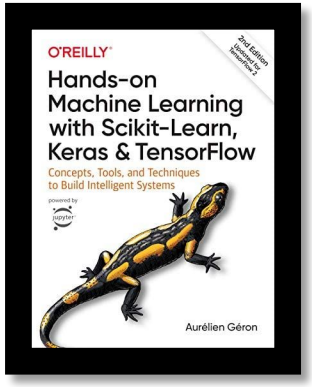
Twitter: iremkomurcu

LinkedIn: iremkomurcu

Kaynaklar



Deep Learning Türkiye

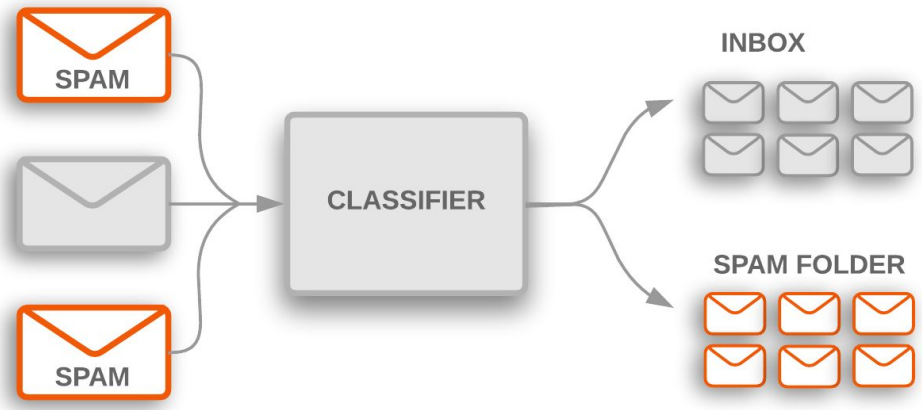


Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Aurelien Geron

Denetimli Öğrenme (Supervised Learning)

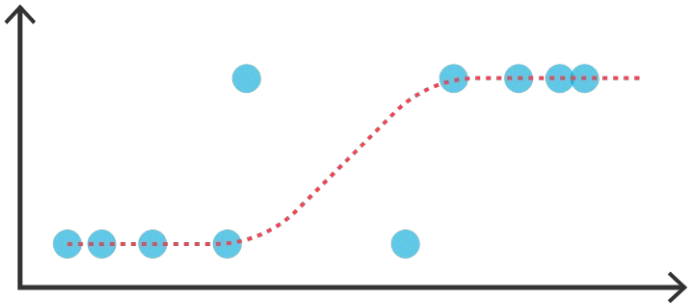
Supervised Learning; etiketlenmiş veri kümelerini bağımlı ve bağımsız değişkenler kullanarak işleme sokan makine öğrenmesi algoritmalarını içerir.

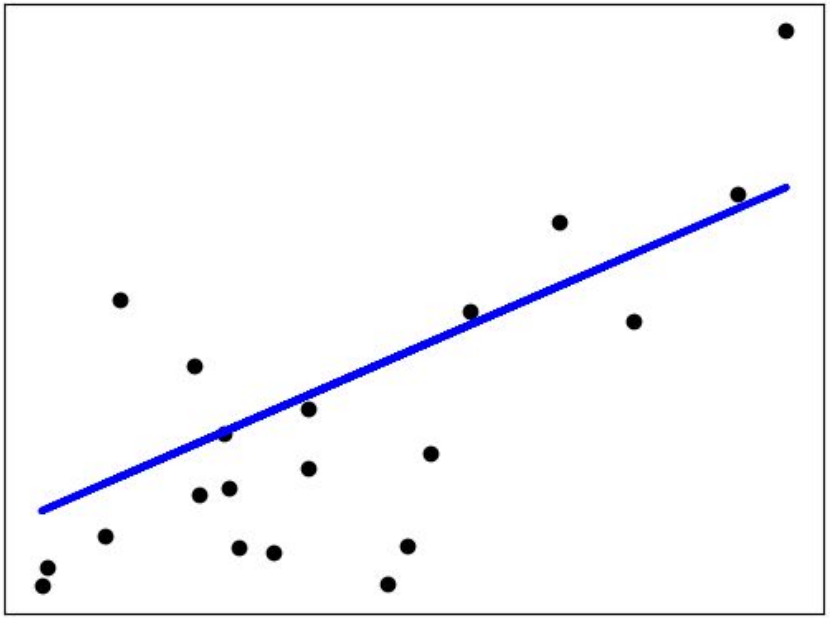
İstenmeyen e-posta filtresi buna iyi bir örnektir: Denetimli Öğrenme sınıflarıyla birlikte birçok örnek e-posta ile eğitilmiştir ve yeni e-postaları nasıl sınıflandıracağını öğrenmelidir.



Denetimli Öğrenme Algoritma ve Mimarileri

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural Networks

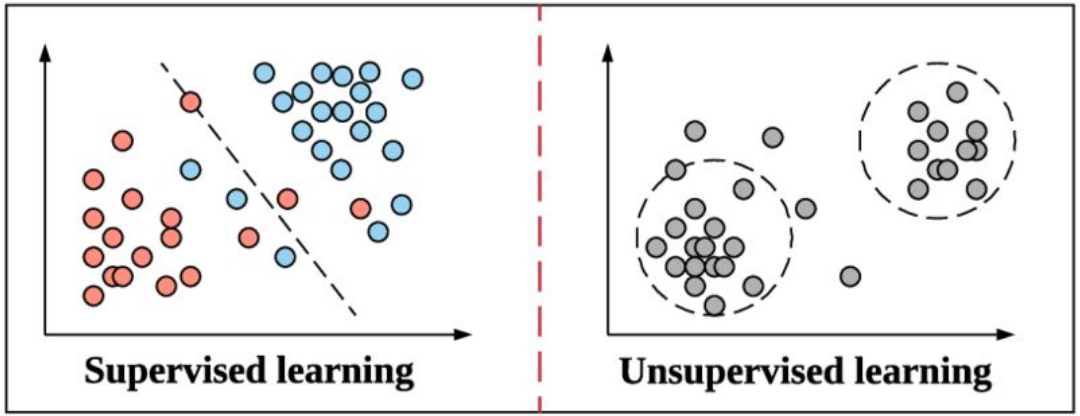




Denetimsiz Öğrenme (Unsupervised Learning)

Unsupervised Learning, etiketlenmemiş veri kümelerini analiz etmek ve kümelemek için makine öğrenimi algoritmalarını kullanır.

Bu algoritmalar, insan müdahalesine ihtiyaç duymadan gizli kalıpları veya veri gruplamalarını keşfeder.



Neden Unsupervised Learning Kullanıyoruz?

Unsupervised Learning için en yaygın görevler;

- Kümeleme
- Yoğunluk tahmini
- Temsili öğrenme

Unsupervised Learning Algoritmaları;

- Küme sayımızın bilinmediği
- Etiketli eğitim verimizin bulunmadığı

durumlarda kullanılabilir.

