Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2026

Hafta 7
Regresyon Modelleri

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

01

Doğrusal ve
Lojistik Regresyon

02

MSE, RMSE

03

MAE, R^2

04

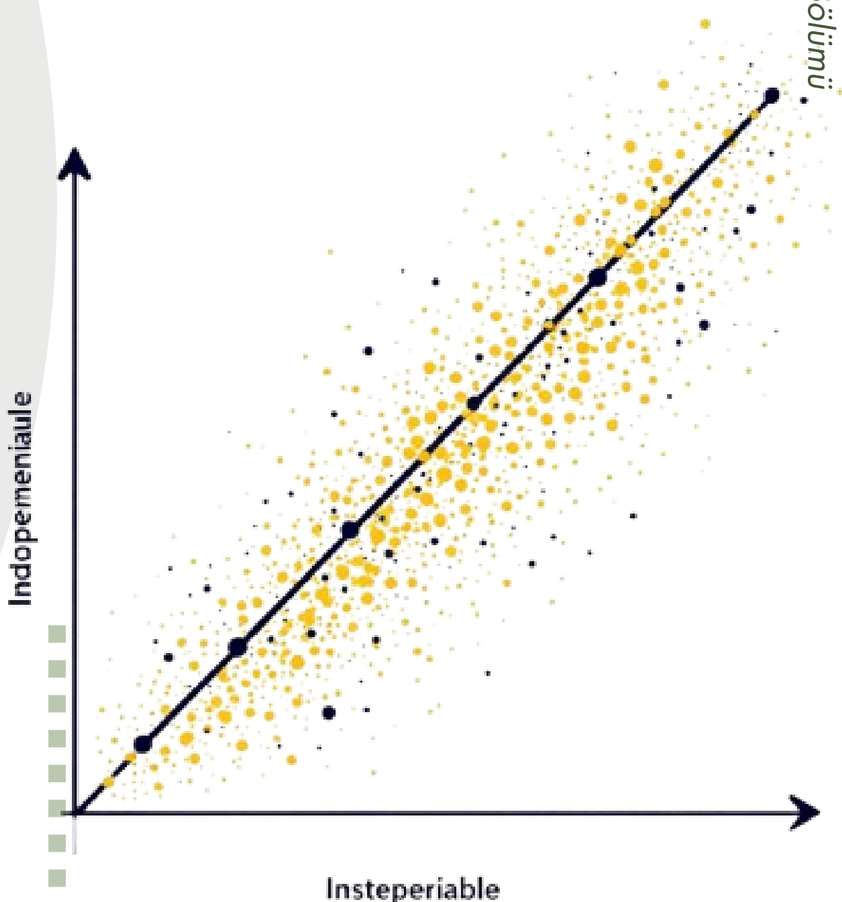
Uygulama

Regresyon Nedir?

Regresyon, bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişkiyi analiz etmek için kullanılan istatistiksel bir tekniktir.Bağımsız değişkenler, bağımlı değişkeni etkileyen faktölerdir. Örneğin ev fiyatını tahmin ederken büyüklük, konum ve oda sayısı bağımsız değişken olabilir.

Regresyon analizi değişkenler arasındaki ilişkiyi inceleyen bir istatistiksel araçtır.

Regresyon modelleri, veri kümeleri içindeki kalıpları ve trendleri belirlemeye yardımcı olur. Bu modeller, bağımlı değişkenin gelecekteki değerlerini tahmin etmek ve bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini ölçmek için kullanılır.



Regresyon Modelleri

Doğrusal Regresyon

Doğrusal regresyon, bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkiyi doğrusal (düz bir çizgi) bir modelle ifade eden yöntemdir. Amaç, veri noktaları arasından geçen en uygun doğruyu bulmaktır. Tahmin edilen çıktı sürekli bir sayısal değerdir. Örneğin, ev fiyatlarını oda sayısı ve metrekaresine göre tahmin etmek doğrusal regresyonla yapılabilir.

Lojistik Regresyon

Lojistik regresyon, adında "regresyon" geçmesine rağmen sınıflandırma problemleri için kullanılır. Çıktı, belirli bir sınıfa ait olma olasılığıdır (örneğin, evet/hayır, hasta/sağlıklı). Model, veriyi doğrusal bir fonksiyonla işler ancak sonuçları sigmoid fonksiyonundan geçirerek 0 ile 1 arasında bir olasılık değeri üretir. Örneğin, bir öğrencinin sınavı geçme ihtimalini çalıştığı saatlere göre tahmin etmek lojistik regresyonla yapılabilir.

