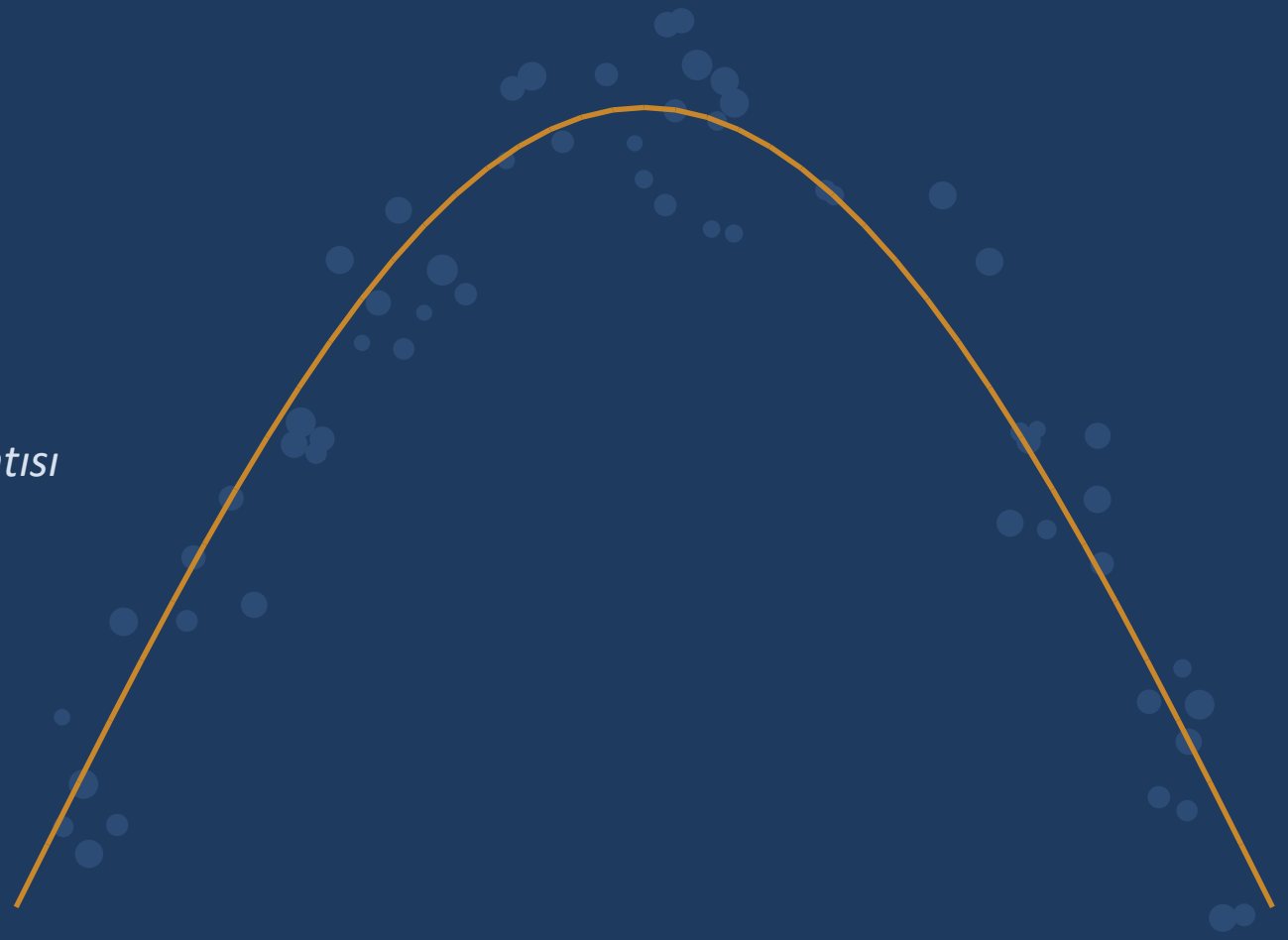
EKONOMİ VE FİNANSTA MAKİNE ÖĞRENMESİ

HAFTA 3

Model Ne Kadar İyi?

Ezberlemek ve öğrenmek

Aşırı uyum, yanlılık-varyans, çapraz doğrulama ve veri sızıntısı



Prof. Dr. Mehmet Songur

Dicle Üniversitesi · İİBF · İktisat Bölümü

Geçen haftadan

Bir problemi makine öğrenmesinin diline çevirdik



Öğrenme türü

Soruyu belirler



Hedef ve öznitelik

Gözlem birimi, y ve X



Veri türü

Kesit, zaman serisi, panel



Veri hazırlama

Eksik, aykırı, ölçek, kodlama

Bu hafta: Modeli kurduk diyelim. Modelin iyi olup olmadığını nasıl anlayacağız?

Bugün ne konuşacağız?

1

Aşırı uyum

Ezberlemek ile öğrenmek arasındaki fark

2

Yanlılık ve varyans

Bir model neden hata yapar?

3

Çapraz doğrulama

Doğru modeli nasıl seçeriz?