Örnekleri

Güvenlik:

Veri kümelerindeki olağandışı veri noktalarını tanımlandığı kümeleme anormalliği algılaması

Pazarlama:

Veri noktaları arasındaki ilişkileri bulduğu ilişki madenciliği

Unsupervised Learning Algoritmaları

Clustering

- Centroid-based
- Density-based Spatial Clustering (DBSCAN)
- Hierarchical Cluster Analysis (HCA)
- Affinity Propagation

Visualization and dimensionality reduction

- Principal Component Analysis (PCA)
- Kernel PCA
- Locally Linear Embedding (LLE)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

Anomaly Detection and Novelty Detection

- One-class SVM
- Isolation Forest

Association rule learning

- Apriori
- Eclat

Veriler Arasındaki Mesafeyi Hesaplama Yöntemleri

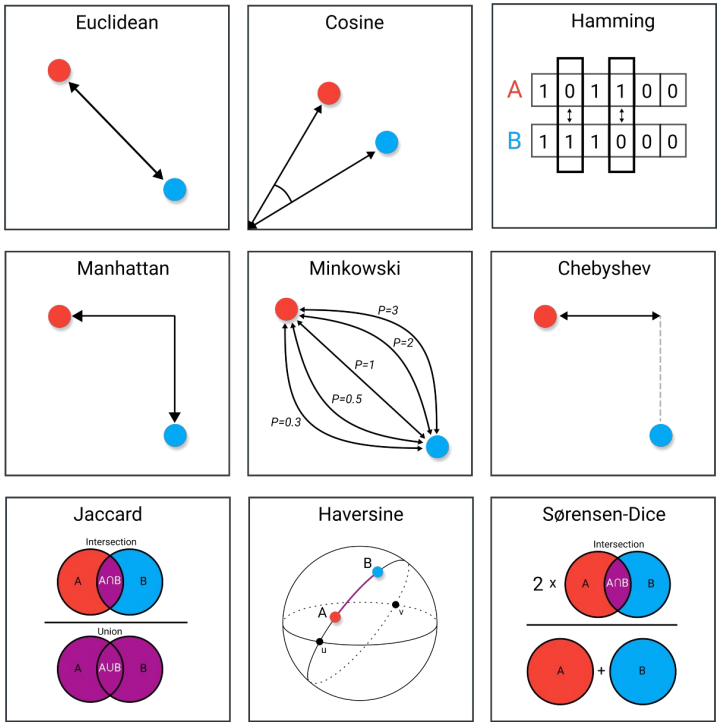
K-means kümelemede de olduğu gibi, birçok Makine Öğrenmesi algoritması uzaklık ölçümleri kullanır.

Veriler arasındaki mesafeleri ölçmek için çeşitli yöntemler vardır ve bu yöntemler veriye, probleme göre seçilmektedir. Aşağıda sık kullanılan bazı veriler arası uzaklık ölçme yöntemi verilmiştir;

Hamming Distance: İki ikili(binary) vektör arasındaki mesafeyi hesaplar

Euclidean Distance: İki gerçek değerli vektör arasındaki mesafeyi hesaplar

Manhattan Distance: Nesneleri tek bir ızgara üzerinde tanımlayan vektörler için kullanışlı bir uzaklık belirleme yöntemidir.



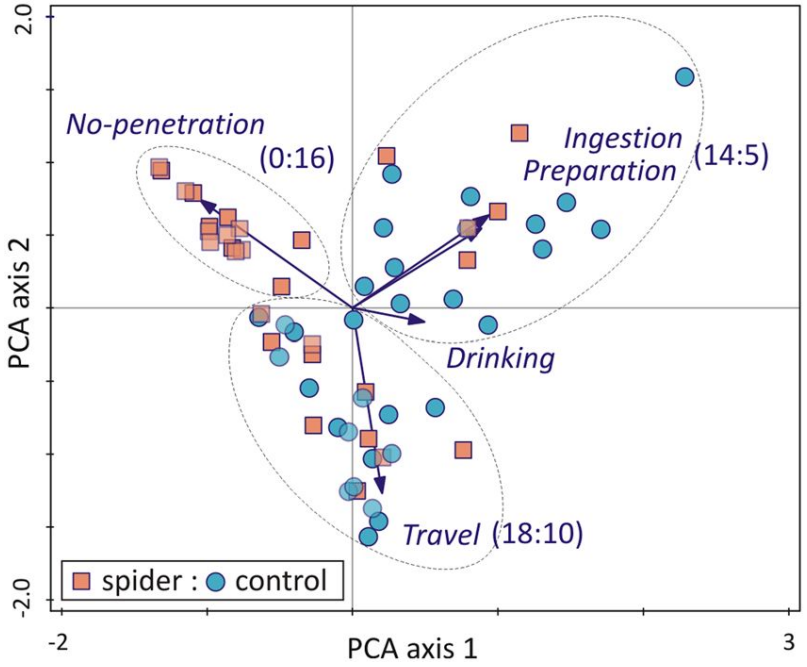
Görselleştirme ve Boyut Azaltma

Principal Component Analysis (PCA)

Temel Bileşen Analizi (PCA), karmaşık bir veri kümesinin daha düşük bir boyuta nasıl indirgeneceğine dair bir yol haritası sağlar.

PCA algoritması verilerdeki temel özellikleri yakalayarak daha az sayıda değişken ile göstermeye çalışır.

Büyük ve karmaşık veri kümelerinden ilgili bilgileri çıkartmak için boyut değiştirme, döndürme, boyut eğimi değiştirme gibi yöntemler kullanır.



Adım Adım PCA

Adım 1: Veri Kümesini standartlaştırın

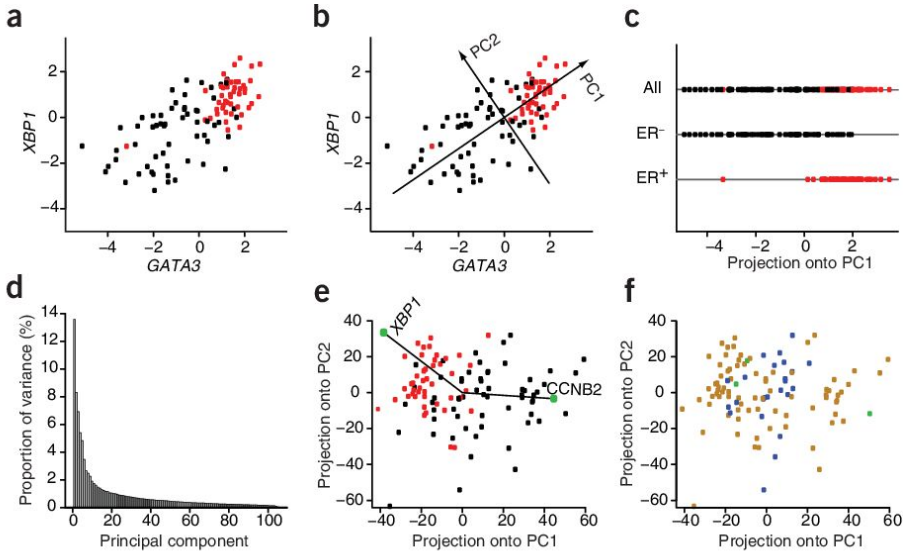
Adım 2: Veri kümesindeki özellikler için Covariance matrisini hesaplayın