Özellik	Doğrusal Regresyon	Lojistik Regresyon
Amaç	Sürekli sayısal değer tahmini yapmak	Kategorik sınıflandırma yapmak
Çıktı Tipi	Herhangi bir reel sayı	0 ile 1 arasında olasılık
Kullanım Alanı	Ev fiyatı, sıcaklık tahmini, gelir tahmini	E-posta spam mi değil mi, hasta/sağlıklı, evet/hayır
Model Yapısı	Doğrusal denklem (y = ax + b)	Sigmoid fonksiyon ile olasılık üretir
Hata Ölçütü	Ortalama Kare Hata (MSE) gibi metrikler	Doğruluk, F1 Skoru, ROC-AUC gibi metrikler
Doğruluk Yorumu	Tahmin edilen değer ile gerçek değer arasındaki fark üzerinden değerlendirilir	Olasılık tahminlerinin sınıf etiketleriyle ne kadar örtüştüğüne göre değerlendirilir

Regresyon Modeli Oluşturma Adımları

Veri Hazırlama → Bu adımda veri temizlenir ve eksik değerler tamamlanır. Veri kalitesi modelin performansına doğrudan etkiler

Model Seçimi → Doğrusal, lojistik veya polinomal resgresyon gibi çeşitli modeller arasından veri yapısına en uygun olanı seçilir.

Model Eğitimi → Seçilen model, veri kümesinde kullanılarak eğitilmelidir. Eğitim sürecinde, modelin parametreleri veri kümesine uydurulur ve en uygun değerler belirlenir.

Regresyon Modeli Oluşturma Adımları

Model Değerlenirmesi → Eğitim tamamlandıktan sonra, modelin performansı değerlendirilmelidir.Bu, modelin tahmin doğruluğunu ve genel performansını ölçmek için gerçekleştirilir.

Tahmin → Model değerlendirildikten sonra, yeni veri noktaları için tahminlerde bulunmak üzere kullanılabilir. Bu, modelin eğitildiği aynı veri kümesinde veya yeni bir veri kümesinde gerçekleştirilebilir.

Regresyon Modeli Değerlendirme ve Performans Ölçütleri

MSE (Ortalama Kare Hatası):

Tahmin edilen değer ile gerçek değer arasındaki farkların karelerinin ortalamasıdır. Hata değerlerini karesini alarak pozitif hale getirir ve büyük hataları daha fazla cezalandırır. Değeri ne kadar küçükse modelin başarısı o kadar yüksektir.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

MAE (Ortalama Mutlak Hata):

Tahmin edilen değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasını ifade eder. Bu metrik, hataların yönünü dikkate almaz, yalnızca büyüklüğüne odaklanır. Ayrıca, uç değerlerden (outlier) daha az etkilenir ve büyük hataları MSE kadar ağır cezalandırmaz.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Regresyon Modeli Değerlendirme ve Performans Ölçütleri

RMSE (Kök Ortalama Kare Hatası):

MSE'nin kareköküdür. Hata birimini, tahmin edilen değerlerin birimi ile aynı hale getirir. Bu sayede yorumlaması daha kolaydır. RMSE değeri düşükse model daha iyi tahmin yapıyor demektir.

