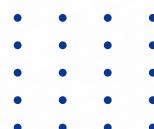
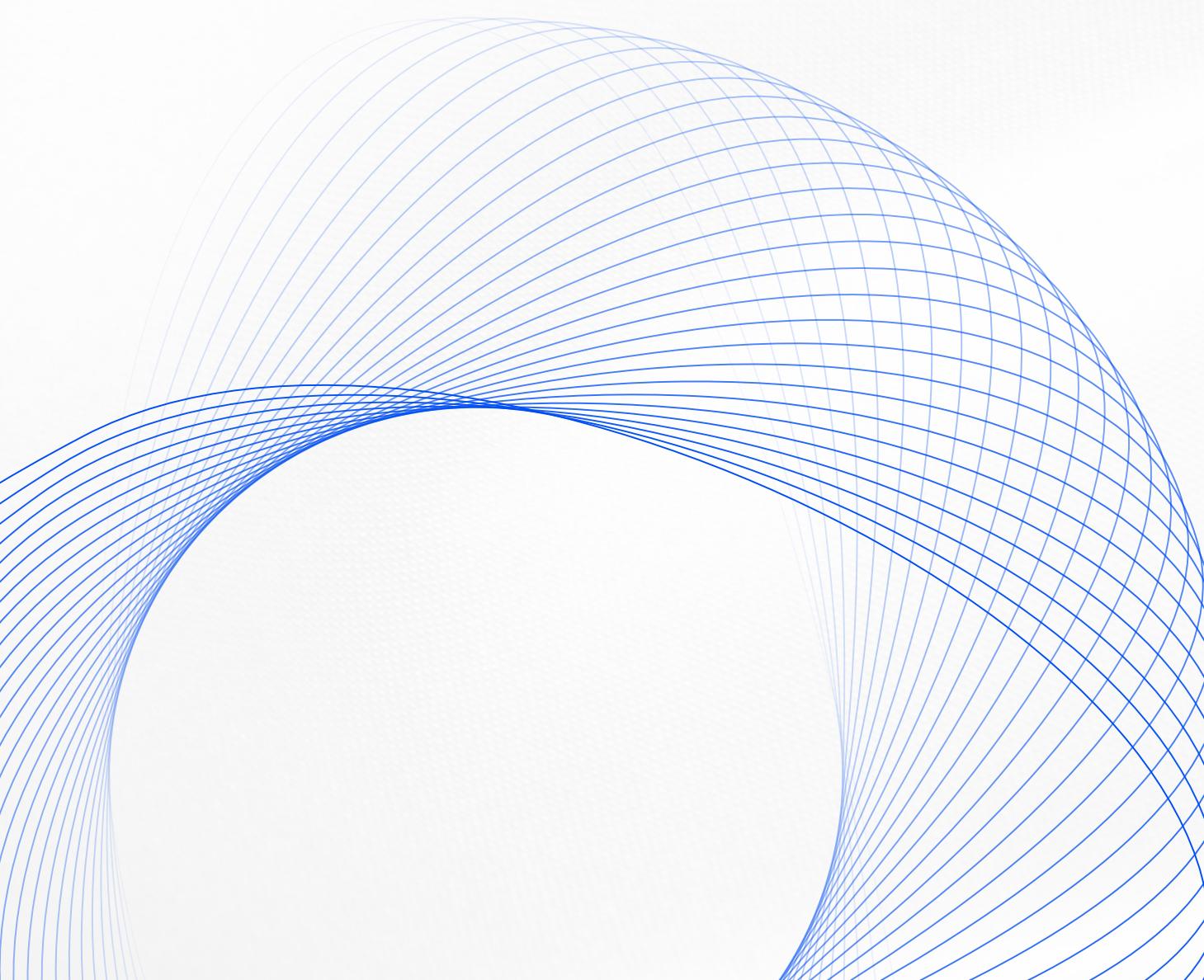