Adım 3: Covariance matrisi için özdeğerleri ve özvektörleri hesaplayın

Adım 4: Özdeğerleri ve bunlara karşılık gelen özvektörleri hesaplayın

Adım 5: k özdeğerini seçin ve bir özvektör matrisi oluşturun

Adım 6: Orjinal matrisi dönüştürün



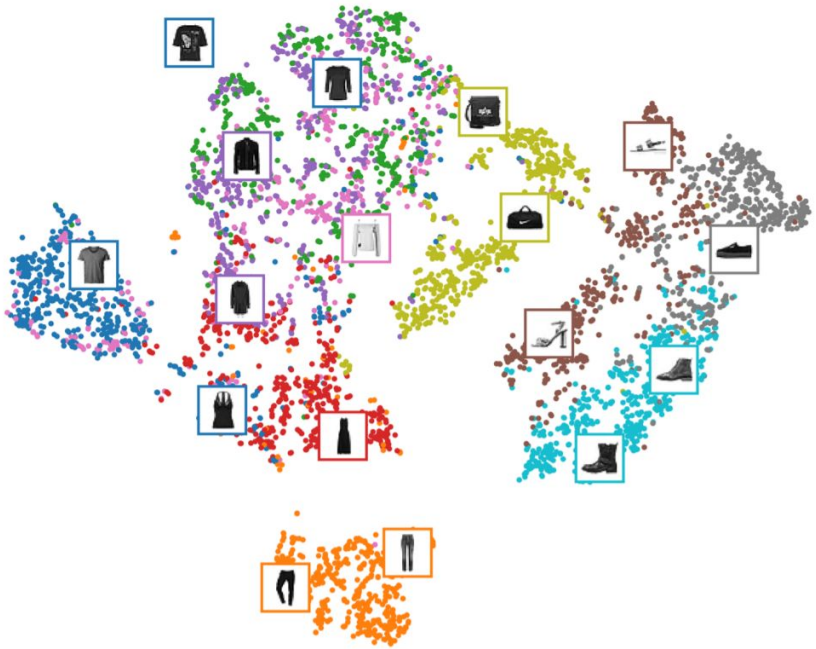
t-Distributed Stochastic Neighbor Embedding (t-SNE)

Benzer örnekleri yakın, farklı örnekleri ayrı tutmaya çalışırken boyutluluğu azaltır.

t-SNE algoritmasının ana fikri, noktalar arası uzaklıkları olabildiğince korumak ve düşük boyutlu bir temsil bulmaktır.

Çoğunlukla görselleştirme için, özellikle yüksek boyutlu uzayda örnek kümelerini görselleştirmek için kullanılır.

MNIST fashion görüntülerini 2D olarak görselleştirmek örneklerden biridir.



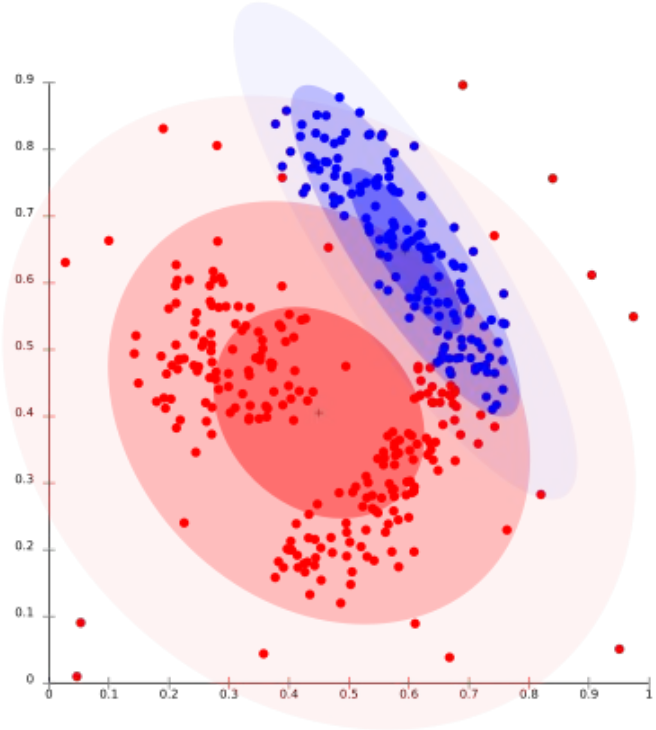
Kümeleme

Kümeleme (Clustering)

Verileri benzer olan gruplar halinde organize etme sürecidir.

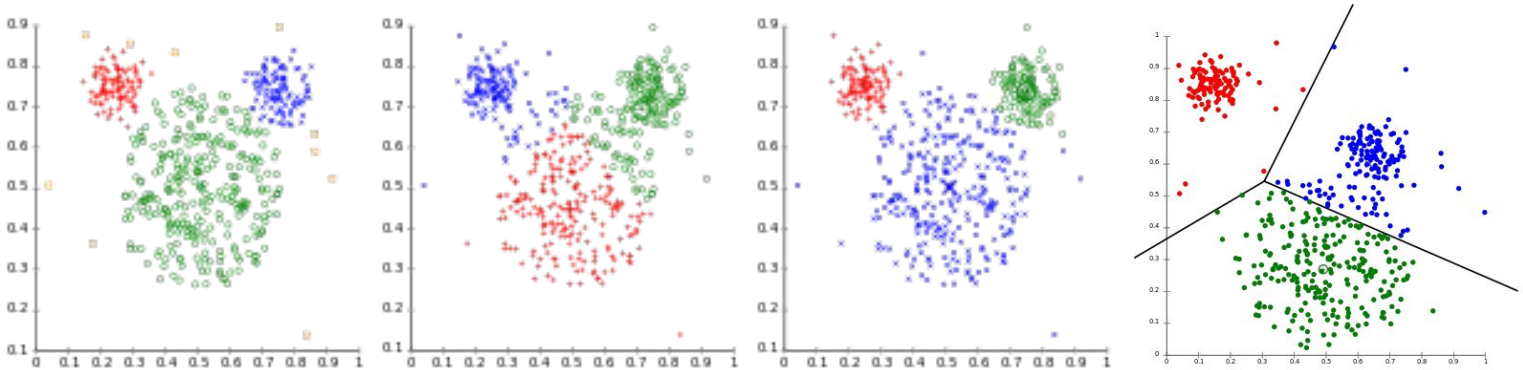
Kümeleme sayesinde etiketlenmemiş veriler tanımlanır ve aralarında benzerlik bularak gruplandırma yapılır.

Olası benzerliklere sahip nesneler bir grup oluştururken, benzerliği az ya da hiç bulunmayan veriler farklı gruplarda kalır.