4

Veri sızıntısı

İyi görünen ama işe yaramayan modeller

İki öğrenci

Kim daha başarılı olur?



Ayşe

Geçmiş yılların sınav sorularını cevaplarıyla birlikte ezberledi.



Mehmet

Aynı soruları çözerek konunun mantığını öğrendi.

İki öğrenci



Ayşe

Geçmiş yılların sınav sorularını cevaplarıyla birlikte ezberledi.

- **Deneme sınavı (eski sorular): tam puan**
- **Gerçek sınav (yeni sorular): başarısız**



Mehmet

Aynı soruları çözerek konunun mantığını öğrendi.

- **Deneme sınavı (eski sorular): birkaç hata**
- **Gerçek sınav (yeni sorular): başarılı**

Sınıfta	Makine öğrenmesinde
Geçmiş yılların soruları	Eğitim verisi
Bu yılın soruları	Test verisi
Ezberlemek	Aşırı uyum

Bir model, eğitildiği veride değil, hiç görmediği veride sınanır.

1

Model başarısını ölçmek

Eğitim verisi, test verisi ve hata

Örneklem içi uyum neden yetmez?

Ekonometriden bildiğiniz bir gerçek

Regresyona değişken ekledikçe R^2…

- Asla düşmez; alakasız bir değişken bile R^2'yi biraz artırır.
- Gözlem sayısı kadar değişken eklenirse $R^2 = 1$ olur.
- Bu "mükemmel" model yeni gözlemlerde büyük olasılıkla çok kötü öngörür.

Ekonometrideki çözüm

Düzeltilmiş R^2, AIC, BIC: her ek parametreye bir "ceza". ARDL'de gecikme seçimi.

Makine öğrenmesinin çözümü

Başarıyı formülle tahmin etmek yerine, ayrılmış bir veride doğrudan ölçmek.

Veriyi ikiye ayırmak



Eğitim verisi

Model burada kurulur: katsayılar bu veriden kestirilir.

Test verisi

Model kurulurken hiç kullanılmaz. En sonda, yeni gözlemlerdeki başarıyı ölçmek için kullanılır.

Oranlar bir kural değil, bir alışkanlıktır. Ekonometrideki "örneklem dışı öngörü" ile aynı fikir.

Hatayı nasıl ölçeriz?

Sayısal bir hedef için en yaygın ölçü

$$\text{HKO} = (1/n)\,\Sigma\,(y_i - \hat{y}_i)^2 \qquad \text{Kök HKO} = \sqrt{\text{HKO}}$$

HKO: hata kareleri ortalaması (kalıntı kareler toplamının ortalaması)

Eğitim hatası

HKO'nun eğitim verisinde hesaplanmış hâli

Test hatası

Aynı formül, test verisinde. Asıl ilgilendiğimiz ölçü.

Kök HKO

Hedefin kendi biriminde: kök HKO = 1,2 ise öngörüler tipik olarak 1,2 puan sapıyor.

2

Aşırı uyum

Sinyali öğrenmek mi, gürültüyü ezberlemek mi?

Ortak örneğimiz: ücret ve deneyim

Mincer denklemi: log ücret deneyimle önce artar, sonra düzleşir

Yapay bir dünya kurduk

- Gerçek ilişkiyi biz belirledik: ters U biçimli.
- Her bireyin ücretine rastgele bir sapma ekledik.
- Bu dünyadan 30 kişilik bir eğitim örneklemi çektik.
- Testi 2.000 yeni kişiyle yapacağız.