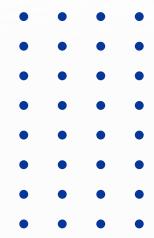
Regresyon Modelleri

Presented by Juliana Silva

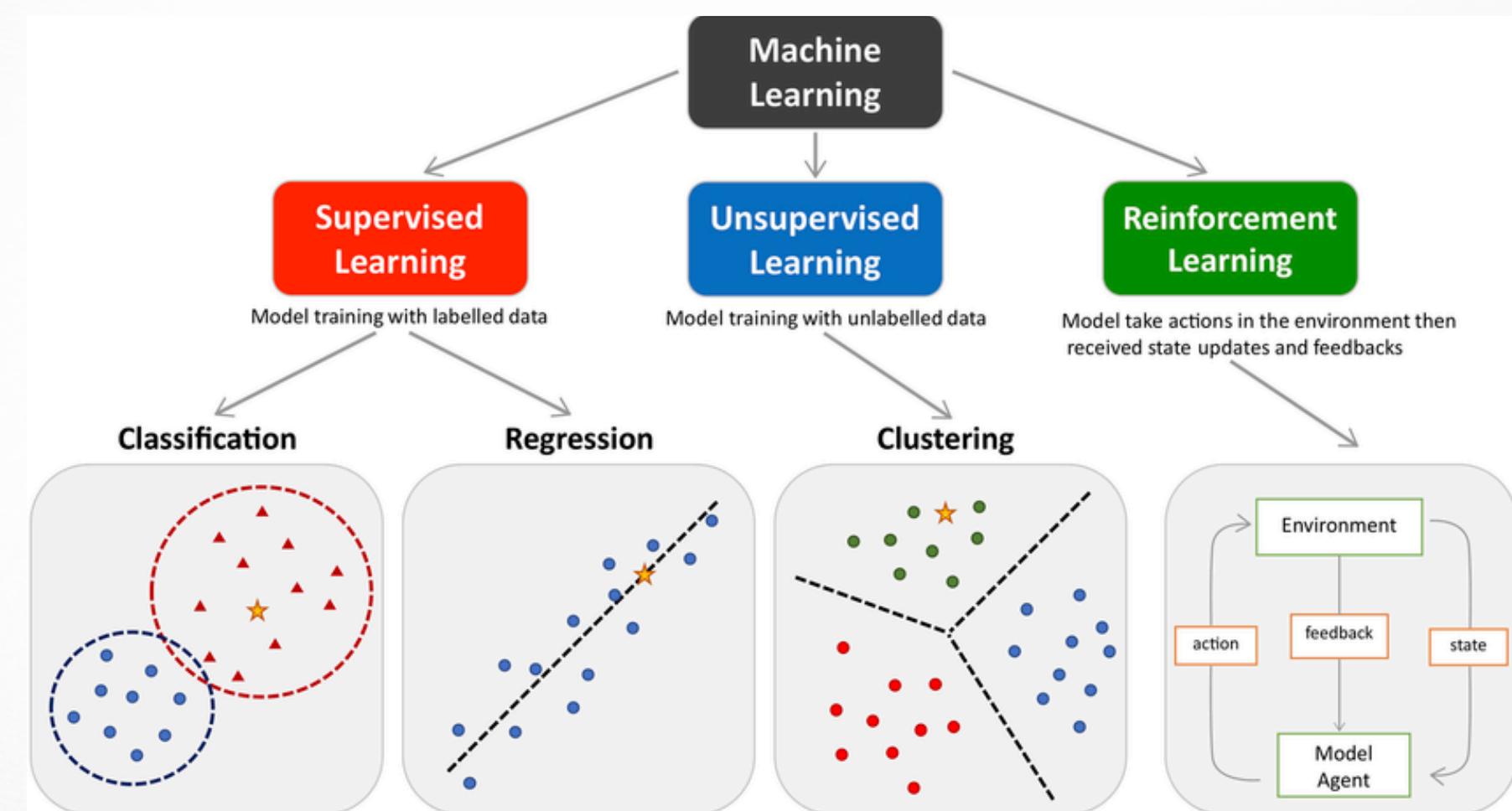


SUNUM AKIŞI

- 
- 1 Makine Öğrenmesi ve Regresyon
 - 2 Basit Doğrusal Regresyon
 - 3 Model Başarısını Nasıl Ölçeriz?
 - 4 Çoklu Doğrusal Regresyon
 - 5 Polinom Regresyonu
 - 6 Doğrusal Olmayan Modeller
 - 7 Hangi Modeller Ne Zaman Kullanılır?
 - 8 Notebook Session



Bölüm 1: Makine Öğrenmesi ve Regresyon



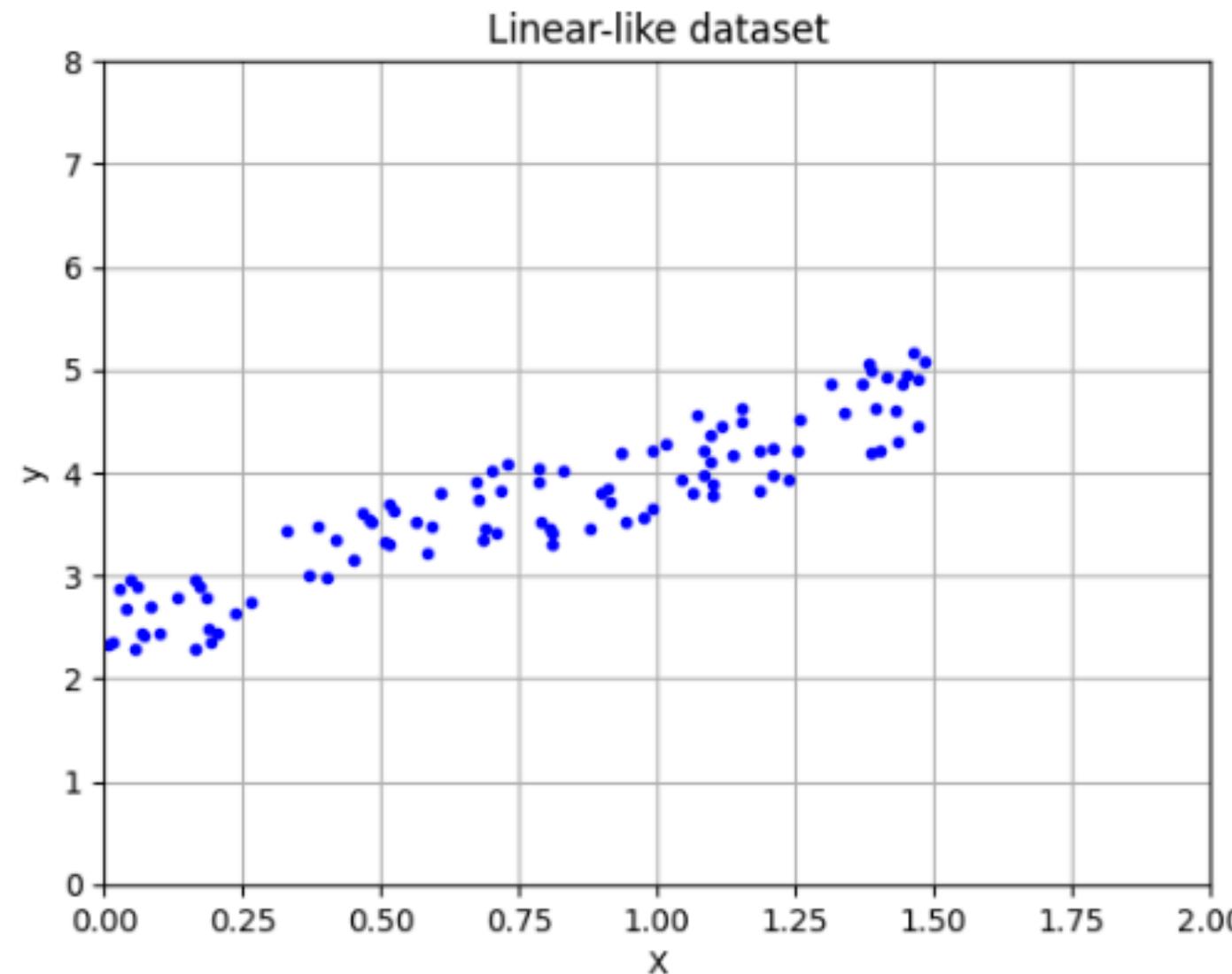
Sınıflandırma vs. Regresyon

- Denetimli öğrenme modelleri: Sınıflandırma ve Regresyon.
- Sınıflandırma: Kategori tahmini (Kedi mi Köpek mi?).
- Regresyon: Sayısal değer tahmini (Otomobil fiyatı).

Gerçek Hayattan Regresyon Örnekleri

- Emlak fiyat tahmini.
- Satış tahminleri.
- Hava durumu tahmini.
- Araç yakıt tüketimi tahmini.

Bölüm 2: Basit Doğrusal Regresyon (Simple Linear Regression)



Regresyon algoritmalarında bağımsız değişkenler (X) ve bağımlı değişken (Y) arasındaki ilişkiyi formüle etmeye çalışırız.