Clustering Tipleri

- Affinity Propagation
- Hierarchical Cluster Analysis (HCA)
- Density-based Spatial Clustering (DBSCAN)
- Centroid-based



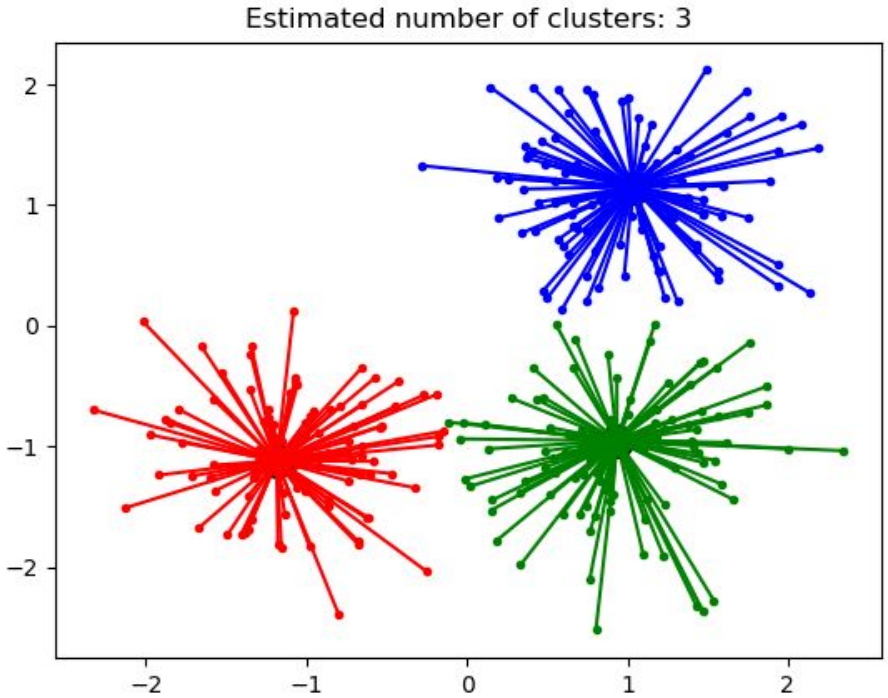
Affinity Propagation

Özellikle örnek ve küme sayımızın bilinmediği problemlerde kullandığımız bir kümeleme algoritmasıdır.

Her veri noktasını ağdaki bir düğüm olarak görür ve tüm veri noktalarını potansiyel örnekler olarak değerlendirir.

Veri noktaları arasındaki benzerliklere dayalı olarak çalışır.

Çok ve gereksiz sayıda küme bulmak, çok veri ile kötü çalışmak gibi sıkıntıları vardır.

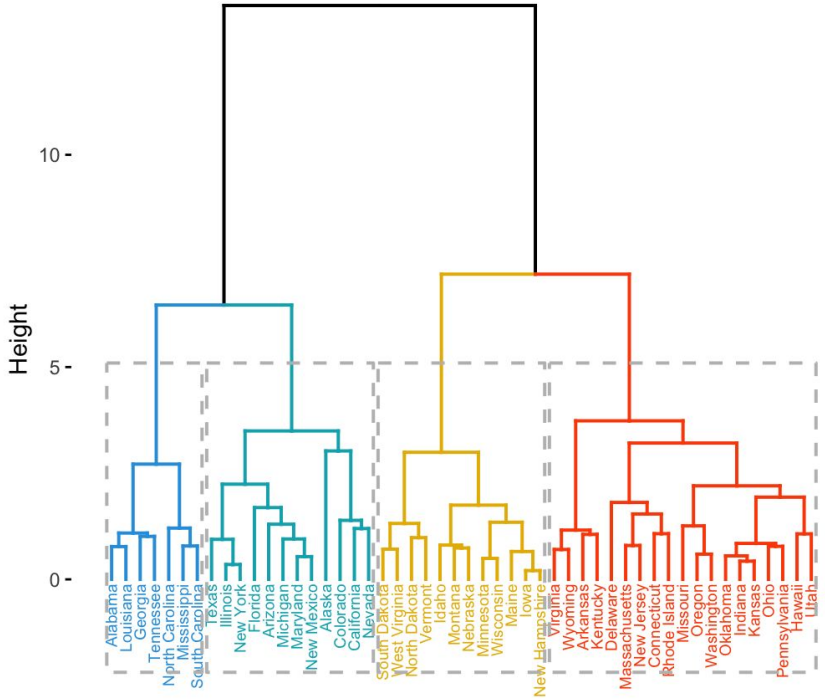


Hierarchical Clustering

Verileri benzerliklerine göre seviye ve hiyerarşilere ayırır.

Oluşturulan hiyerarşi sonucu Dendrogram adı verilen ağaç benzeri hiyerarşik yapılar oluşur.

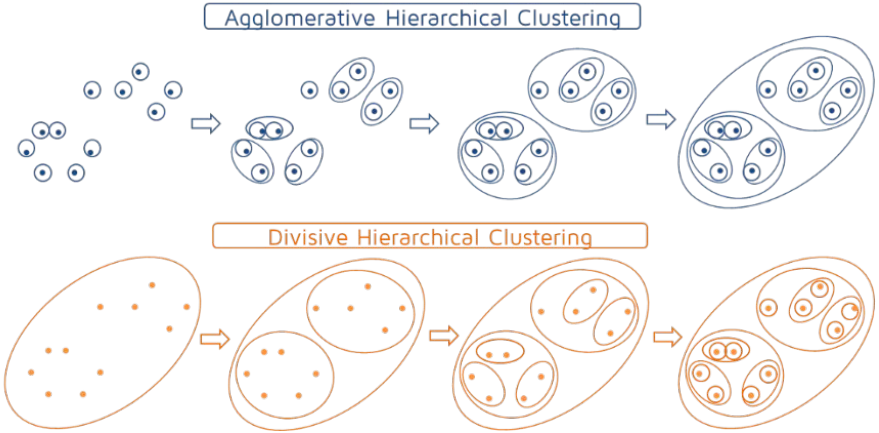
Oluşan Dendrogram yapısının kolları doğru seviyede kesilerek istenilen sayıda küme seçilebilir.