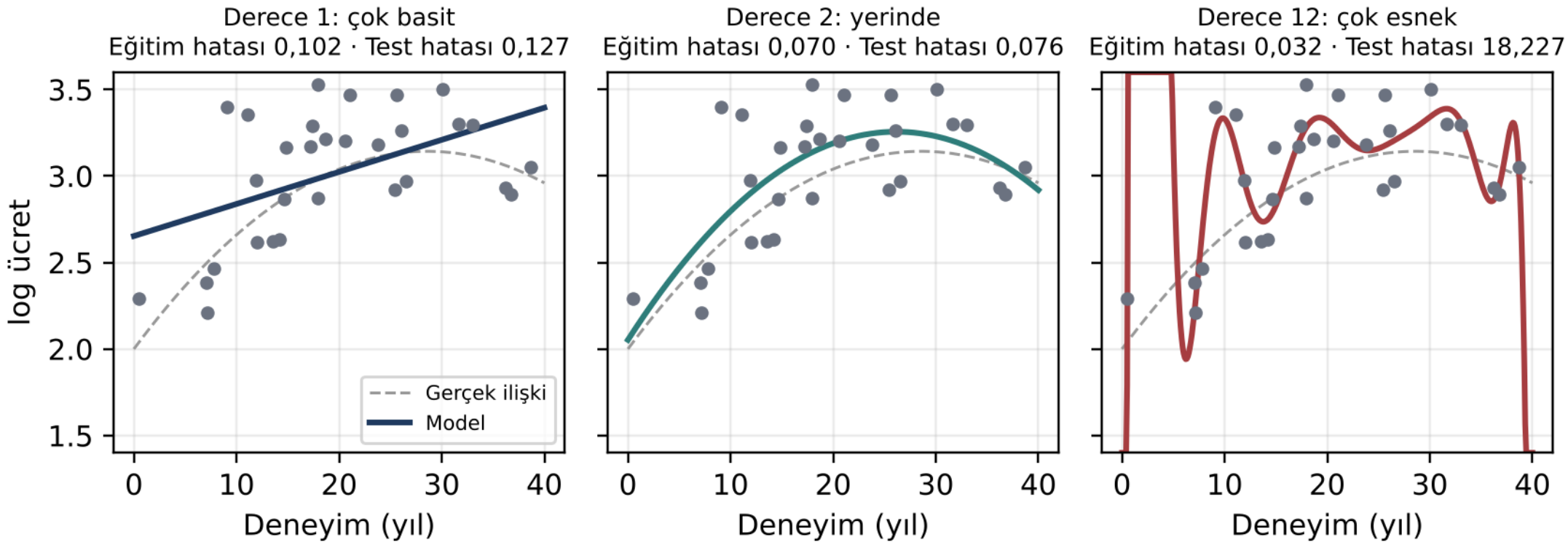
Tek aracımız: polinom regresyon

$$\log \text{ücret} = b_0 + b_1 \cdot d + b_2 \cdot d^2 + \ldots + b_k \cdot d^k$$

(d = deneyim; en küçük karelerle kestirilir)

- Polinom derecesi k, modelin esnekliğini ayarlayan düğmedir.
- k = 1: düz bir doğru. k büyüdükçe eğri daha çok kıvrılabilir.

Aynı 30 gözlem, üç farklı esneklik



Derece 12: eğitim hatası en düşük (0,032), test hatası felaket (18,227). Derece 2: iki hata da düşük ve birbirine yakın.

Eksik uyum ve aşırı uyum

Eksik uyum

Model gerçek ilişkinin şeklini yakalayamayacak kadar basit.

Eğitim hatası yüksek, test hatası yüksek

Derece 1: ters U'yu düz doğruyla anlatmaya çalışmak

Yerinde uyum

Model verideki sinyali yakalıyor.

İki hata da düşük ve birbirine yakın

Derece 2

Aşırı uyum

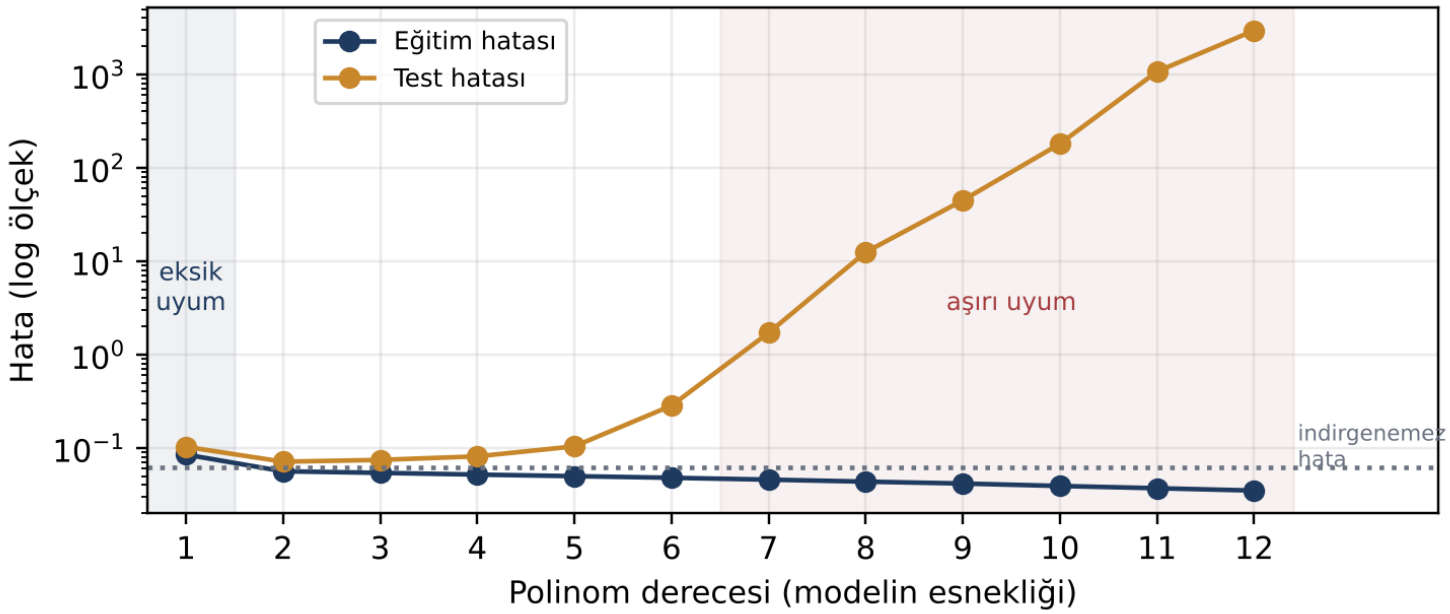
Model sinyalle birlikte gürültüyü de ezberliyor.

