MAKİNA ÖĞRENMESİNDE SINIFLANDIRMA YÖNTEMLERİ

1. Lojistik Regresyon (Logistic Regression)

Tanım:

Lojistik regresyon, ikili sınıflandırma problemlerinde kullanılır ve bağımlı değişkenin belirli sınıflara ait olma olasılığını tahmin eder. Bir doğrusal modelin çıktısını lojistik (sigmoid) fonksiyon ile dönüştürerek olasılıkları hesaplar.

Kullanım Alanları:

Hastalık teşhisi (örneğin, hastanın belirli bir hastalığa sahip olup olmadığını tahmin etme)

Pazarlama kampanyalarında müşteri dönüşüm oranlarını tahmin etme

Avantajlar:

Kolay yorumlanabilirlik

Hızlı ve verimli hesaplama

Dezavantajlar:

Doğrusal ayrım gerektiren problemlerle sınırlıdır

Karmaşık ilişkileri modellemede yetersiz kalabilir

2. Karar Ağaçları (Decision Trees)

Tanım:

Karar ağaçları, verileri sınıflandırmak için dallanma yapısı kullanır. Her düğüm, bir özniteliği (feature) kontrol eder ve dallar bu özniteliğin olası değerlerini temsil eder. Ağaç yapısı, nihai yapraklarda sınıf etiketleriyle sonlanır.

Kullanım Alanları:

Müşteri segmentasyonu

Kredi risk değerlendirmesi

Avantajlar:

Kolay anlaşılır ve görselleştirilebilir

Hem kategorik hem de sayısal verilerle çalışabilir

Dezavantajlar:

Aşırı öğrenmeye (overfitting) yatkın

Veri gürültüsüne duyarlı

3. Rassal Ormanlar (Random Forests)

Tanım:

Rassal ormanlar, birden fazla karar ağacının bir araya gelmesiyle oluşan bir topluluk yöntemidir. Her ağaç, veri setinin ve özniteliklerin farklı bir alt kümesinde eğitilir. Sonuçta, tüm ağaçların tahminlerinin çoğunluğu kullanılarak nihai karar verilir.

Kullanım Alanları:

Hastalık teşhisi

Hisse senedi fiyat tahmini

Avantajlar:

Aşırı öğrenmeye karşı daha dayanıklıdır

Genellikle yüksek doğruluk sağlar

Dezavantajlar:

Hesaplama maliyeti yüksektir

Yorumlanabilirlik zordur

4. Destek Vektör Makineleri (Support Vector Machines - SVM)

Tanım:

SVM, verileri en iyi ayıran hiperdüzlemi bulmaya çalışır. Lineer olarak ayrılabilir veriler için uygundur, ancak çekirdek (kernel) fonksiyonları kullanarak lineer olmayan ayrımlar da yapılabilir.

Kullanım Alanları:

Görüntü tanıma

Genomik veri analizi

Avantajlar:

Yüksek boyutlu veriler için idealdir

Genel olarak yüksek performans gösterir

Dezavantajlar:

Büyük veri setlerinde eğitim süresi uzundur

Hiperparametre ayarı karmaşıktır

5. K-En Yakın Komşu (K-Nearest Neighbors - KNN)

Tanım:

KNN, yeni bir veri noktasının sınıfını tahmin etmek için, en yakın

k komşusunun sınıflarına bakar. Komşuların çoğunluğuna göre sınıf etiketi belirlenir.

Kullanım Alanları:

Desen tanıma

Müşteri segmentasyonu

Avantajlar:

Kolay ve sezgisel bir yöntemdir

Parametrik olmayan bir yöntemdir

Dezavantajlar:

Büyük veri setlerinde hesaplama maliyeti yüksektir

Gürültüye duyarlıdır

6. Naive Bayes

Tanım:

Naive Bayes, Bayes teoremine dayanır ve özniteliklerin birbirinden bağımsız olduğu varsayımını yapar. Bu basit varsayıma rağmen, birçok uygulamada oldukça etkilidir.

Kullanım Alanları:

E-posta spam tespiti

Doküman sınıflandırma

Avantajlar:

Hızlı ve verimli hesaplama

Az sayıda veri ile iyi performans gösterir

Dezavantajlar:

Öznitelikler arasındaki bağımlılıkları dikkate almaz

Gerçek dünya verilerinde varsayımları genellikle geçerli değildir

7. Yapay Sinir Ağları (Artificial Neural Networks)

Tanım:

Yapay sinir ağları, biyolojik sinir ağlarından esinlenerek geliştirilmiş modellerdir. Çok katmanlı yapılar, karmaşık ve derin ilişkileri modellemek için kullanılır. Derin öğrenme yöntemleri, daha büyük ve daha karmaşık sinir ağı modelleridir.

Kullanım Alanları:

Görüntü sınıflandırma

Doğal dil işleme

Avantajlar:

Karmaşık ve yüksek boyutlu veri setleri için güçlüdür

Özellikle derin öğrenme, çok katmanlı yapıların öğrenilmesini sağlar

Dezavantajlar:

Eğitim süresi uzundur

Büyük veri setleri ve yüksek hesaplama gücü gerektirir

8. Gradyan Artırma (Gradient Boosting)

Tanım:

Gradyan artırma, zayıf öğrenicileri (weak learners) kullanarak güçlü bir sınıflandırıcı oluşturmayı amaçlar. Her yeni model, önceki modellerin hatalarını düzeltmeye çalışır. XGBoost ve LightGBM gibi kütüphaneler, bu yöntemin güçlü uygulamalarıdır.

Kullanım Alanları:

Tahminleme yarışmaları (Kaggle gibi)

Finansal zaman serisi analizi

Avantajlar:

Yüksek performans

Esnek ve güçlü bir yöntemdir

Dezavantajlar:

Ayarları ve hiperparametre seçimi karmaşıktır

Eğitim süresi uzun olabilir

9. AdaBoost

Tanım:

AdaBoost (Adaptive Boosting), zayıf öğrenicileri ardışık olarak eğiterek, her yeni modelin önceki modelin hatalarını düzeltmeye çalıştığı bir yöntemdir. Yanlış sınıflandırılmış örneklere daha fazla ağırlık verilir.

Kullanım Alanları:

Yüz tanıma

Metin sınıflandırma

Avantajlar:

Hızlı ve verimli hesaplama

Düşük hafıza gereksinimi

Dezavantajlar:

Gürültüye duyarlıdır

Aşırı öğrenme riski vardır

10. Gaussian Mixture Models (GMM)

Tanım:

GMM, verileri belirli sayıda Gauss dağılımına ayıran bir modeldir. Her bir bileşen, verilerin belirli bir alt kümesini temsil eder. Genellikle kümeleme problemleri için kullanılsa da, sınıflandırma için de uygulanabilir.

Kullanım Alanları:

Anomali tespiti

Görüntü segmentasyonu

Avantajlar:

Esnek modelleme yeteneği

Karmaşık veri yapılarında iyi performans gösterir

Dezavantajlar:

Veri dağılımının Gaussian olması gereklidir

Eğitim süresi uzun olabilir