

K-En Yakın Komşu Algoritması (K-Nearest Neighbour Algorithm)

K-En Yakın Komşu (KNN), hem sınıflandırma hem de regresyon problemlerinde kullanılan basit ve sezgisel bir makine öğrenimi algoritmasıdır. Temel prensip, yeni bir veri noktasını sınıflandırmak veya tahmin etmek için eğitim veri setindeki en yakın K komşusuna bakmaktır.

Temel Adımlar ve Prensipler

1. K Değerinin Seçimi:

- KNN algoritmasında, K değeri kaç komşunun dikkate alınacağını belirler. K değeri genellikle tek sayıdır (örneğin, K=3, K=5).

2. Mesafe Hesaplama:

- Yeni veri noktası ile tüm eğitim veri noktaları arasındaki mesafeler hesaplanır. Genellikle Öklidyen mesafe kullanılır.
- Öklidyen Mesafe Formülü:

$$d(p,q)=(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2$$

- Manhattan mesafesi ve Minkowski mesafesi gibi diğer mesafe ölçümleri de kullanılabilir.

3. En Yakın K Komşunun Seçimi:

- Hesaplanan mesafelere göre en yakın K komşu seçilir.

4. Sınıflandırma veya Tahmin:

- Sınıflandırma:** K komşunun sınıf etiketlerine bakılır ve çoğunluk sınıf yeni veri noktasının sınıfı olarak atanır.
- Regresyon:** K komşunun değerlerinin ortalaması alınır ve bu ortalama yeni veri noktasının tahmini değeri olarak atanır.

Örnek 1: Sınıflandırma Problemi

Bir veri seti düşünelim: Elma ve portakal resimlerini içeren bir meyve veri setimiz var ve bu meyveleri boyut ve renk gibi özelliklerle sınıflandırmak istiyoruz.

1. Adım 1 - Eğitim Verileri:

Meyve	Renk Yoğunluğu	Boyut	Sınıf
Elma	5	7	Elma
Elma	6	6	Elma
Portakal	3	8	Portakal
Portakal	2	9	Portakal
Elma	4	6	Elma

- Elma ve portakal resimlerinden oluşan veri setinde her resmin özellikleri belirlenmiş olsun (örneğin, renk yoğunluğu ve boyut).

2. Adım 2 - Yeni Veri Noktası:

- Renk yoğunluğu 4 ve boyut 7 olan yeni bir meyve verimiz var. Bu meyvenin elma mı yoksa portakal mı olduğunu belirlemek istiyoruz.

3. Adım 3 - Mesafe Hesaplama:

- Yeni meyve ile eğitim veri setindeki tüm meyveler arasındaki Öklidyen mesafeleri hesaplıyoruz.

Meyve	Mesafe
Elma	1.41
Elma	2.24
Portakal	1.41
Portakal	2.83
Elma	1.00

4. Adım 4 - En Yakın K Komşunun Seçimi:

- K=3 olarak seçildiğinde en yakın 3 komşu:
 - Elma (1.00)
 - Elma (1.41)
 - Portakal (1.41)

5. Adım 5 - Sınıflandırma:

- K=3 komşunun çoğunluk sınıfı "Elma" olduğu için, yeni meyve "Elma" olarak sınıflandırılır.

Örnek 2: Regresyon Problemi

Bir veri seti düşünelim: Bir evin fiyatını tahmin etmek istiyoruz ve elimizde evin metrekaresi ve odası sayısı gibi özellikler var.

1. Adım 1 - Eğitim Verileri:

- Elimizde evlerin metrekaresi, oda sayısı ve fiyat bilgisi var.

Metrekare	Oda Sayısı	Fiyat
100	3	200000
150	4	250000
200	5	300000
120	3	220000
180	4	270000

2. Adım 2 - Yeni Veri Noktası:

- 160 metrekare ve 4 odalı yeni bir evin fiyatını tahmin etmek istiyoruz.

3. Adım 3 - Mesafe Hesaplama:

- Yeni ev ile eğitim veri setindeki evler arasındaki Öklidyen mesafeleri hesaplıyoruz.

Metrekare	Oda Sayısı	Mesafe
100	3	60.83
150	4	10.00
200	5	44.72
120	3	40.00
180	4	20.00

4. Adım 4 - En Yakın K Komşunun Seçimi:

- K=3 olarak seçildiğinde en yakın 3 komşu:
 - 150 metrekare, 4 odalı ev (10.00)
 - 180 metrekare, 4 odalı ev (20.00)
 - 120 metrekare, 3 odalı ev (40.00)

5. Adım 5 - Regresyon (Tahmin):

- Bu üç komşunun fiyatlarının ortalaması alınır:

Tahmin Edilen Fiyat=250000+270000+220000/3=246666.67
Tahmin Edilen Fiyat=3250000+270000+220000=246666.67

- Yeni evin tahmin edilen fiyatı yaklaşık 246666.67 TL olur.

KNN Algoritmasının Avantajları ve Dezavantajları

Avantajları:

- **Basit ve Kolay Anlaşılır:** Algoritmanın mantığı oldukça sezgiseldir.
- **Parametrik Olmayan:** Veri dağılımı hakkında önceden bilgi gerektirmez.
- **Etkili:** Küçük ve orta ölçekli veri setlerinde iyi performans gösterir.

Dezavantajları:

- **Hesaplama Maliyeti:** Büyük veri setlerinde mesafe hesaplamaları yavaş olabilir.
- **Hafıza Gereksinimi:** Tüm veri setini saklaması gerektiği için hafıza açısından maliyetlidir.
- **K ve Mesafe Metriği Seçimi:** Doğru K değeri ve mesafe metriğini seçmek zordur ve sonuçları büyük ölçüde etkileyebilir.

Bu örnekler ve açıklamalarla, K-En Yakın Komşu algoritmasının nasıl çalıştığını ve nasıl kullanıldığını daha iyi anlayabilirsiniz. KNN, özellikle veri seti küçük olduğunda ve hesaplama gücü sınırlı olmadığında oldukça etkili ve basit bir yöntemdir.