- 1. Bağımlı Değişken (Target - y):** Her zaman tahmin etmek istediğimiz şeydir. Örneğin, bir otomobilin fiyatı. Y, bağımsız değişkenlere (X'lere) bağımlıdır.
- 2. Bağımsız Değişkenler (Feature - x):** Bağımlı değişkenin (Y) tahmin edilmesini sağlayan diğer verilerdir. Örneğin, marka, model, kilometre, ağır hasar kaydı gibi özelliklerdir.

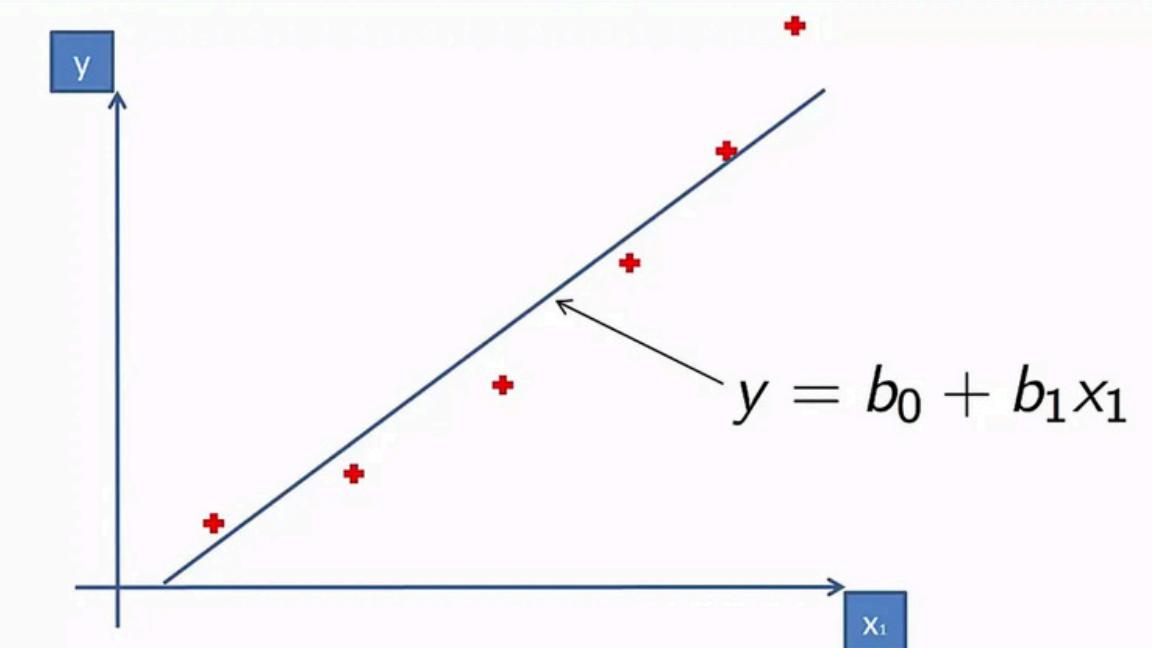
Matematiksel Formül

Basit doğrusal regresyon formülü: $Y = Ax + B$

A (Eğim/Slope): Bağımsız değişkenin (X) katsayısıdır. Bu katsayı, her verinin Y'yi ne oranda etkilediğini gösterir. Örneğin, otomobilin yılı bir oranda etkilerken, kilometresi daha farklı bir oranda etkileyebilir. Eğim ne kadar dik olursa, tahminler X'teki küçük değişikliklere o kadar duyarlı olur.

B (Kesişim/Intercept): Bir sabittir. Bazen bazı değerler sabittir ve bu sabiti bulup değişkenlere ekleyebilmemiz gereklidir. Örneğin, Migros alışverişinde poşet fiyatı gibi sabit bir değer olabilir. Regresyon doğrusunun y-eksenini kestiği noktası (y-axis intercept).

Simple Linear Regression

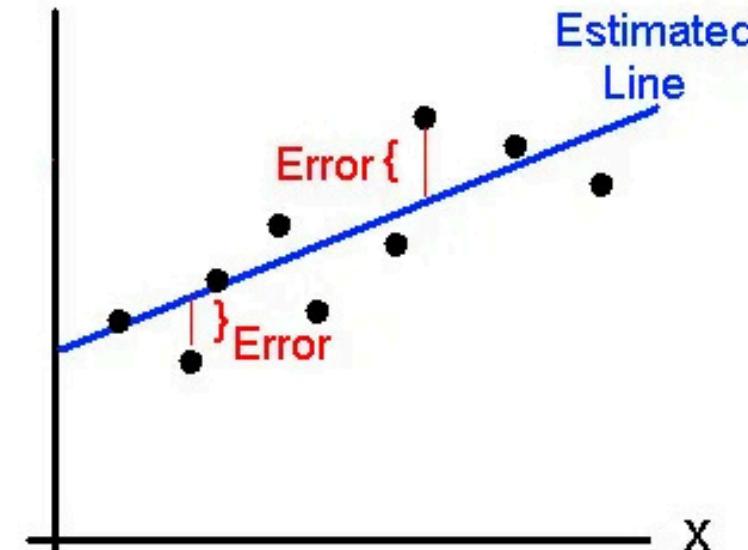


Hata Kareleri Toplamı

Elimizdeki veriler, çizdiğimiz doğru üzerinde değildir; bir dağılım gösterir. Bir noktanın gerçek değeri ile modelin tahmin ettiği değer arasındaki mesafeye **artık (residual)** denir.

$$\hat{Y}_i = b_0 + b_1 X_i$$

Estimated (or predicted) Y value for observation i
Estimate of the regression intercept
Estimate of the regression slope
Value of X for observation i



Model Bu Doğruyu Nasıl Çizer?

Doğrusal regresyonun standart yöntemi olan **OLS (Ordinary Least Squares)**, hataların kareleriyle ilgilenir. OLS, bu artıkların **karelerinin toplamını en aza indirmeye** çalışan bir yöntemdir.

Bu yöntemin amacı, **en küçük hata karelerinin toplamı** sonucunu veren doğruya bularak verilerden geçen en iyi regresyon çizgisini belirlemektir.

Modelin çizdiği doğrunun "**en iyi uyan**"(best fit line) doğru olarak nitelendirilebimesi için, veriye olan uzaklıklarının minimum olması gereklidir.

Bölüm 3: Başarıyı Nasıl Ölçeriz? (Metrikler)

R-Kare (R-Square)

R^2 skoru, modelin bağımlı değişkenindeki (Y) toplam varyansı ne kadar iyi açıkladığını gösterir.