Eğitim hatası çok düşük, test hatası çok yüksek

Derece 12: her noktanın yakınından geçmek için kıvrılmak

Esneklik ve hata: U biçimli eğri

Deney 300 kez tekrarlandı; ortalama hatalar (dikey eksen logaritmik)



Eğitim hatası

Esneklik arttıkça hep düşer. Tıpkı R^2 gibi. Model seçimi için kullanılamaz.

Test hatası

Önce düşer, sonra hızla yükselir. En iyi model aradadır: derece 2.

Ekonometriden tanıdık bir sorun



Onlarca spesifikasyon deneyip “en iyi çalışanı” raporlamak

Seçilen model o örneklemin tesadüflerine uyarlanmıştır. Leamer (1983) bu tür aramaların ekonometrik sonuçların güvenilirliğini zayıflattığını eleştirmişti.



Makine öğrenmesinin katkısı: bunu bir kurala bağlamak

Model ne kadar aranırsa aransın, son değerlendirme arama sürecinde hiç kullanılmamış bir veride yapılır.

İki yaygın yanılgı



“Daha karmaşık model her zaman daha iyidir.”

U eğrisi tam tersini gösteriyor. Karmaşıklık, ancak verideki gerçek yapıyı yakalamak için gerektiği kadar faydalıdır.



“Veri büyüdükçe aşırı uyum tamamen kaybolur.”

Azalır ama kaybolmaz. Daha fazla veri daha esnek modellere izin verir; değişken sayısı da artarsa sorun geri gelir.

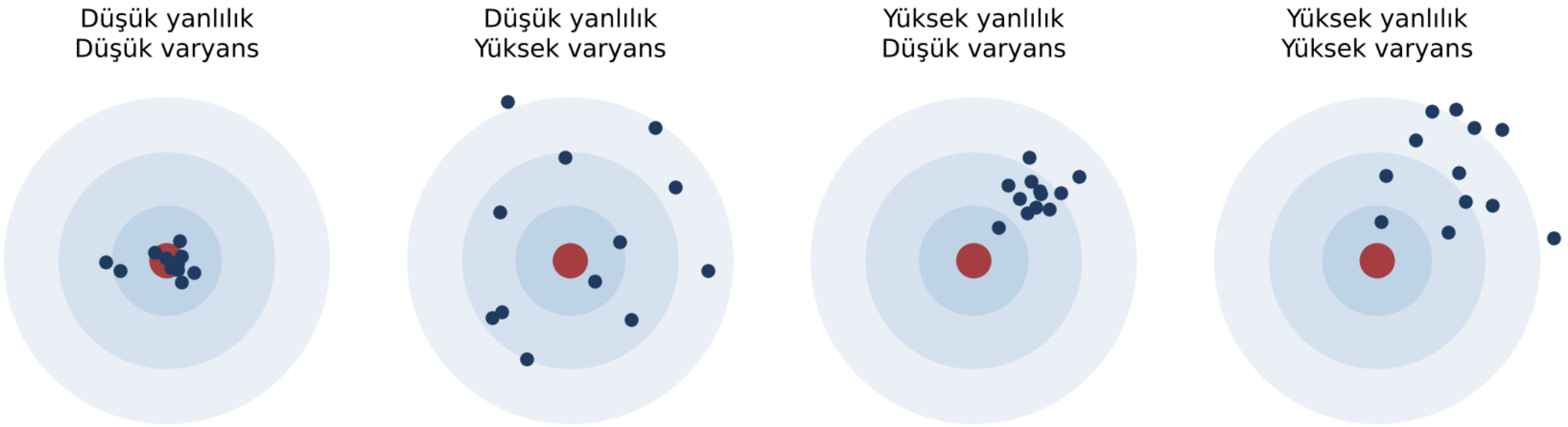
3

Yanlılık ve varyans

Bir model neden hata yapar?