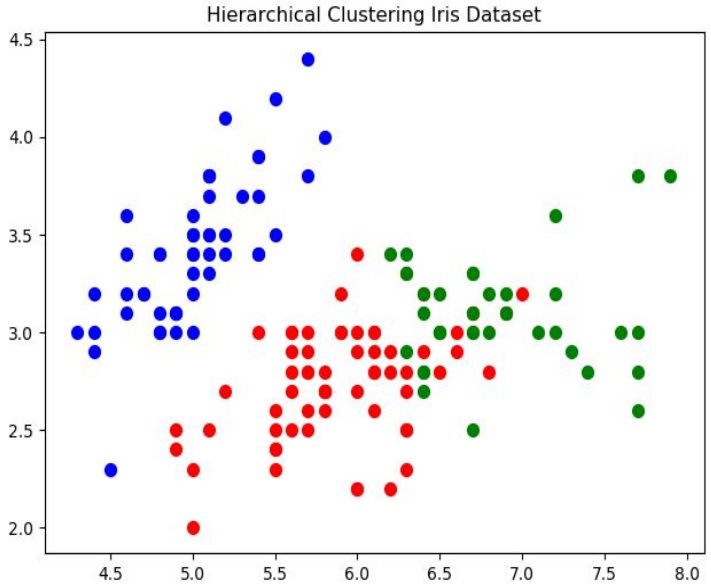
Hierarchical Clustering

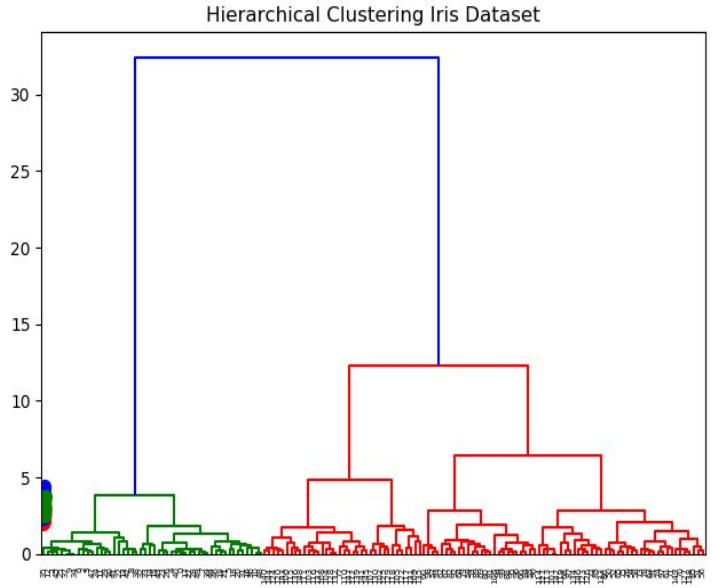
Hierarchical Clustering birbirine benzer mantıkta iki farklı çalışma yapısına sahiptir;

- **Agglomerative Clustering**; Parçadan bütüne prensibi ile çalışır. İlk başta tüm verileri birer küme olarak görür. Birbirine mesafe olarak yakın olan kümeler birleştirilerek kümeler oluşturulur
- **Divisive Clustering;** Bütünden parçaya prensibi ile çalışır. Tüm veriler tek bir küme oluşturur. Ardından küme içinde yakınlığa bağlı yeni kümeleme işlemi yapılır



Hierarchical Clustering and Dendrogram





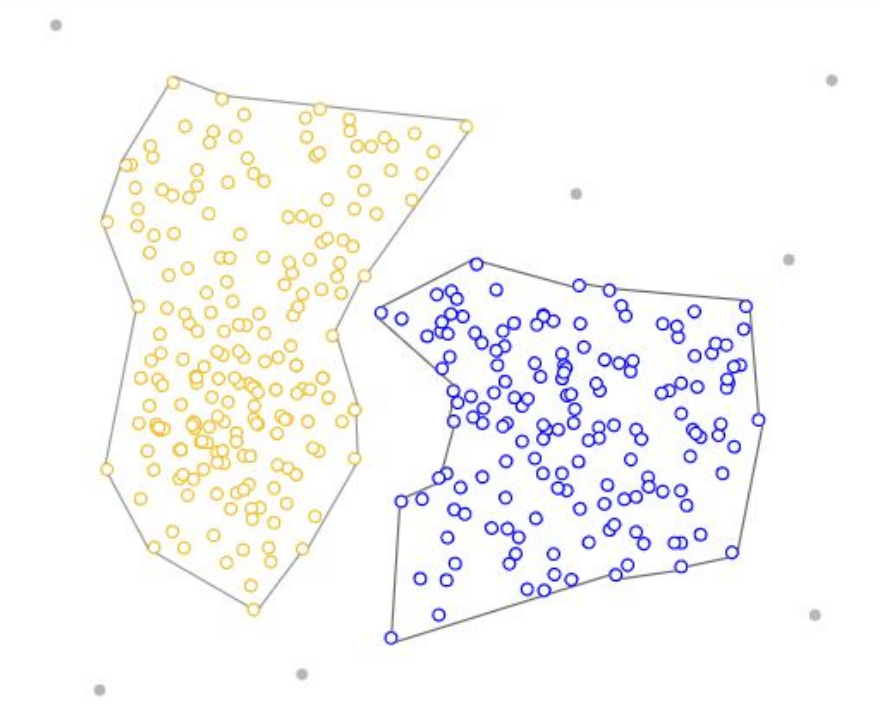
Hierarchical Clustering; veri noktalarını benzerliklerine göre seviyelere / hiyerarşilere ayırır. Kümeleme sonrası, bu hiyerarşi Dendrogram adı verilen ağaç benzeri bir yapı oluşturur.

Density-based Clustering

Verilerin Gauss dağılımları gibi dağılımlardan oluştuğunu varsayar ve yoğunluğa dayalı kümeleme işlemi yapar.

Dağılımın merkezine olan uzaklık arttıkça, bir noktanın dağılıma ait olma olasılığı azalır.

Küme sayısı gerektirmez, verilerdeki yoğunluğa dayalı küme sayısını çıkartır.

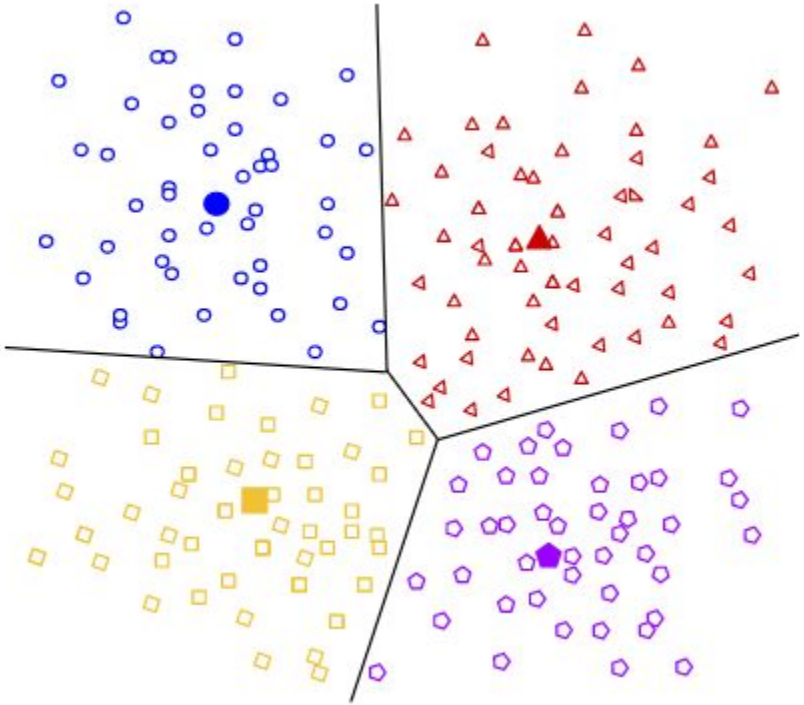


Centroid-based Clustering

Centroid Based Clustering, verileri n sayıda centroid yardımıyla n sayıda kümeye böler.

k-means, en yaygın kullanılan centroid tabanlı kümeleme algoritmasıdır.

Centroid-based kümeleme başlangıç koşullarına ve aykırı değerlere karşı hassastır.



