3.Hafta

MAKİNE ÖĞRENMESİNE GİRİŞ VE REGRESYONLAR

1) ML Nedir? Süreç Nasıl İşler?

Makine Öğrenmesi (ML), bilgisayar sistemlerinin verilerdeki kalıpları ve ilişkileri otomatik olarak öğrenerek belirli bir görevde performanslarını iyileştirmelerini sağlayan bir alandır. Temel amacı, geçmiş verilerden genelleme yaparak gelecekteki veriler hakkında tahminlerde bulunmak veya kararlar almaktır.

Makine Öğrenmesi Süreci:

1. **Veri Toplama (Data Collection):** Makine öğrenmesi modelleri için kaliteli ve yeterli veri olmazsa olmazdır. Bu veriler, metin, sayılar, resimler veya ses kayıtları olabilir.
2. **Veri Hazırlama (Data Preprocessing):** Toplanan veriler genellikle ham ve dağınıktır. Bu aşamada veriler temizlenir (eksik değerleri doldurma, hatalıları düzeltme), dönüştürülür (sayısal hale getirme, ölçeklendirme), özellik mühendisliği (mevcut özelliklerden yeni özellikler türetme) yapılır. Bu, modelin veriyi daha iyi anlamasını sağlar.
3. **Model Seçimi (Model Selection):** Çözmek istediğiniz probleme ve sahip olduğunuz veri tipine göre uygun bir makine öğrenmesi algoritması seçilir. Örneğin, sınıflandırma için lojistik regresyon veya karar ağaçları, regresyon için doğrusal regresyon gibi.
4. **Model Eğitimi (Model Training):** Hazırlanan veriler (eğitim verisi) kullanılarak seçilen algoritma eğitilir. Model, veri içindeki kalıpları öğrenmek için bir dizi yineleme (iterasyon) boyunca ayarlamalar yapar.
5. **Model Değerlendirme (Model Evaluation):** Eğitilen modelin ne kadar iyi performans gösterdiğini belirlemek için test verisi kullanılır. Modelin genelleme yeteneği (daha önce görmediği verilere ne kadar doğru tepki verdiği) bu aşamada ölçülür. R^2 skoru, ortalama karesel hata (MSE) gibi metrikler kullanılır.
6. **Model Optimizasyonu (Model Optimization/Hyperparameter Tuning):** Modelin performansı yeterli değilse, hiperparametreler (modelin öğrenme süreciyle ilgili ayarlamalar, veriyle ilgili olmayanlar) ayarlanır veya farklı bir model denenebilir.
7. **Dağıtım (Deployment):** Eğitilmiş ve doğrulanmış model, gerçek dünya uygulamalarında kullanılmak üzere dağıtılır.

2) Supervised (Denetimli) vs. Unsupervised (Denetimsiz) Öğrenme

Makine öğrenmesindeki algoritmalar genellikle iki ana kategoriye ayrılır:

1. Supervised Learning (Denetimli Öğrenme)

- **Tanım:** Bu öğrenme türünde, modelimize **etiketli (labeled) veri** setleri sunulur. Yani, hem girdi (özellikler) hem de bu girdilere karşılık gelen doğru çıktı (etiketler/hedefler) bilinir. Model, girdi ile çıktı arasındaki ilişkiyi bu etiketli verilerden öğrenir.
- **Amaç:** Modelin yeni, daha önce görmediği verilere dayanarak doğru tahminler yapabilmesini sağlamaktır.
- **Kullanım Alanları:**
    - **Sınıflandırma (Classification):** Girdinin hangi kategoriye ait olduğunu tahmin etme (örneğin, e-postanın spam mı değil mi olduğunu tahmin etme, resimdeki hayvan türünü belirleme).
    - **Regresyon (Regression):** Sürekli bir sayısal değeri tahmin etme (örneğin, ev fiyatını, hava sıcaklığını, hisse senedi fiyatını tahmin etme).

2. Unsupervised Learning (Denetimsiz Öğrenme)

- **Tanım:** Bu öğrenme türünde, modelimize **etiketsiz (unlabeled) veri** setleri sunulur. Modelin, verilerin içindeki gizli yapıları, kalıpları veya ilişkileri kendi başına keşfetmesi beklenir. Modelin neyi araması gerektiği konusunda önceden bir ipucu (etiket) verilmez.
- **Amaç:** Veri setindeki gizli yapıları bulmak, veriyi özetlemek veya benzer veri noktalarını gruplamaktır.
- **Kullanım Alanları:**
    - **Kümeleme (Clustering):** Benzer veri noktalarını gruplara ayırma (örneğin, müşteri segmentasyonu, belge sınıflandırması).
    - **Boyut Azaltma (Dimensionality Reduction):** Veri setindeki özellik sayısını azaltarak daha anlaşılır ve yönetilebilir hale getirme (örneğin, PCA - Temel Bileşen Analizi).
    - **Birliktelik Kuralı Madenciliği (Association Rule Mining):** Veri setindeki öğeler arasındaki ilişkileri bulma (örneğin, "birlikte alınan ürünler" önerileri).