İki hata kaynağı: dart tahtası

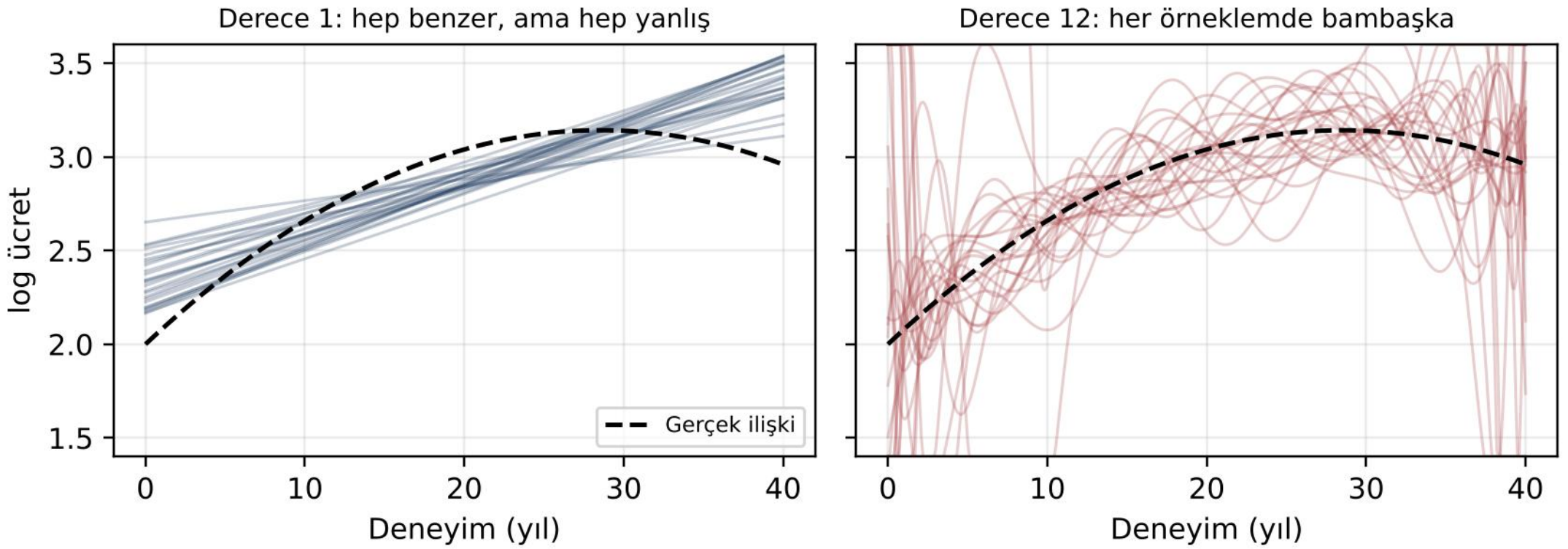
Merkez gerçek değer; her nokta farklı bir örneklemle eğitilmiş modelin öngörüsü



Yanlılık: atışlar merkezden sistematik olarak sapıyor. **Varyans:** atışlar birbirinden dağınık. Ekonometriden tanıdık iki kavram; bu kez öngörüler için.

Aynı fikir, ücret örneğinde

25 farklı 30 kişilik örneklemle eğitilmiş modeller



Derece 1: yüksek yanlılık, düşük varyans

Derece 12: düşük yanlılık, yüksek varyans

Hatanın üç parçası

$$\text{Beklenen test hatası} = \text{Yanlılık}^2 + \text{Varyans} + \text{İndirgenemez hata}$$

Yanlılık^2

Model ortalamada gerçek ilişkiden ne kadar sapıyor?

Esneklik ↑ → azalır

Varyans

Model örneklemden örnekleme ne kadar oynuyor?

Esneklik ↑ → artar

İndirgenemez hata

Hiçbir modelin öngöremeyeceği rastlantısallık

Esneklikten bağımsız; taban

Yanlılık-varyans ödünleşimi: ikisini birden sıfıra indiremeyiz. U eğrisi bu ödünleşimin görüntüsüdür.

İndirgenemez hata: bir alçakgönüllülük dersi

Aynı deneyime sahip iki kişinin ücreti yine de farklıdır.

Simülasyonumuzda bu taban 0,0625. Hiçbir polinom, ne kadar iyi seçilirse seçilsin, test hatasını bunun altına indiremez (U eğrisindeki noktalı çizgi).

İktisatta

Enflasyonu ya da kuru belirleyen faktörlerin bir kısmı (beklenmedik şoklar, siyasi gelişmeler) öngörülemez. Bir modelden bu tabanın altında hata beklemek, modelin değil beklentinin hatasıdır.

Yansız olmak yeterli mi?

Ekonometri ile makine öğrenmesinin farklı öncelikleri

$\hat{\beta}$

Ekonometri

Gauss-Markov: EKK, yansız kestiriciler arasında en küçük varyanslıdır. Katsayının ortalamada doğru olması esastır.

$\hat{y}$

Makine öğrenmesi

Amaç toplam hatayı küçültmek. Biraz yanlılığı kabul edip varyansı çok azaltmak toplam hatayı düşürebilir.

Dart benzetmesiyle: merkezin biraz yanına ama hep aynı yere atan oyuncu, dağınık atan oyuncudan daha çok puan alabilir. Bu fikri kullanan yöntemler: gelecek hafta.