Eğer R^2 değeri 1'e yakınsa, modelin veriye çok iyi uyduğu anlamına gelir (Ancak bu durum yüksek varyans ve aşırı uyum riskini de getirebilir). Eğer model eğitim verisindeki kalıpları etkili bir şekilde öğrenirse ve yeni verilere iyi genelleme yaparsa "iyi" kabul edilir.

Hata Metrikleri (MAE, MSE, RMSE)

- **MSE (Mean Squared Error)**: Modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkın **karelerinin ortalamasıdır**. **Büyük hataları cezalandırır**: Hataların karesi alındığı için, büyük hataların (aykırı değerlerin) modelin toplam hatasına etkisi daha fazla olur.
- **MAE (Mean Absolute Error)**: Ortalama mutlak hata, hataların mutlak değerlerinin ortalamasını alır. Büyük hataları MSE kadar cezalandırmaz.
- **RMSE (Root Mean Squared Error)** : Kök ortalama kare hata, MSE değerinin kareköküdür.

Bölüm 4: Çoklu Doğrusal Regresyon

Gerçek hayatın problemleri çok daha karmaşıktır. Bu nedenle, asıl kullanılan model, **Çoklu Doğrusal Regresyon (Multiple Linear Regression)** modelidir.

- **Birden Fazla Bağımsız Değişken:** Çoklu regresyonda artık bir tane değil, birden fazla bağımsız değişken (X) vardır. Veri setinde yüzlerce boyut olabilir.

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Bu model tipi, daha fazla özellik kullanarak tahmin yapmamızı sağlar, ancak bu durum beraberinde yüksek varyans (aşırı uyum) gibi sorunları da getirebilir. Bu sorunları çözmek için *Regülarizasyon* teknikleri (L1 ve L2) kullanılır. L2 (Ridge) regülarizasyonu, yüksek varyans sorununu çözmek için hataların karelerinin toplamına, katsayıların karelerinin toplamı üzerinden bir ceza ekler

Temel Varsayımlar (Assumptions)

★ Doğrusallık (Linearity):

- Doğrusal regresyonun temel varsayıımı, bağımlı değişken (Y) ile bağımsız değişkenler (X 'ler) arasında doğrusal bir ilişki olduğudur.
- Eğer verideki ilişki doğrusal değilse, doğrusal regresyon modeli yetersiz kalır ve underfitting sorununa yol açar.
bkz. Polynomial Regression

★ 2. Çoklu Bağlantı Sorunu (Multicollinearity):

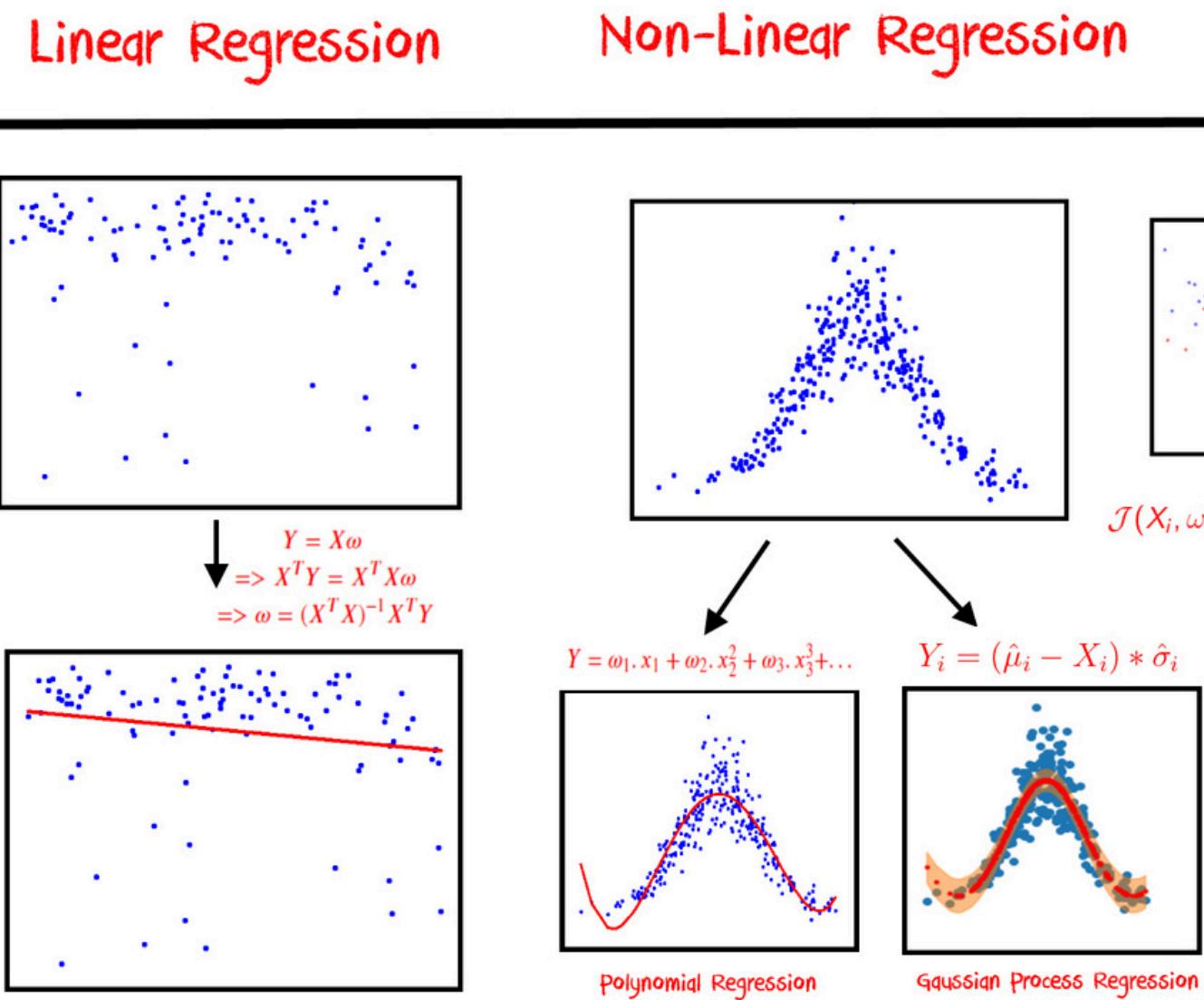
Çoklu bağlantı, bir regresyon modelindeki iki veya daha fazla bağımsız değişkenin birbirleriyle önemli ölçüde bağlantılı olması, yani aralarında yüksek korelasyon bulunması sorunudur.

bkz. Korelasyon Matrisi

bkz. Varyans Şisirme Faktörü (VIF)



Bölüm 5: Polinom Regresyonu (Polynomial Regression)



Veri setindeki ilişki kuadratik (eğrisel) bir yapıdaysa, basit bir doğrusal regresyon modeli yetersiz kalır ve **eksik uyum (underfitting)** sorununa yol açar.

