K-Means Clustering Algoritması

Atıl Samancıoğlu

1 Giriş

K-Means, gözetimsiz (*unsupervised*) öğrenmenin en bilinen algoritmalarından biridir. Amacı, benzer veri noktalarını gruplara ayırmak yani **clustering** (kümeleme) yapmaktır.

Bu algoritma, veri setini k adet kümeye ayırır. Her küme bir **centroid** (merkez) etrafında şekillenir. K-means'in ismi de buradan gelir: *k tane ortalama merkez*.

2 Geometrik Sezgi

İki boyutlu bir düzlemde veri noktalarınız olduğunu düşünün. Şekil olarak şöyle olabilir:

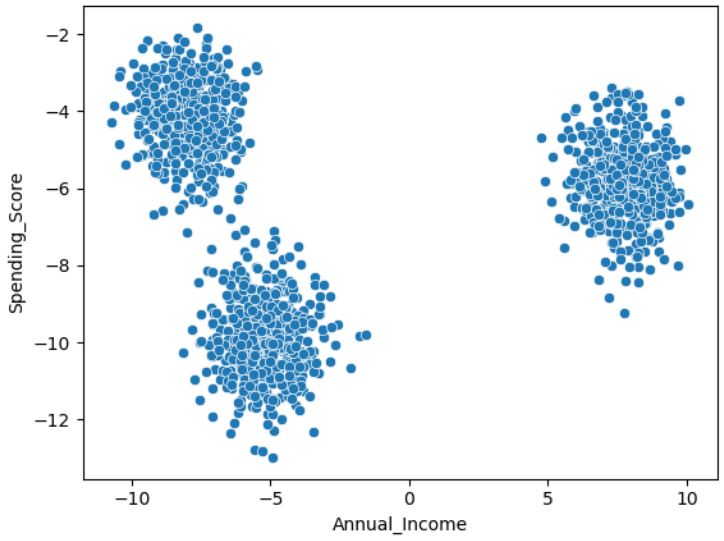

Figure 1: Kümelenmemiş veri noktaları

Bu noktaların aslında üç grup halinde toplandığını gözle görebiliyoruz. K-Means uygulandıktan sonra veri bu şekilde kümelenmiş olur:

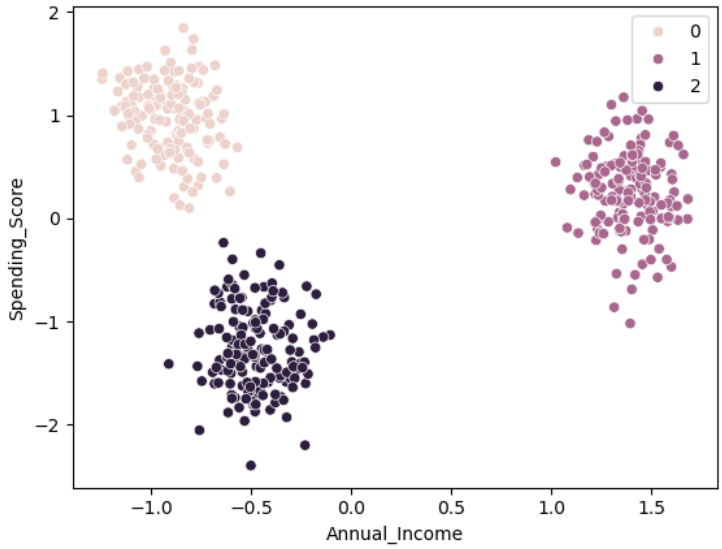

Figure 2: K-means sonrası kümelenmiş veriler ve centroid'ler

Kırmızı ve mavi renkli noktalar farklı kümeleri, büyük noktalar da bu kümelerin merkezlerini (centroid) temsil eder.

3 Algoritma Adımları

K-Means üç temel adımı tekrar eder:

1. Centroid'lerin Başlatılması

İlk adımda k tane centroid rastgele oluşturulur. Örneğin:

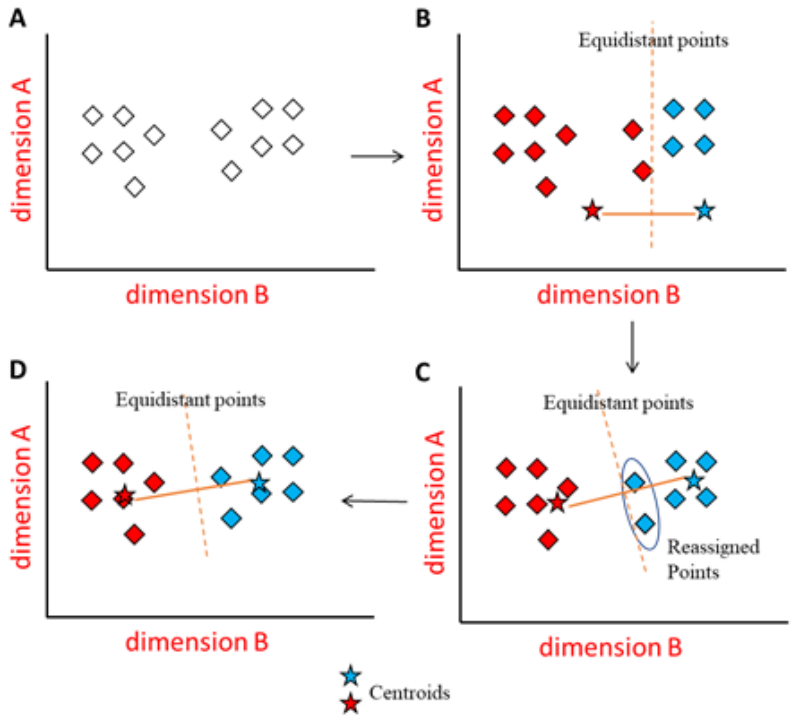

Figure 3: Oxford Protein Grup Kmeans Visual

2. Her Noktayı En Yakın Centroid'e Ata

Burada genellikle **Euclidean mesafesi** (öklidyen uzaklık) kullanılır:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Daha önce KNN algoritmasında gördüğümüz gibi **Manhattan mesafesi** de kullanılabilir:

$$d = |x_2 - x_1| + |y_2 - y_1|$$

3. Centroid'leri Güncelle

Her kümedeki noktaların ortalaması alınarak yeni centroid hesaplanır. Bu işlem tüm noktalar yeniden atanana veya centroid'ler değişmeyene kadar tekrar edilir. Tüm noktalar en yakın centroid'e atanır. Yeni centroid'ler hesaplanır ve adımlar tekrar edilir.

4 K Değeri Nasıl Seçilir?

Doğru k değeri için **Elbow Method** kullanılır. Bu yöntemde farklı k değerleri için **WCSS** (Within-Cluster Sum of Squares) hesaplanır:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Buradaki μ_i her bir kümenin centroid'i, x ise o kümedeki bir veri noktasıdır.

Buradaki WCSS değeri şu şekilde hesaplanır: K sayısı 1 seçilirse, 1 tane centroid'ten tüm noktalara uzaklık hesaplanır. Tek bir centroid olduğu için bu sayı büyük olacaktır. Sonrasında 2 centroid için hesaplanır. Mantıken bazı noktalar 2. centroid'e daha yakın olacağı için toplam uzaklık, yani WCSS, daha düşük olacaktır. Bu centroid sayısını arttırdıkça düşecektir. Fakat bir süre sonra (ideal grup sayısını geçtikten sonra) bu azalma daha az miktarda olmaya başlayacaktır. İşte bu kırılım noktası bizim optimal sayımızdır.

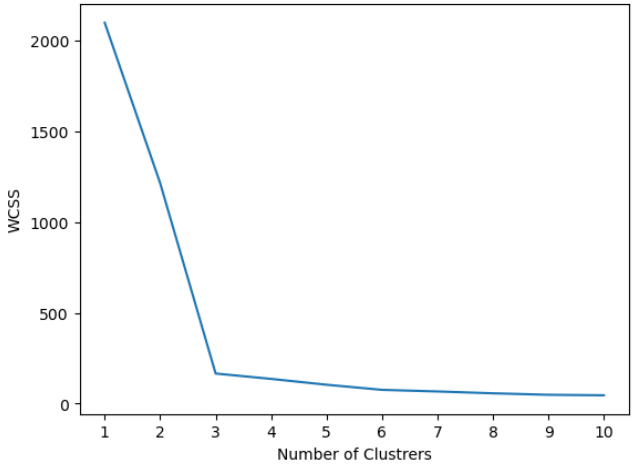

Figure 4: Elbow Yöntemi ile optimal k seçimi

Grafikteki "dirsek" noktası optimal k değeridir.

5 Random Initialization Trap ve K-Means++

Centroid'lerin rastgele başlatılması sonucu bazı durumlarda kötü kümeler elde edilebilir. Bu duruma **Random Initialization Trap** denir.

K-Means++, centroid'leri birbirinden olabildiğince uzak başlatır. Böylece algoritmanın daha hızlı ve doğru sonuç vermesi sağlanır. Rastgele seçimlerde olduğundan daha az veya daha fazla küme elde etmek bazı veri setleri için kaçınılmaz olur. K-Means++ ise bu sorunu en ideal şekilde çözmeye çalışır.

6 Kapanış

K-Means, basit ve güçlü bir kümeleme algoritmasıdır. En önemli avantajı:

- Kolay uygulanabilir,
- Büyük veri setlerinde hızlı çalışır.

Ancak dezavantajları da vardır:

- Küme sayısı (k) önceden verilmelidir,

Yine de çoğu uygulamada ilk denenecek yöntemlerden biridir.