4

Çapraz doğrulama

Doğru modeli seçmek

Test verisi yalnızca bir kez kullanılır

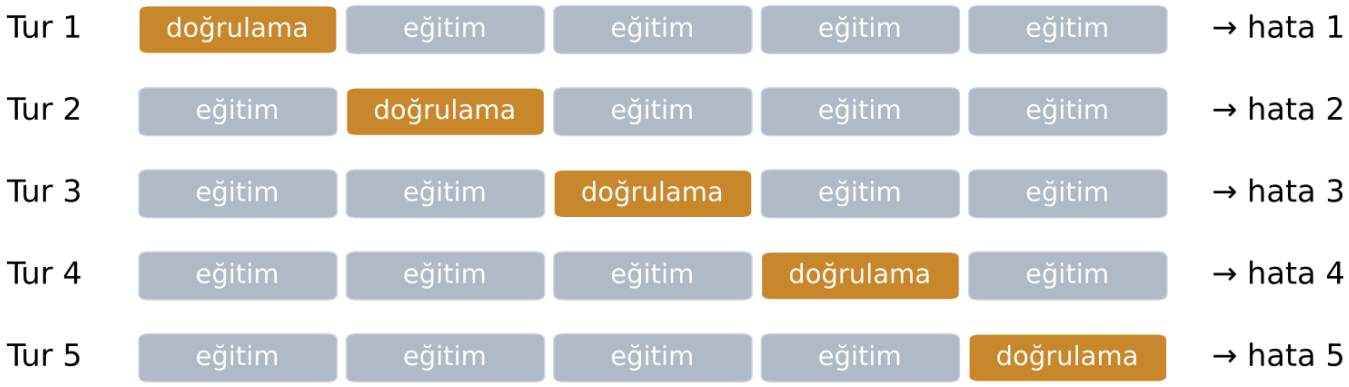
Test verisiyle model seçersek, test artık bağımsız bir sınav değildir

1	**Eğitim verisi**	Her aday modelin katsayılarını kestirmek	*Ders çalışmak*
2	**Doğrulama verisi**	Aday modelleri karşılaştırıp birini seçmek (ör. polinom derecesi)	*Deneme sınavları*
3	**Test verisi**	Seçilen modeli en sonda, bir kez değerlendirmek	*Gerçek sınav*

Polinom derecesi gibi, veriden doğrudan kestirilmeyen ayarlara hiperparametre denir.

k-katlı çapraz doğrulama

30 kişiden 6'sını ayırmak hem eğitimi küçültür hem sonucu şansa bırakır. Çözüm: veriyi dönüşümlü kullanmak.



Çapraz doğrulama hatası = beş hatanın ortalaması

1. Eğitim verisini k parçaya böl (k = 5 ya da 10).
2. Her turda bir parçayı ayır, modeli kalanlarla kur, ayrılan parçada hatayı ölç.
3. Hataların ortalaması: çapraz doğrulama hatası.
4. Her aday model için tekrarla; en düşük olanı seç.
5. Seçilen modeli tüm eğitim verisiyle kur, testte bir kez değerlendir.

Ücret örneğinde çapraz doğrulama

30 kişilik eğitim örneklemi, 5-katlı çapraz doğrulama

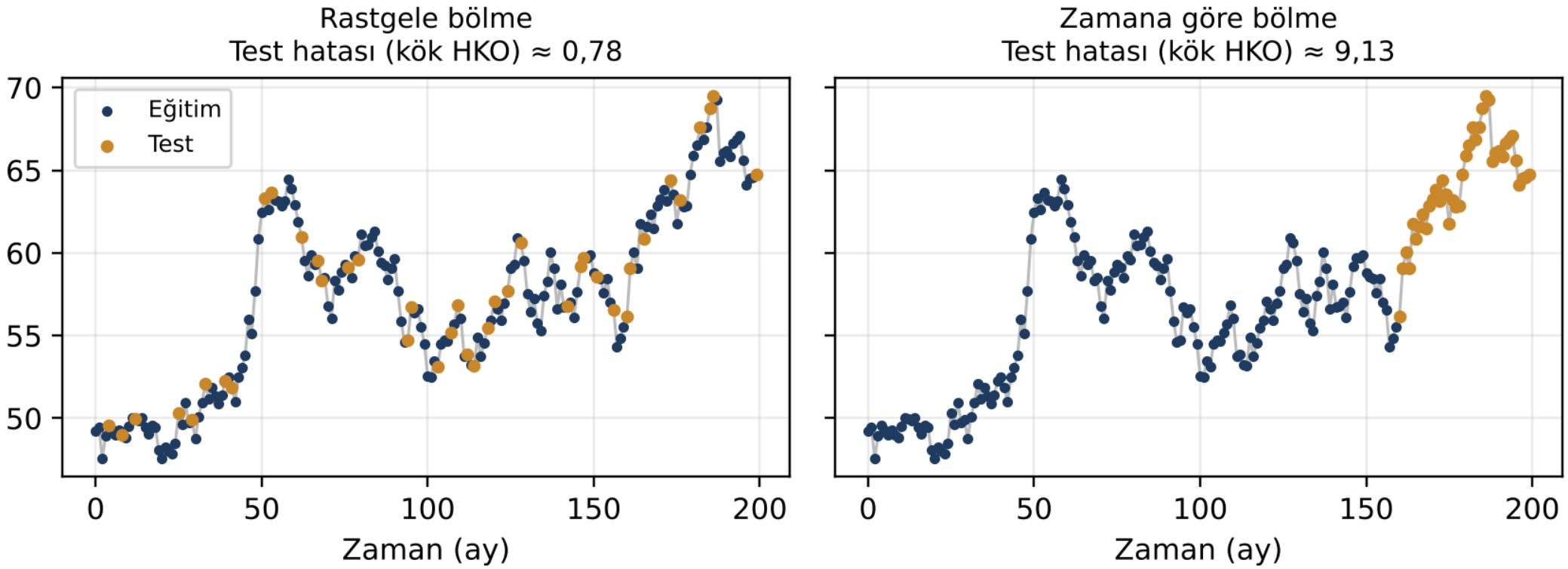
Derece	Eğitim hatası	Çapraz doğrulama hatası	Test hatası (2.000 yeni kişi)
1	0,102	0,140	0,127
2	**0,070**	**0,075**	**0,076**
3	0,069	0,118	0,078
4	0,068	0,134	0,077
5	0,068	0,418	0,077
6	**0,066**	2,644	0,079