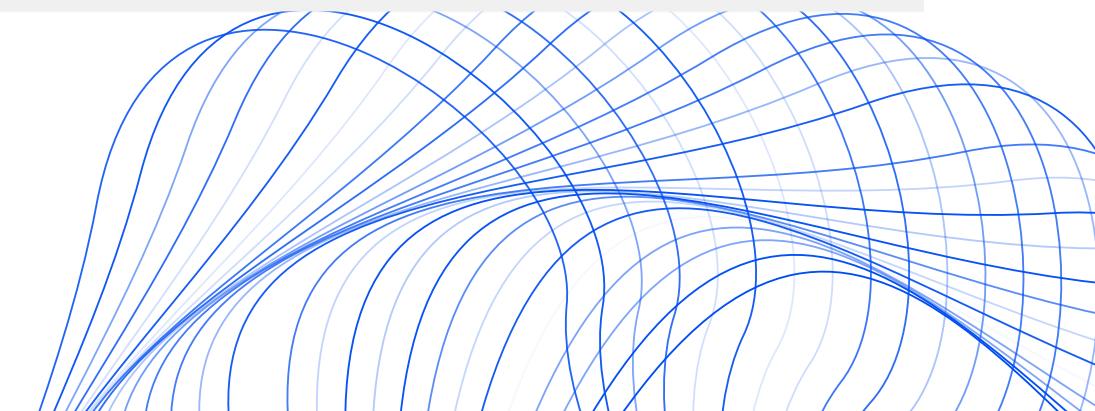
**** Degree artması** = polynomial regresyon modeline daha fazla polinom terimi eklenmesi demektir.

Polynomial regresyonda iki feature (x_1 ve x_2) için degree = 3 seçildiğinde model;
 x_1 ve x_2 'nin kendileri,
kareleri (x_1^2, x_2^2),
küpleri (x_1^3, x_2^3) ile birlikte tüm etkileşim terimlerini **$x_1 \cdot x_2, x_1^2 \cdot x_2$ ve $x_1 \cdot x_2^2$** ekleyerek toplam 10 özellik üretir. Bu, modelin daha karmaşık veri ilişkilerini yakalamasını sağlar.

Aşırı Öğrenme (Overfitting) ve Eksik Öğrenme (Underfitting)

Makine öğrenmesi modellerinin temel amacı, hem eğitim verisindeki desenleri etkili bir şekilde öğrenmek hem de **yeni, görülmeyen verilere** **iyi genelleme** yapmaktadır.

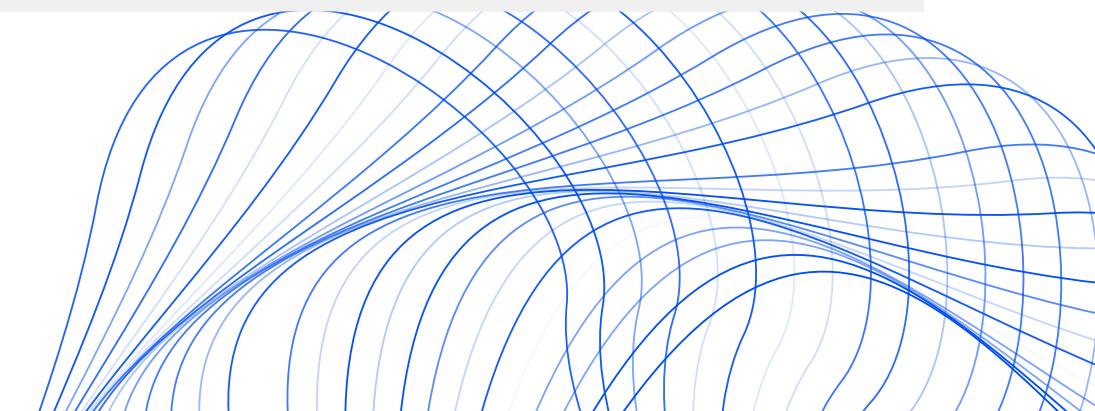
- **Eksik Öğrenme (Underfitting):** Modelin **çok basit** olması durumudur. Model, verideki baskın desenleri yakalamakta başarısız olur ve hem eğitim hem de test verilerinde tutarlı bir şekilde **yüksek hata** verir. Bu, yüksek yanılık (high bias) ve düşük varyans (low variance) ile karakterize edilir.
- **Aşırı Öğrenme (Overfitting):** Modelin **çok karmaşık** olması durumudur. Model, eğitim verisindeki rastgele gürültüyü, aykırı değerleri ve alakasız detayları da dahil ederek veriyi **ezberler**.



Aşırı Öğrenme (Overfitting) ve Eksik Öğrenme (Underfitting)

Makine öğrenmesi modellerinin temel amacı, hem eğitim verisindeki desenleri etkili bir şekilde öğrenmek hem de **yeni, görülmeyen verilere** **iyi genelleme** yapmaktadır.

- **Eksik Öğrenme (Underfitting):** Modelin **çok basit** olması durumudur. Model, verideki baskın desenleri yakalamakta başarısız olur ve hem eğitim hem de test verilerinde tutarlı bir şekilde **yüksek hata** verir. Bu, yüksek yanılık (high bias) ve düşük varyans (low variance) ile karakterize edilir.
- **Aşırı Öğrenme (Overfitting):** Modelin **çok karmaşık** olması durumudur. Model, eğitim verisindeki rastgele gürültüyü, aykırı değerleri ve alakasız detayları da dahil ederek veriyi **ezberler**.