K-Means Clustering

K-Means Clustering

K-Means algoritması, örnekleri eşit varyanslı n grupta ayırmaya çalışır.

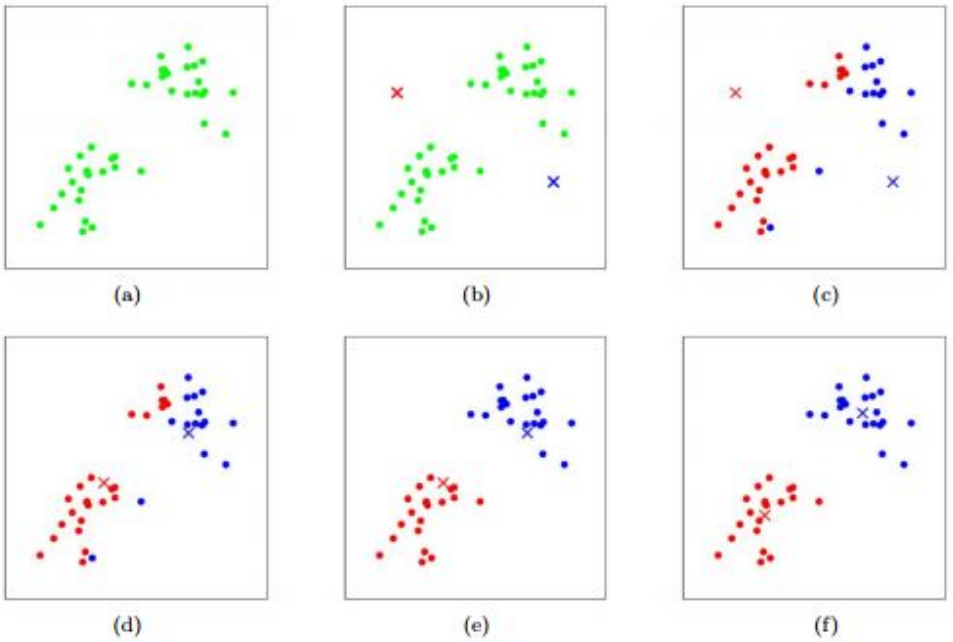
Bu algoritma, küme sayısının belirtilmesini gerektirir. Çok sayıda örneğe iyi ölçeklenir ve birçok farklı alanda çok çeşitli uygulama alanlarında kullanılmıştır.

K-means algoritması, karesi alınmış hata fonksiyonu olarak bilinen bir amaç fonksiyonunu en aza indirmeyi amaçlar:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - m_i \right|^2$$

K-Means Clustering

- Küme merkezi (centroid), kümeye ait tüm noktaların aritmetik ortalamasıdır
- Her nokta kendi küme merkezine diğer küme merkezlerinden daha yakındır



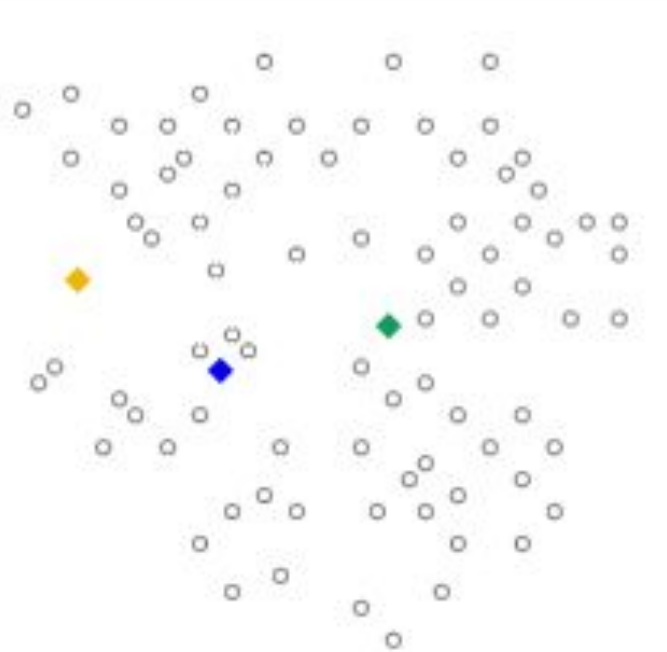
Adım Adım K-Means Clustering

Adım 1:

Centroidlerin rastgele atanması

Algoritma, her küme için rastgele bir centroid seçer.

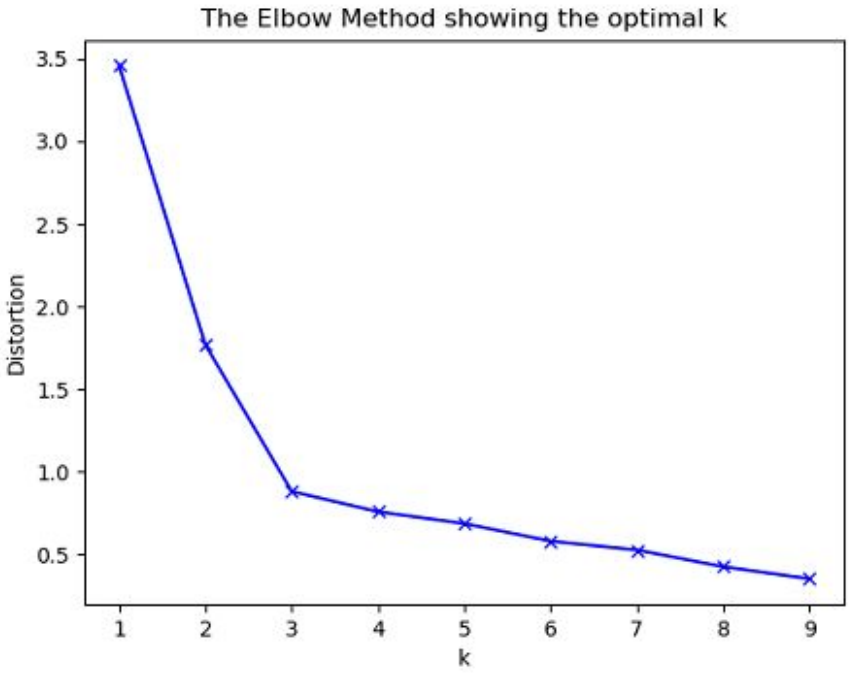
Örneğimizde, k değeri 3 seçildiği için algoritma 3 centroidi dağılıma rastgele yerleştiriyor.



Optimum k Değerini Belirlemek: Elbow Method

K-Means algoritması kullanılırken küme sayısını belirlemek için kullanılan bir yöntemdir.