Çapraz doğrulama teste hiç bakmadan derece 2'yi seçti.

Eğitim hatasına bakan derece 6'yı seçerdi. Not: küçük örneklemde çapraz doğrulama gürültülüdür.

Zaman serisinde rastgele bölme geleceği gösterir

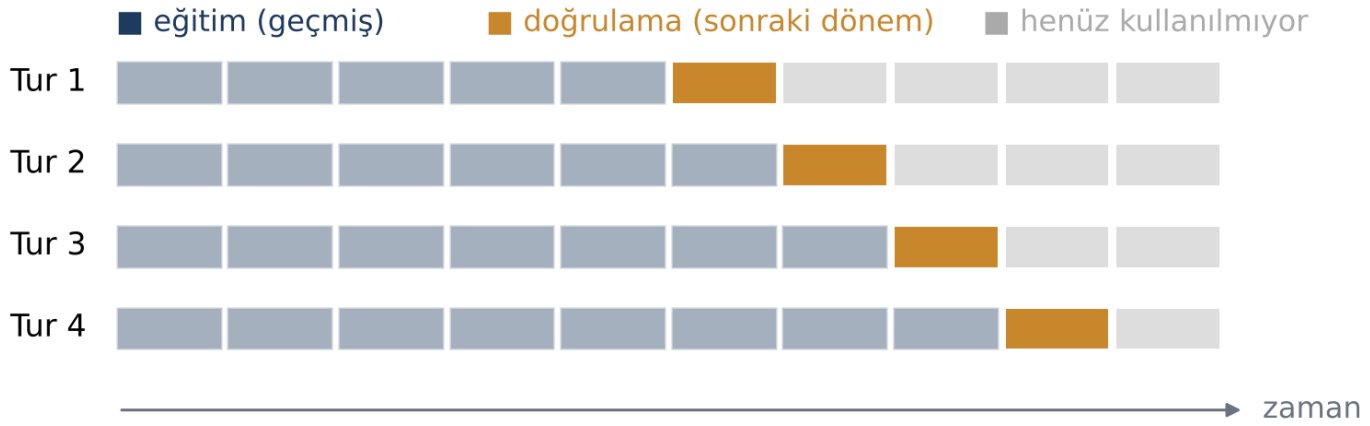
200 aylık yapay bir seri, aynı basit model, iki farklı bölme



Solda model test ayının sonrasını da görmüş: hata 0,78, sağda 9,13. Yaklaşık 12 kat daha başarılı görünüyor; gerçekte gelecek elimizde yoktur.

Zaman serisinde doğrulama: genişleyen pencere

Her turda yalnızca geçmişle eğit, bir sonraki dönemi doğrula



Neden?

Öngörü gerçek hayattaki gibi yapılır: her adımda yalnızca o ana kadar gözlenmiş veri kullanılır.

Ekonometrideki “yinelemeli örneklem dışı öngörü” ile aynı fikir.

Kural: zaman serisinde geçmişle eğit, gelecekle doğrula.

Panelde katları nasıl oluştururuz? Soruya bağlı

81 ilin 2004–2023 verisiyle bir model kuruyoruz

Soru: Gelecek yılları öngörmek

Zamana göre: genişleyen pencere



Soru: Hiç görülmemiş bir ili öngörmek

İllere göre: bir ilin tüm yılları aynı katta



Aynı ilin yakın yıllarını rastgele katlara dağıtmak başarıyı abartır: model test edilen ilin önceki ve sonraki yılını görmüştür.

5

Veri sızıntısı

İnanılmayacak kadar iyi görünen modeller

Veri sızıntısı nedir?

Modelin eğitim sırasında, gerçek kullanımda elinde olmayacak bir bilgiye erişmesi.



Ezberci öğrenci benzetmesiyle

Sınav soruları önceden sızmıştır. Deneme puanı yüksek, gerçek başarı yok.



Belirtisi

Doğrulama ve test başarısı olağanüstü; gerçek hayatta model çöküyor. Bilimsel yayınlarda da yaygın bir sorun.