3) Linear Regression (Doğrusal Regresyon)

Doğrusal Regresyon, hem istatistikte hem de makine öğrenmesinde en temel ve en yaygın kullanılan regresyon algoritmalarından biridir. Amacı, bir bağımlı değişken (hedef) ile bir veya daha fazla bağımsız değişken (özellik) arasında doğrusal bir ilişki modellemektir.

Sezgisel Anlatım (Intuition)

Hayal edin ki bir grup öğrencinin ders çalışma saatleri ile sınav notları arasındaki ilişkiyi incelemek istiyorsunuz.

- **Bağımsız Değişken (X):** Ders çalışma saatleri
- **Bağımlı Değişken (Y):** Sınav notu

Doğrusal regresyon, bu noktalar arasına en uygun **doğruyu** çizmeye çalışır. Bu doğru,

Y = aX + b (tek özellik için) veya

Y = b0 + b1X1 + b2X2 + ... + bnXn (çoklu özellik için) denklemiyle ifade edilir.

- **b0:** Y-kesen (intercept), tüm X değerleri 0 iken Y'nin tahmini değeridir.
- **b1, b2, ... bn:** Katsayılar (coefficients), ilgili X özelliğindeki bir birimlik değişimin Y üzerindeki etkisini gösterir.

Modelin amacı, bu b katsayılarını (doğrunun eğimi ve y-keseni) öyle bir şekilde bulmaktır ki, çizilen doğru, gerçek veri noktalarına en yakın konumda olsun. "En yakın" olmak, genellikle **Hata Kareleri Ortalaması (Mean Squared Error - MSE)** gibi bir ölçütle belirlenir. Model, tahmin edilen değerler ile gerçek değerler arasındaki farkların karelerini minimize etmeye çalışır.

Cost Function (Maliyet Fonksiyonu)

Maliyet fonksiyonu (veya kayıp fonksiyonu), bir makine öğrenmesi modelinin **tahminlerinin gerçek değerlerden ne kadar saptığını ölçen** fonksiyondur. Regresyon modellerinde genellikle **Ortalama Kare Hatası (MSE)** kullanılır. Amaç, bu fonksiyonun değerini **minimize ederek** modelin en iyi tahminleri yapmasını sağlayan parametreleri bulmaktır.

R^2 Skoru (Belirleme Katsayısı):

R^2 Skoru, regresyon modelinizin bağımlı değişkendeki **toplam varyansın ne kadarını bağımsız değişkenler tarafından açıklandığını** gösteren bir performans metriğidir. Kısaca, modelin verilere ne kadar **iyi uyduğunu** belirtir. 0 ile 1 arasında değişir; 1'e ne kadar yakınsa, model o kadar iyi uyum sağlamıştır.

$R^2 = 1 - (SStot / SSres)$

- **SSres:** Hata Kareleri Toplamı (modelin açıklayamadığı varyans)
- **SStot:** Toplam Kareler Toplamı (bağımlı değişkendeki toplam varyans)

Ridge ve Lasso Regresyon:

Doğrusal regresyonda **aşırı uyum (overfitting)** problemini çözmek ve modelin genelleme yeteneğini artırmak için **düzenlileştirme (regularization)** teknikleri kullanılır. Ridge ve Lasso regresyon, bu amaçla maliyet fonksiyonuna bir "ceza" terimi ekler. Bu cezalar, model katsayılarının çok büyük olmasını engelleyerek modeli basitleştirir.

1. Ridge Regresyon (L2 Düzenlileştirme)

Ridge regresyon, maliyet fonksiyonuna **katsayıların karelerinin toplamını** (L2 normu) cezalandıran bir terim ekler. Bu, katsayıları **sıfıra yaklaştırır** ancak hiçbir katsayıyı tam olarak sıfır yapmaz. Yani, tüm özellikleri modelde tutar ancak etkilerini azaltır.

2. Lasso Regresyon (L1 Düzenlileştirme)

Lasso regresyon, maliyet fonksiyonuna **katsayıların mutlak değerlerinin toplamını** (L1 normu) cezalandıran bir terim ekler. Lasso'nun en önemli özelliği, bazı katsayıları **tamamen sıfır yapabilmesidir**. Bu, etkisiz özellikleri modelden otomatik olarak çıkararak **özellik seçimi** yapılmasına olanak tanır.

ElasticNet Açıklaması

ElasticNet Regresyon, hem **Ridge (L2)** hem de **Lasso (L1)** düzenlileştirmenin bir **kombinasyonudur**.

- **Maliyet fonksiyonu,** hem L1 hem de L2 cezalandırma terimlerini içerir.
- **Avantajı:** Hem Lasso'nun özellik seçimi yeteneğini (gereksiz katsayıları sıfırlama) hem de Ridge'in yüksek korelasyonlu özellik gruplarını daha iyi yönetme (katsayıları birlikte küçültme) avantajlarını sunar.
- **Ayarlama:** alpha (toplam ceza miktarı) ve l1_ratio (L1 ve L2 arasındaki oran) olmak üzere iki ana parametre ile ayarlanır.

  l1_ratio = 1 olduğunda Lasso'ya,

  l1_ratio = 0 olduğunda Ridge'e dönüşür.

ElasticNet, özellikle çok sayıda özellik olduğunda, bu özellikler arasında yüksek korelasyon bulunduğunda ve hem özellik seçimi yapmak hem de aşırı uyumu kontrol etmek istediğinizde tercih edilir.