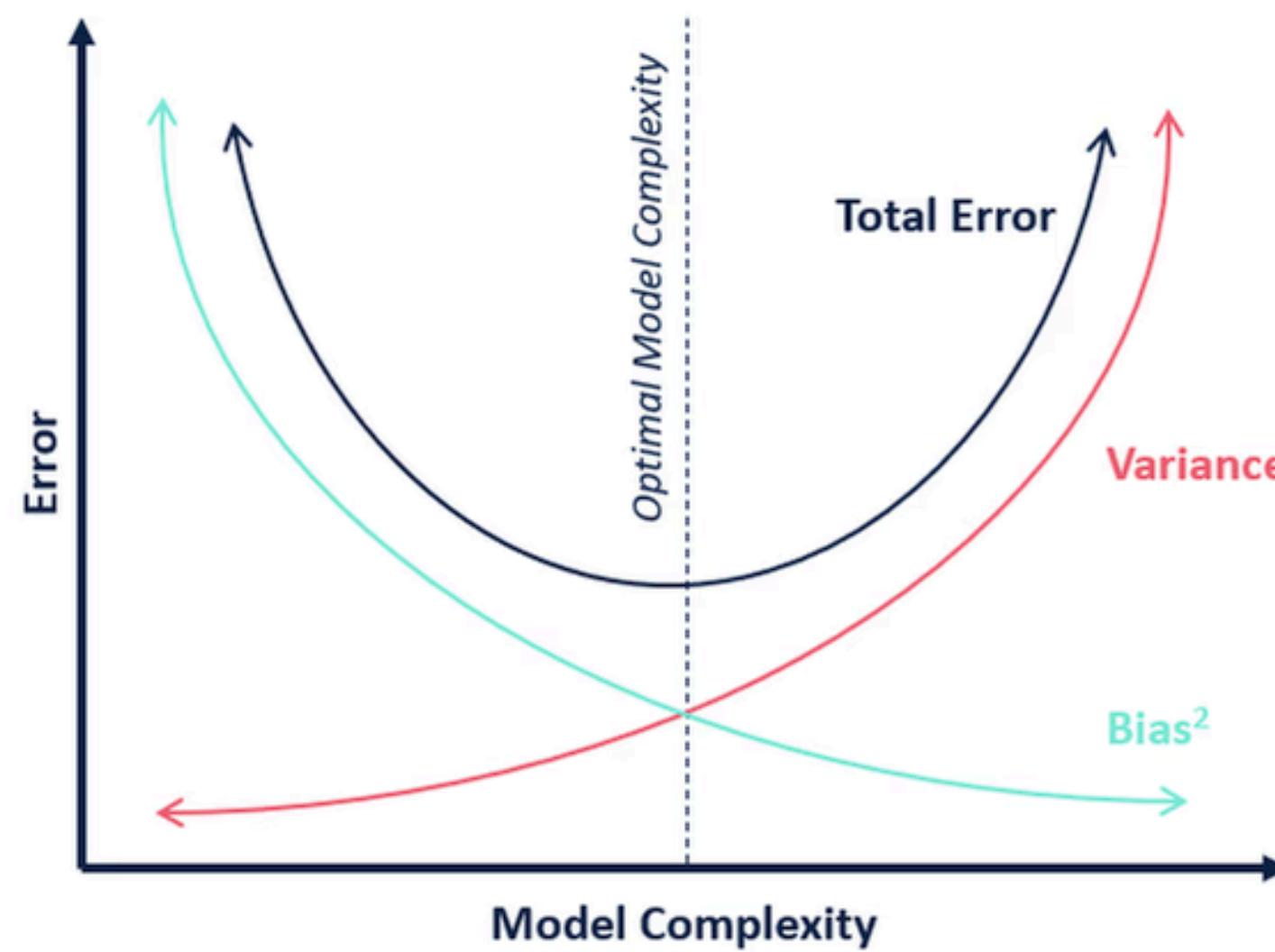


Bias-Variance (Yanlılık-Varyans) Trade-off:

Bias ve Variance model performansını ve genelleme yeteneğini doğrudan etkileyen iki temel hata kaynağıdır.

Amaç: Hem yanlılığı hem de varyansı en aza indiren optimal dengeyi bulmaktır.

Trade-off: Model karmaşıklığını artırmak yanlılığını azaltır, ancak varyansı artırır (aşırı öğrenme riski). Model karmaşıklığını azaltmak ise varyansı azaltır, ancak yanlılığını artırır (eksik öğrenme riski)