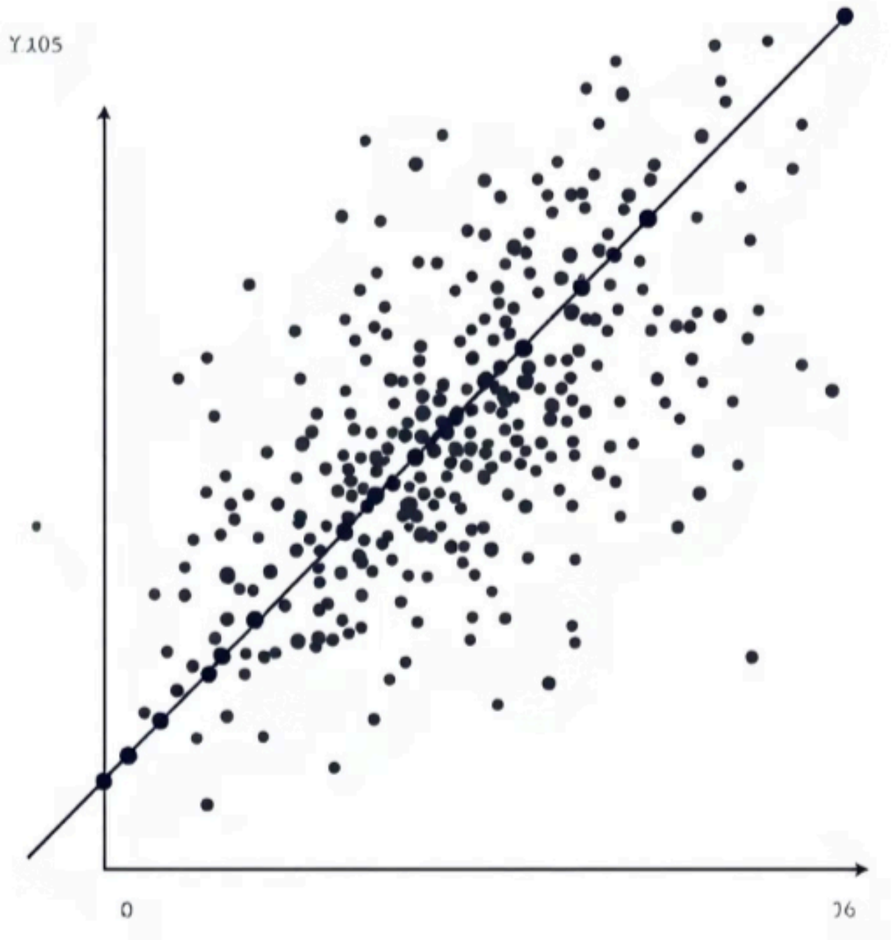
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

R^2 (Belirleme Katsayısı):

Modelin bağımlı değişkendeki toplam varyansın ne kadarını açıkladığını gösterir. 0 ile 1 arasında bir değer alır. 1'e yakın olması, modelin veriye iyi uyduğunu; 0'a yakın olması ise modelin varyansı açıklayamadığını gösterir. Negatif bir değer, modelin rastgele tahminden bile kötü performans gösterdiğini ifade eder.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Basit Doğrusal Regresyon



Temel Denklem

Stokastik (Olasılıklı) bir model olan ve ana kütledeki ilişkiyi gösteren basit doğrusal regresyon denklemi şöyle ifade edilir: y=B0 + B1X + ε

Parametreler

B0 = Doğrunun y-eksenini kestiği yer ve regresyon sabitidir.
B1 = Doğrunun eğimi veya regresyon katsayısıdır.
ε = Rastgele hata değeridir.

Uygulama

Basit doğrusal regresyon ; bağımsız değişken ile bağımlı değişkendeki değişimi açıklamayı, bağımsız değişkendeki bir birimlik değişimin bağımlı değişken üzerindeki etkisini ölçmeyi amaçlar

Regresyon Modeli Varsayımlar ve Tahminler

Doğrusallık

Doğrusal regresyon modelleri, bağımsız değişkenler ile bağımlı değişkenler arasında doğrusal bir ilişki olduğunu varsayar. Bu varsayım, modelin tahmin doğruluğunu etkileyebilir.

Bağımsızlık

Regresyon modelleri, hata terimlerinin birbirinden bağımsız olduğunu varsayar. Bu varsayım, modelin tahmin doğruluğunu etkileyebilir ve zaman serisi verilerinde özellikle önelidir.

Normal Dağılım

Regresyon modelleri, hata terimlerinin normal bir dağılıma sahip olduğunu varsayar.Bu varsayım, modelin tahmin doğruluğunu etkileyebilir ve istatistiksel testlerin geçerliliği için önemlidir.

Sabit Varyans

Regresyon modelleri, hata terimlerinin tüm bağımsız değişken değerleri için aynı varyansa sahip olduğunu varsayar. Bu varsayım, modelin tahmin doğruluğunu etkileyebilir.

Regresyon Modelinin Avantajları ve Sınırları

Avantajları

- Veri içindeki trendleri ve kalıpları belirleme
- Gelecekteki değerleri tahmin etme
- Bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini ölçme
- Kolay uygulanabilirlik ve yorumlanabilirlik.

Sınırları

- Doğrusal olmayan ilişkilere uyum sağlamakta zorluk
- Çoklu değişkenli modellerde değişken seçimi zorluğu
- Veri kalitesine duyarlılık
- Varsayımların ihlali modelin doğruluğunu etkileyebilir.