- Farklı k değerleri için çizilen cost fonksiyonunun değer grafiğini ifade eder
- Çizilen grafikte dirsek noktası alınır
- Bu nokta, genellikle optimum küme sayısına yakındır



Silhouette Score

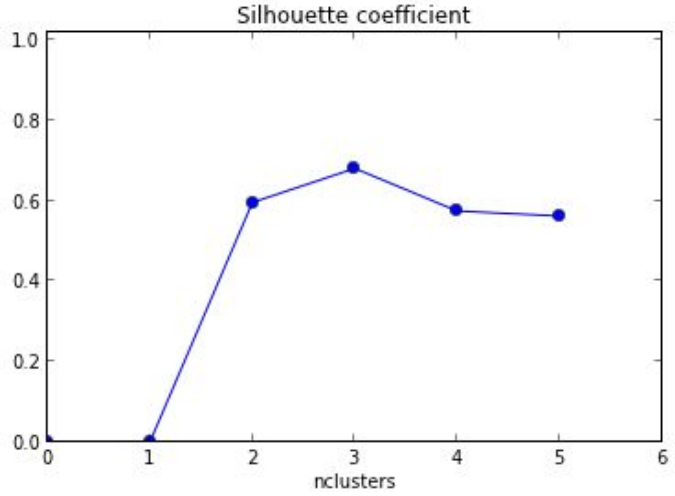
Silhouette Score, Katsayı veya Siluet puanı;

Bir kümeleme tekniğinin ne kadar iyi kümelediğini hesaplamak için kullanılan bir ölçüdür. Değeri -1 ile 1 arasında değişir.

1: Kümelerin birbirinden oldukça uzak olduğu ve açıkça ayırt edildiği anlamına gelir.

0: Kümelerin birbirine olan mesafenin azaldığı veya kümeler arasındaki mesafenin önemli olmadığı anlamına gelir.

-1: Kümelerin yanlış şekilde atandığı anlamına gelir.



$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

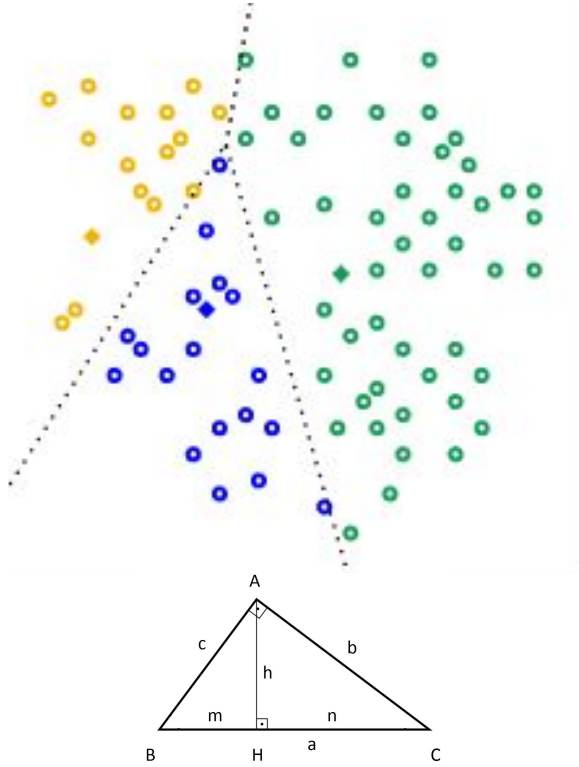
Adım Adım K-Means

Adım 2:

İlk Kümelerin Oluşması

Algoritma, ilk k (cluster sayısı) tane küme elde etmek için her noktayı en yakın centroid noktasının kümesine dahil eder.

Centroidlere atama konusunda mesafe göz önüne alınır ve kullanılan en yaygın yöntemlerden biri Öklid formülüdür.



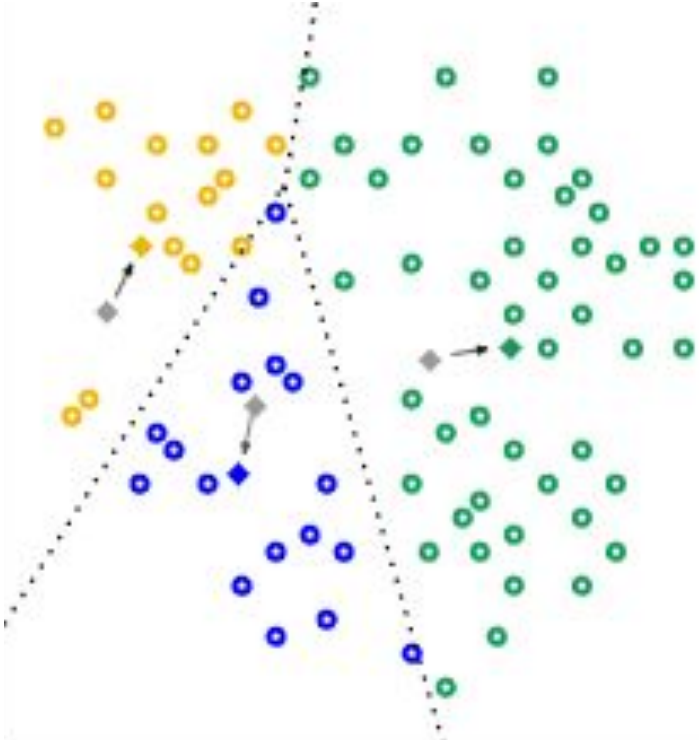
Adım Adım K-Means

Adım 3:

Ağırlık Merkezinin Yeniden Hesaplanması

Algoritma, her küme için kümedeki tüm noktaların ortalamasını alarak ağırlık merkezini yeniden hesaplar.

Merkezlerdeki değişiklikler şekilde oklarla gösterilmiştir. Merkezler değiştiğinden, algoritma, noktaları en yakın merkeze yeniden atar.

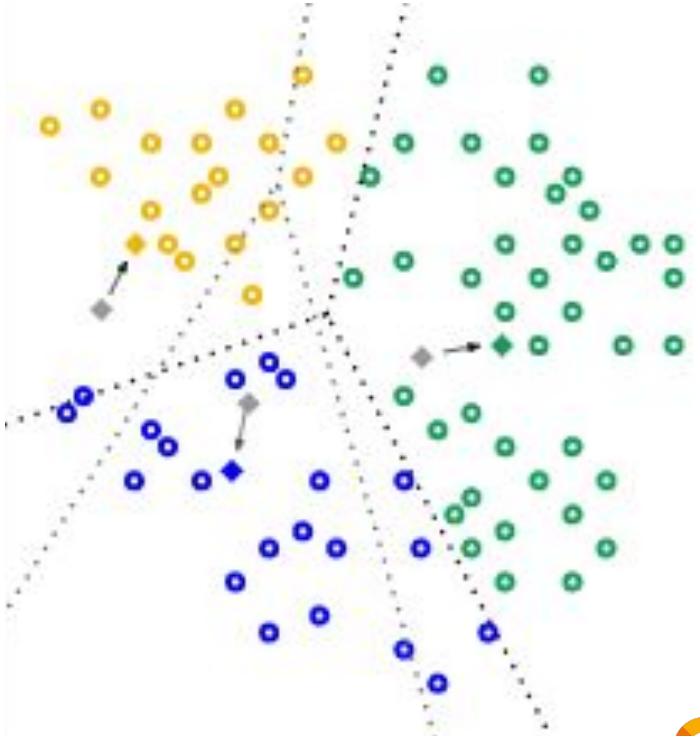


Adım Adım K-Means

Adım 4:

Yeniden Atamalar ve Kümelendirmeler

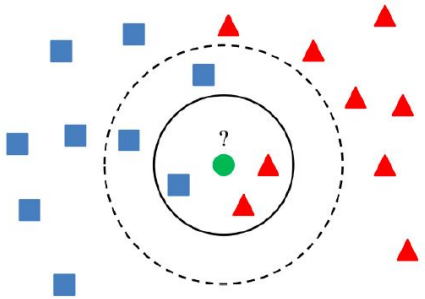
Algoritma, noktalar kümeleri değiştirmeyi durdurana kadar merkez noktalarının hesaplanmasını ve noktaların atanmasını tekrarlar. Büyük veri kümelerini kümelerken, bunun yerine diğer kriterleri kullanarak yakınsamaya ulaşmadan önce algoritmayı durdurursunuz.



