Udemy - (50 Saat) Python A-Z™: Veri Bilimi ve Machine Learning

11. Bölüm - Makine Öğrenmesi III – Doğrusal Olmayan Regresyon Modelleri (291. Video) - Vahit Keskin anlattı.

K-En Yakın Komşu (KNN)

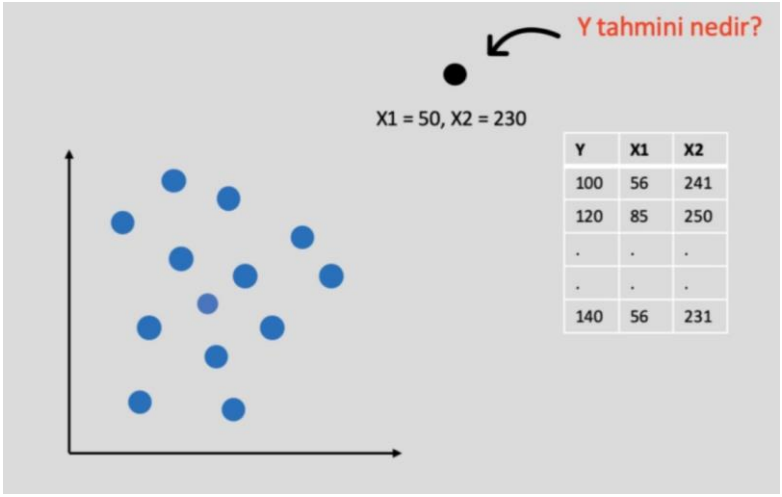
Tahminler gözlem benzerliğine göre yapılır. "Bana arkadaşını söyle sana kim olduğunu söyleyeyim." Parametrik olmayan bir öğrenme türüdür. Büyük veri setlerinde performans açısından çok da iyi olduğu söylenemez ama sınıflandırma problemleri için ortaya çıkmış daha sonra da regresyon problemlerine uyarlanmıştır. Yine de kullanılması, uygulaması kolay olduğundan tercih edilebilen kullanılabilir algoritmalardan birisidir.

"Tahminler gözlem benzerliğine göre yapılır." ne demek? Bir gözlem birimi geldiğinde bu gözlem birimine en benzer olan arkadaşlarına bakıyor, bu gözlem birimini en yakın olan 3 tane, 5 tane, 10 tane k-en yakın komşu gözlem birimini inceleyip kendisine en yakın olanlara göre bu gözlem biriminin bağımlı değişken değerini ortaya koymuş oluyor.

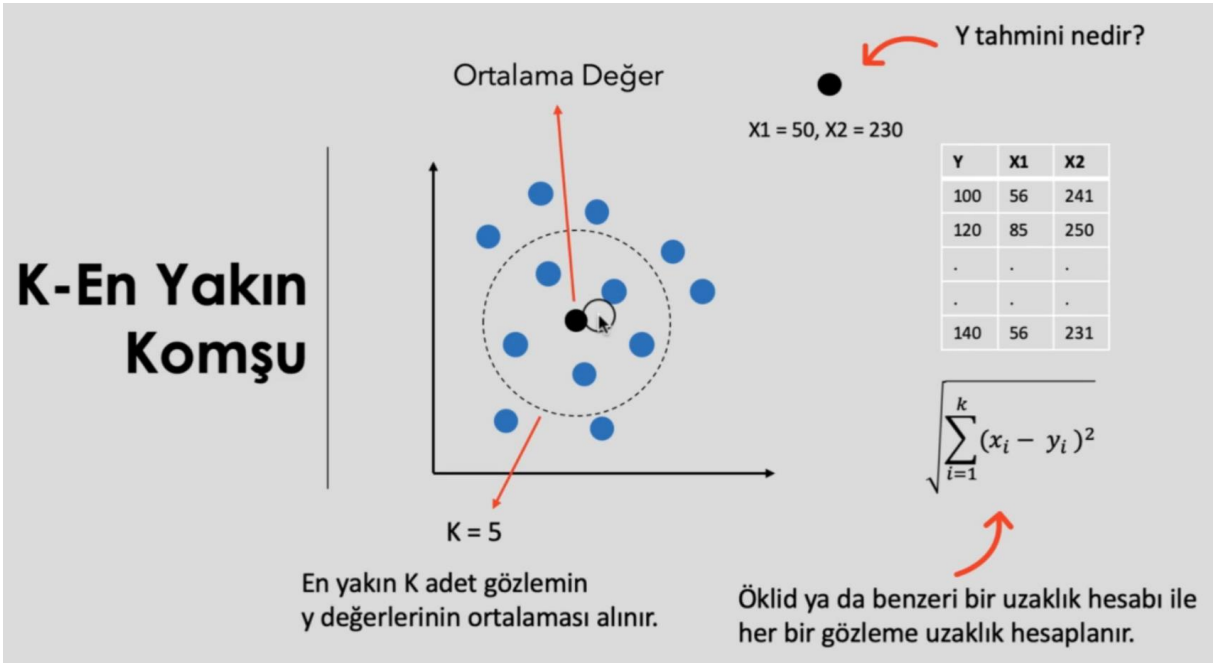
Bir örnek üzerinden bu durumu biraz daha açıklamaya çalışalım:

Gözlemlerin burada bu şekilde bir dağılımı olsun. Bunu y ve x olarak da düşünebilirsiniz, x'lerden birden fazla olup indirgenmiş şekilde de düşünebilirsiniz. Sağ tarafta ise örnek veri setimiz var. Y bağımlı değişkeni ve X1 ve X2 bağımsız değişkenleri olarak verilmiş.

Elimize yeni bir gözlem birimi geldiğinde diyor ki X1=50, X2=230. Y değerim ne olacak diye bir soru soruyor.



Bu durumda KNN'in yaptığı şey şu olmuş olacak. Öklid ya da benzeri bir uzaklık hesabı ile bu gözlemin (X1=50, X2=230) her bir gözleme uzaklığı hesaplanır. Daha sonra kendisine en yakın olan örneğin K=5 adet gözlemin bağımlı değişken (Y) değerlerinin ortalaması alınır. Bu durumda bu yeni gelen gözlemin bağımlı değişken değeri bu 5 değerin ortalaması olarak atanır.



KNN Basamakları

- Komşu sayısını belirle (K)
- Bilinmeyen nokta ile diğer tüm noktalar ile arasındaki uzaklıkları hesapla
- Uzaklıkları sırala ve belirlenen k sayısına göre en yakın olan k gözlemi seç
- Sınıflandırma ise en sık sınıf, regresyon ise ortalama değeri tahmin değeri olarak ver.