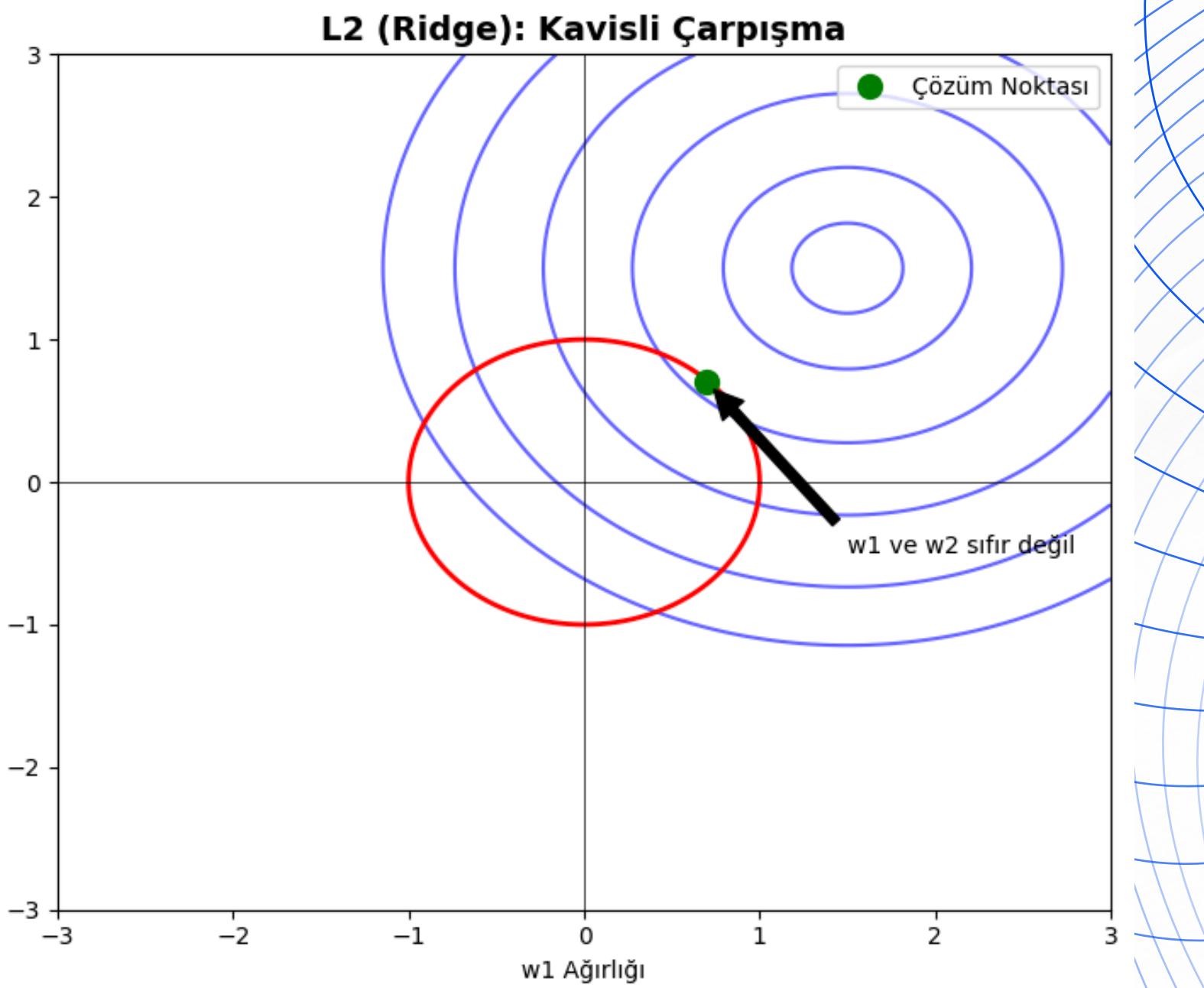
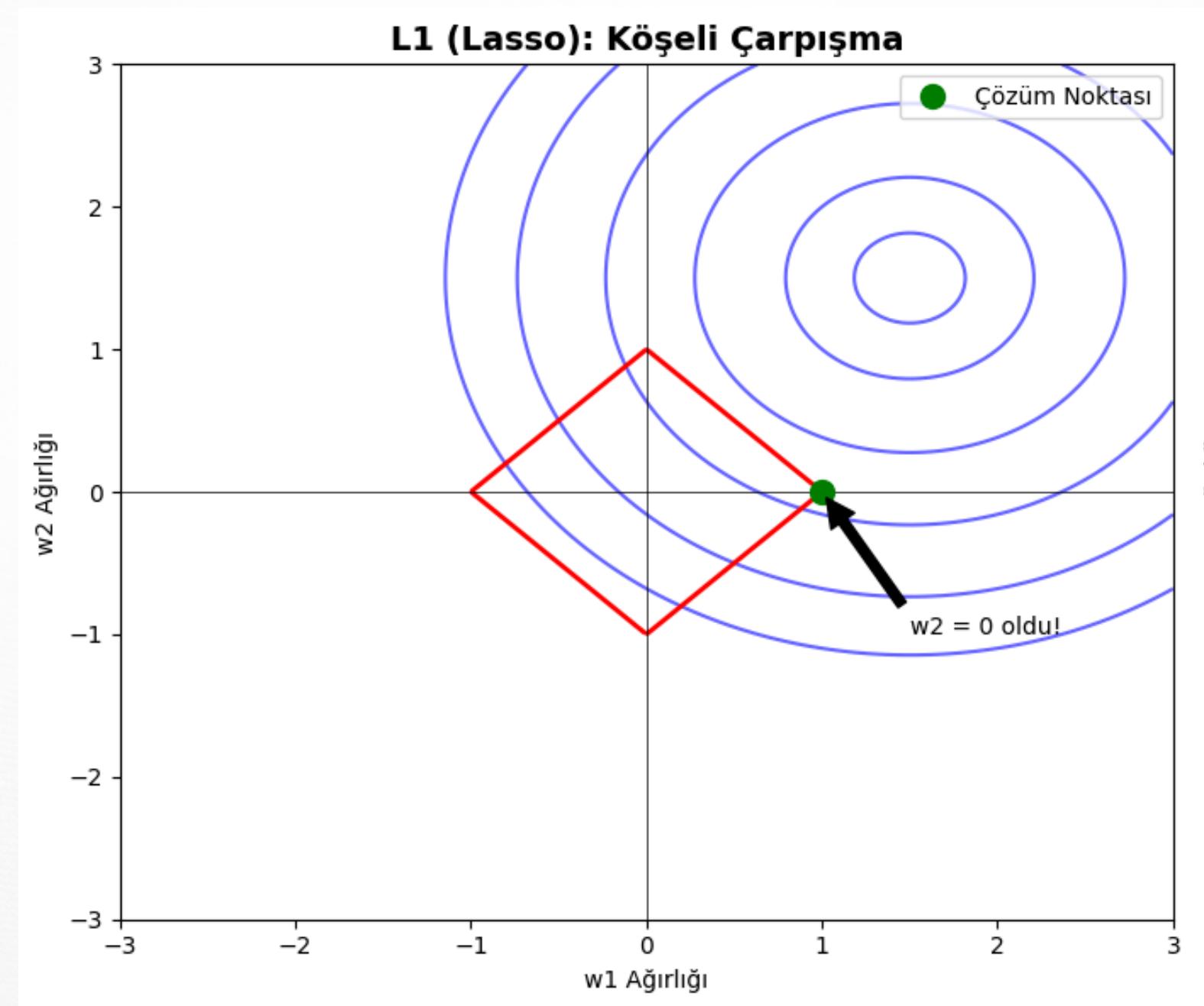


Regularization - Ridge & Lasso

Modelin aşırı karmaşıklaşması (yüksek varyans) ve eğitim verisini ezberlemesi sorununu çözmek için **Düzenlileştirme (Regularization)** teknikleri kullanılır. Regülarizasyon, modelin karmaşıklığını azaltarak ve parametrelerini sınırlayarak aşırı uyum yapmasını engeller.

Model Çok Karmaşıklaşırsa Ne Yaparız? Aşırı uyum durumunda (yüksek varyans), regülarizasyon teknikleri uygulanarak modelin tahminlerinin eğitim verilerine olan hassasiyeti azaltılır. Bu, küçük bir yanlılık (bias) ekleyerek varyansta önemli bir düşüş elde etme ana fikrine dayanır.

Katsayıları Cezalandırma Mantığı: Geleneksel en küçük kareler (OLS) yöntemi, sadece hataların karelerinin toplamını minimize eder. Regülarizasyon ise bu denkleme bir **ceza terimi (penalty term)** ekler. Bu ceza, parametrelerin (β veya A) çok yüksek değerler almamasını engellemeyi amaçlar. Cezanın şiddeti λ (lambda) adı verilen bir hiperparametre ile belirlenir.



L1 Regülarizasyon (Lasso Regresyon)

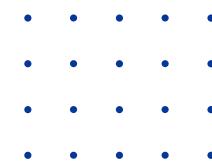
Çalışma Şekli: L1 regülarizasyonu (Lasso Regresyon), hataya katsayılarının mutlak değerlerinin toplamını ekleyerek çalışır.