UYGULAMA

Konut Fiyat Tahmini

Kod Parçası

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import fetch_california_housing
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
housing = fetch_california_housing(as_frame=True)
df = housing.frame
X = df[['AveRooms']]      # Ortalama oda sayısı
y = df['MedHouseVal']     # Ortalama ev değeri (100.000 $ cinsinden)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
print(f"Model Denklemi: MEDV = {model.intercept_:.2f} + {model.coef_[0]:.2f} * AveRooms")

# Test verisiyle tahmin
y_pred = model.predict(X_test)

# Model başarı ölçütleri
print(f"Ortalama Hata Kareleri (MSE): {mean_squared_error(y_test, y_pred):.4f}")
print(f"R-Kare (R²): {r2_score(y_test, y_pred):.4f}")

# Örnek tahmin: Ortalama oda sayısı 6 olursa
test_value = pd.DataFrame({'AveRooms': [6]})
prediction = model.predict(test_value)[0]
print(f"\nOrtalama oda sayısı = 6 için tahmini ev fiyatı: {prediction*100000:.0f} $")
```

Kod Çıktısı

```
Model Denklemi: MEDV = 1.65 + 0.08 * AveRooms
Ortalama Hata Kareleri (MSE): 1.2923
R-Kare (R²): 0.0138

Ortalama oda sayısı = 6 için tahmini ev fiyatı: 211530 $
```

UYGULAMA

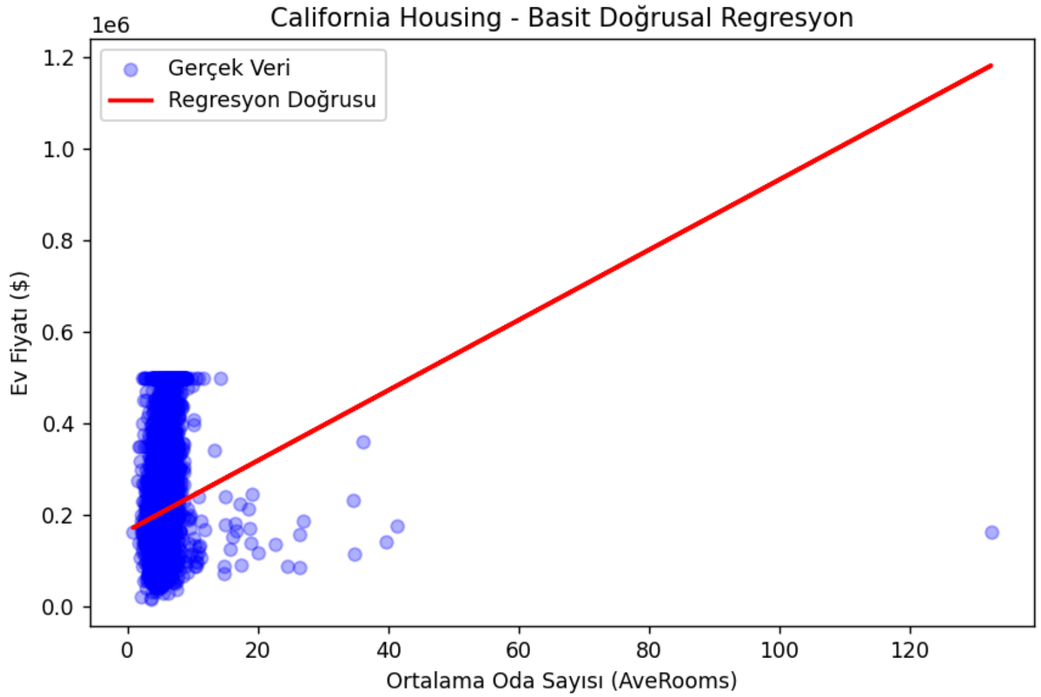
Konut Fiyat Tahmini

Kod Parçası

```python
# Görselleştirme
plt.figure(figsize=(8,5))
plt.scatter(X_test, y_test*100000, alpha=0.3, color="blue", label="Gerçek Veri")
plt.plot(X_test, y_pred*100000, color="red", linewidth=2, label="Regresyon Doğrusu")
plt.xlabel("Ortalama Oda Sayısı (AveRooms)")
plt.ylabel("Ev Fiyatı ($)")
plt.title("California Housing - Basit Doğrusal Regresyon")
plt.legend()
plt.show()

```

Kod Çıktısı



Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2026