K-NN ve K-Means Algoritmaları

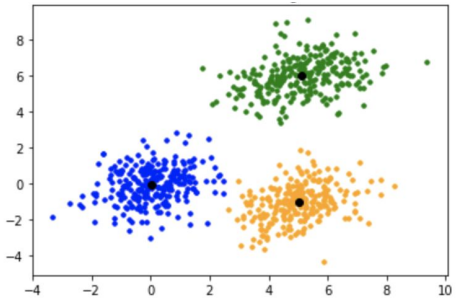
K-Nearest Neighbors Algoritması (k-NN);

- Sınıflandırma ve Regresyon problemlerinde kullanılan bir Denetimli Öğrenme Algoritması
- Temelde özellik benzerliğine dayanır ve buna bağlı işlemler yapar
- Bir nesnenin birden fazla sınıfı olabilir



K-Means Algoritması;

- Kümeleme problemlerinde kullanılan Denetimsiz Öğrenme Algoritması
- Etiketten bağımsız olarak yakınlık, ortalama gibi özelliklere bağlı işlemler yapar
- Bir veri tek bir kümeye aittir



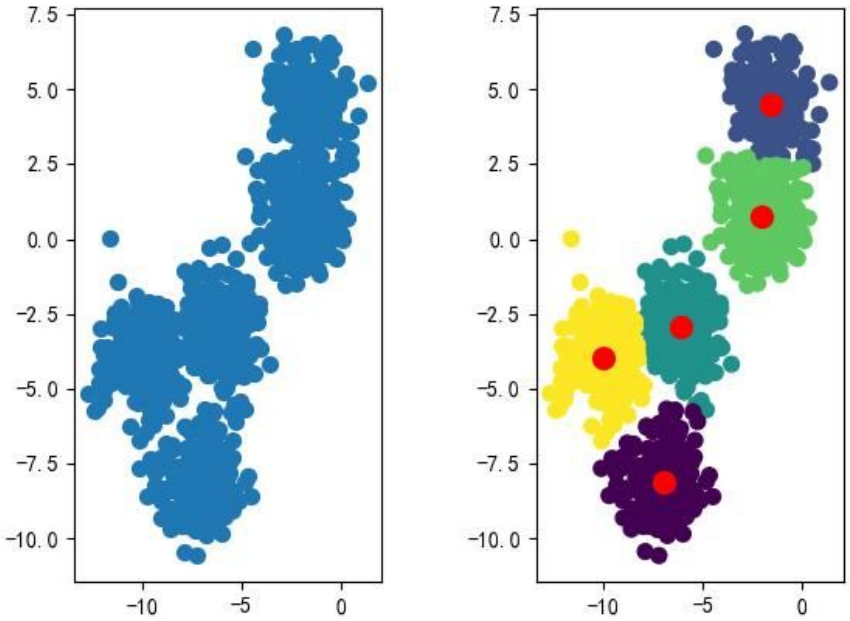
Mini Batch K-Means

Mini Batch K-Means Clustering

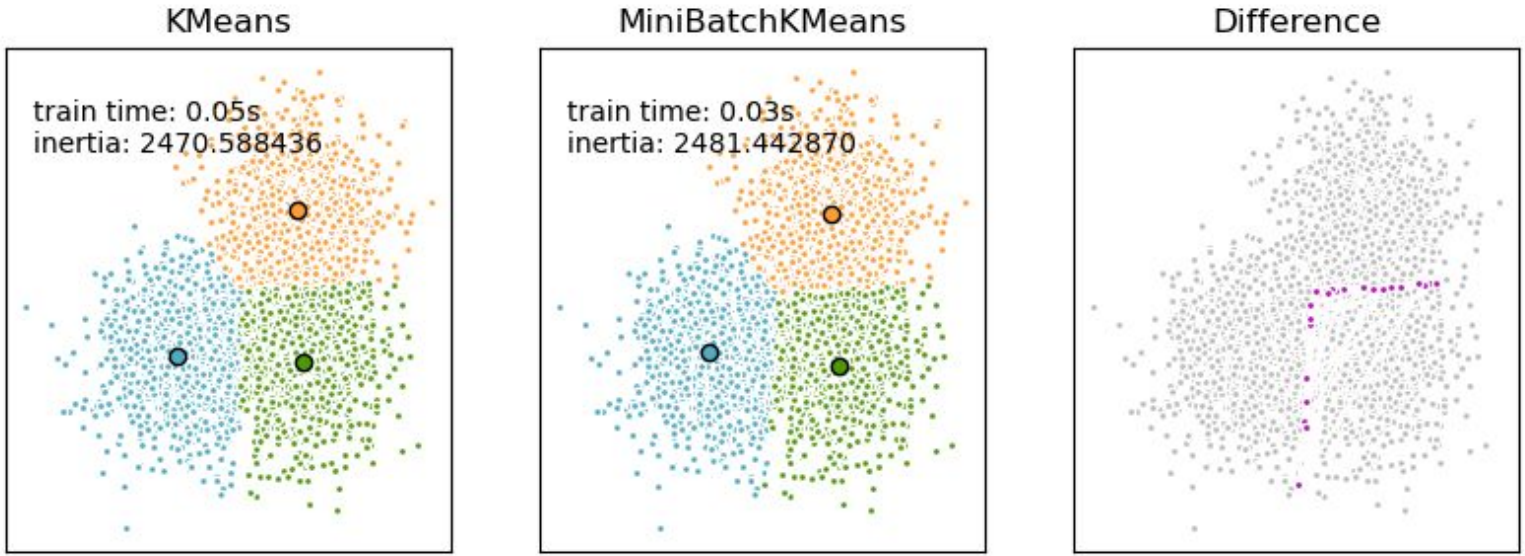
Hesaplama süresini azaltmak için mini-batchler kullanır ve K-Means algoritmasının bir çeşididir.

Mini Batchler, her eğitim yinelemesinde rastgele seçilmiş küçük yığınları alır.

Mini Batch içindeki her veri, küme merkezinin önceki konumlarına bağlı olarak kümelere atanır.



K-Means Clustering ve Mini Batch K-Means Clustering



Mini Batch K-Means daha hızlıdır ancak K-Means algoritmasına oranla biraz farklı sonuçlar verir.