Lasso'nun Özellik Seçimi (Feature Selection) Yapabilmesi: L1'i Ridge'den ayıran en büyük fark, lambda değeri yeterince artırıldığında, Lasso'nun bazı katsayıları tamamen sıfıra indirebilmesidir.

Katsayısı sıfıra inen bir değişken, denklemden etkili bir şekilde hariç tutulmuş olur.

Bu, Lasso'yu gereksiz veya etkisiz değişkenleri denklemden çıkararak modelin daha basit hale gelmesini sağlayan bir özellik seçimi (feature selection) tekniği yapar.

Özetle, L1 (Lasso) gereksiz özellikleri eleyerek modeli basitleştirir.



L2 Regülarizasyon (Ridge Regresyon)

Çalışma Şekli: L2 regülarizasyonu (Ridge Regresyon), hataya katsayılarının karelerinin toplamını ekleyerek çalışır.

Cezalandırma: L2, tüm parametreleri (katsayıları) küçültür. Lambda değeri arttıkça, bu katsayılar sıfıra yaklaşır.

Parametreler: L2, katsayıları küçültürken, sıfıra tam olarak ulaşmaz. Bu, L2'nin tüm parametreleri koruduğu, ancak etkilerini azalttığı anlamına gelir.

Bölüm 6: Doğrusal Olmayan Diğer Modeller

Karar Ağaçları Regresyonu (Decision Tree Regression)

Şimdiye kadar ele aldığımız Doğrusal Regresyon modelleri (Basit, Çoklu, Polinom), veriler arasında **doğrusal bir ilişki** varsayar. Ancak, makine öğrenmesinde sadece düz çizgilerle değil, karmaşık, eğrisel desenlerle (non-linear relationships) karşılaşırız.

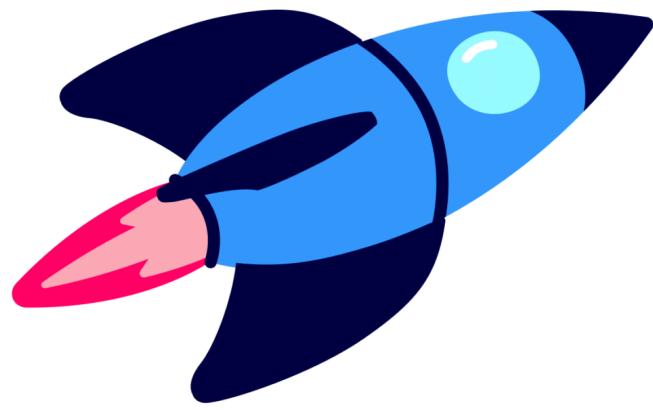
Karar Ağaçları Regresyonu (Decision Tree Regression), hem sınıflandırma hem de **regresyon sorunlarına uygulanabilen** esnek bir denetimli öğrenme yaklaşımıdır.

- **Veriyi Bölerek Tahmin Yapma Mantığı:** Karar ağaçları, akış şemasına benzeyen bir ağaç yapısı kullanır; ana düğümler özellikleri (features), dallar kuralları ve yaprak düğümler ise algoritmanın nihai çıktısını (tahminini) temsil eder.
- Model, tahmin yaparken veriyi bölerek ilerler. **Özellik önemine** göre veriyi en iyi şekilde bölen düğümleri (split) değerlendirir.
- Doğrusal regresyonun aksine, karar ağaçları bağımsız değişkenler arasında belirli bir ilişki varsayımadığı için, **doğrusal olmayan verilerle** (non-linear data) karşılaşlığında (örneğin sinüs dalgası şeklindeki bir veri kümelerinde) Doğrusal Regresyona göre **çok daha düşük hata** (MSE) ile daha iyi uyum sağlayabilir

Bölüm 7: Hangi Model Ne Zaman Kullanılır?

Bir regresyon problemi için en iyi modelin seçimi, **verinin doğası**, **boyutu** ve modelden beklenen **yorumlanabilirlik seviyesine** bağlıdır.

Kriter	Doğrusal Regresyon Linear Regression/OLS)	Regülarizasyon (Ridge/Lasso)	Karar Ağaçları/Topluluk Modelleri (Random Forests)
Veri İlişkisi (Doğrusallık)	İlişkinin doğrusal veya yaklaşık olarak doğrusal olduğu durumlarda.	Doğrusal veya polinom/hafif eğrisel ilişkiler.	İlişkinin doğrusal olmadığı (non-linear) karmaşık verilerde daha iyi performans.
Yorumlanabilirlik İhtiyacı	Yüksek. Katsayılar (A, B_1, \dots) hangi özelliğin ne kadar etkilediğini açıkça gösterir.	Orta/Yüksek. Ridge tüm katsayıları korur, Lasso gereksiz katsayıları sıfırladığı için yorumlamayı basitleştirir (özellik seçimi).	Düşük. Daha az yorumlanabilir, ancak topluluk modelleri (Ensemble) genel hatlarıyla özellik önemini verebilir.
Aşırı Uyma (Overfitting) Riski	Genellikle düşük varyans (daha çok eksik uyma/underfitting riski).	Yüksek varyans sorunu çözümek için idealdir (Model karmaşıklığını ve katsayıları cezalandırır).	Tekil Karar Ağaçları aşırı uyuma duyarlıdır; Topluluk modelleri bu riski azaltır.
Çoklu Bağlantı (Multicollinearity)	Yüksek risk. Birbirine yüksek korelasyonlu değişkenler sorun yaratır.	Çoklu bağlantı sorununu hafifletmeye yardımcı olur.	Daha az etkilenir. Ağaçlar, birbirini tekrar eden özelliklerden sadece birini seçme eğilimindedir.



Notebook Session

Kaynaklar

Derste işlenen notebook linki:

<https://www.kaggle.com/code/eselinildam/simple-linear-polynomial-randomforest-regression>

1. <https://www.kaggle.com/code/farzadnekouei/polynomial-regression-regularization-assumptions>
2. <https://www.kaggle.com/code/jaypradipshah/polynomial-regression-from-scratch>
3. <https://www.kaggle.com/code/alfathterry/simple-linear-regression-from-scratch#1.-Regression-Line-dan-Residual-Analysis>
4. <https://medium.com/data-science/your-guide-to-linear-regression-models-df1d847185db>
5. <https://medium.com/@rndayala/linear-regression-a00514bc45b0>

Sorular ??