Sızıntının dört türü



Hedef sızıntısı

Hedeften sonra oluşan bilgi

Temerrüt modelinde “tahsilat birimi aradı mı?”



Ön işleme sızıntısı

Doldurma, ortalama, ölçek tüm veriden

Geliri tüm örneklemin ortalamasıyla doldurmak



Zaman sızıntısı

Gelecekle geçmişi öngörmek

Zaman serisinde rastgele bölme



Birim sızıntısı

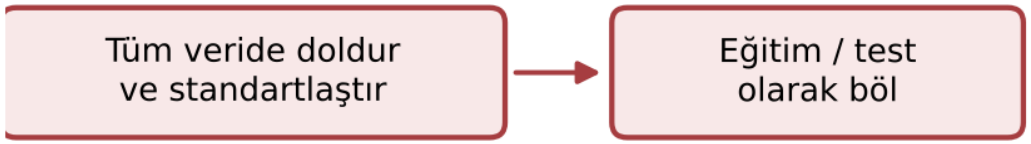
Aynı birim hem eğitimde hem testte

Aynı müşterinin farklı başvuruları iki tarafa düşüyor

Önce böl, sonra hazırla

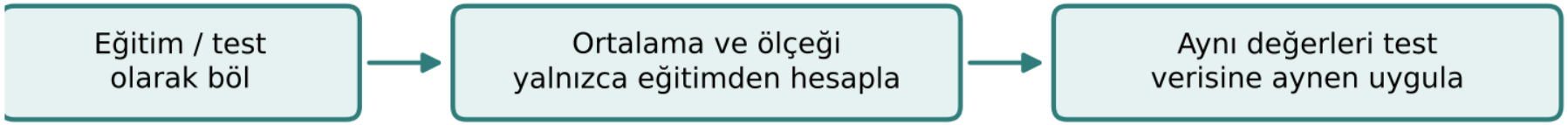
Hafta 2'deki hazırlık adımları da veriden bir şey "öğrenir"

YANLIŞ SIRA



Test verisinin bilgisi eğitime sızdı

DOĞRU SIRA



Doldurma değerleri, ortalamalar ve ölçekler yalnızca eğitim verisinden hesaplanır; test verisine aynen uygulanır. Çapraz doğrulamada her tur için.

İktisatçıya özgü iki tuzak



Veri revizyonları

GSYH ve sanayi üretimi ilk açıklamadan sonra revize edilir. Bugün indirdiğimiz seri, 2019'da gerçekten bilinen değeri değil, düzeltilmiş değeri içerir.

Çözüm: gerçek zamanlı veri setleri (Croushore ve Stark, 2001)



Yayımlanma gecikmesi

Çeyreğin GSYH'si çeyrek bittikten yaklaşık iki ay sonra açıklanır. Çeyrek biter bitmez yapılan bir öngörüde kullanılamaz.

Çözüm: her değişkenin yayımlanma tarihini kontrol etmek

Alıştırma: Sızıntı var mı?

Varsa türü ne?

1. İflas modelinde “bir sonraki yıl denetimde olumsuz görüş”
2. Kur verisi önce standartlaştırılıyor, sonra son 24 ay teste ayrılıyor
3. Aynı evin 2022 ve 2023 satışları rastgele eğitim ve teste dağıtılıyor
4. Enflasyon öngörüsünde yalnızca ay başında yayımlanmış veriler
5. 2015 anlık tahmin modeli bugünkü revize GSYH ile test ediliyor

Cevaplar

İpucu her zaman aynı: tahmin anında bu bilgi elimde olacak mıydı?

1	İflas modelinde “bir sonraki yıl denetimde olumsuz görüş”	**Var: hedef**
2	Kur verisi önce standartlaştırılıyor, sonra son 24 ay teste ayrılıyor	**Var: ön işleme**
3	Aynı evin 2022 ve 2023 satışları rastgele eğitim ve teste dağıtılıyor	**Var: birim**
4	Enflasyon öngörüsünde yalnızca ay başında yayımlanmış veriler	**Yok**
5	2015 anlık tahmin modeli bugünkü revize GSYH ile test ediliyor	**Var: zaman**

Bir bilgi, tahmini yapacağınız anda elinizde olmayacaksa, modelde de olmamalıdır.

Model inanılmayacak kadar iyi görünüyorsa, önce sızıntı arayın.

Bugünün beş mesajı

1. Bir modelin değeri hiç görmediği veride ölçülür; eğitim hatası ve R^2 yanıltır.
2. Çok basit model gerçek ilişkiyi kaçırır, çok esnek model gürültüyü ezberler. Test hatası U biçimlidir.
3. Test hatası = yanlılık² + varyans + indirgenemez hata. Esneklik ↑: yanlılık ↓, varyans ↑.
4. Model doğrulama verisiyle ya da çapraz doğrulamayla seçilir. Test verisi bir kez kullanılır.
5. Tahmin anında elde olmayacak bilgi modele girmez. Önce böl, sonra hazırla.