Yeniden Örnekleme

A. Talha Yalta
TOBB Ekonomi ve Teknoloji Üniversitesi

İKT-457 Ekonomi ve Finans İçin Yapay Zeka 1
Sürüm 0,92 (Güz 2020)

Açık Lisans Bilgisi

A. Talha Yalta
TOBB Ekonomi ve Teknoloji Üniversitesi

Ders Planı

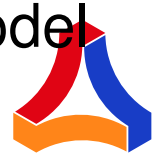
1. Çapraz-Geçerleme
   - Geçerleme-seti yaklaşımı
   - Bir-eksiltmeli çapraz-geçerleme
   - K-kat çapraz-geçerleme
   - Sınıflandırma için çapraz-geçerleme
2. Özyetinim
   - Örnek uygulama
   - Maksimum entropi özyetinimi

Yeniden Örnekleme

- Yeniden örnekleme (resampling), bir örneklem veri setinden yeni örneklemler seçme işlemine denir.
- Bu şekilde tek veri setinden çok sayıda tahminler üretebiliriz.
- Böylece, eğitim verilerini bir kez kullanarak elde edebileceğimizden daha fazla bilgiye ulaşabiliriz.
- Yeniden örnekleme yoğun hesaplama içeren bir işlemdir. Ancak modern bilgisayarlar günlük uygulamalarda yeterli olmaktadır.
- Bu bölümde aşağıdaki iki temel yöntem üzerinde duracağız:
  1. Çapraz-geçerleme
  2. Özyetinim
- Bunlar günümüzde istatistiksel öğrenme için son derece önemli çözümleme yaklaşımlarıdır.

Çapraz-Geçerleme

- Çapraz-geçerleme (cross-validation), eğitim verilerini kullanarak *test hata oranı* tahminleri üreten bir kesinlik ölçüm aracıdır.
- Eğitim hataları ile test hatalarını 2. Bölümde tartışmıştık.
- Belli bir istatistiksel öğrenme yöntemini eldeki eğitim verileriyle kullanınca ortaya çıkan hata oranına eğitim hata oranı denir.
- Test hata oranı ise elimizde olmayan, yeni veriler kullandığımız zaman göreceğimiz hata oranıdır.
- Eğitim hata oranı her zaman düşük çıkar ve yanıltıcıdır. Ancak genellikle elimizde test verileri de yoktur. Bu yüzden test hata oranını bilmek çoğu zaman olanaksızdır.
- İşte, çapraz-geçerleme bu noktada yararlı bir tahmin aracı sunar.
- Bu aracı hem model seçimi hem de model değerlendirmesi için kullanabiliriz.
- Model seçimi (model selection), bir istatistiksel öğrenme yöntemi için en uygun esnekliği belirlemektir. Model değerlendirme (model assessment) ise tahminlerin kesinlik performansını ölçmektir.



Ders Planı

Geçerleme-Seti (1)

- İlk olarak, geçerleme-seti (validation-set) adı verilen görece basit çapraz-geçerleme yöntemini ele alalım.
- Bu yöntem eldeki verileri test seti (test set) ve geçerleme seti (validation set) şeklinde iki parçaya bölmeye dayanır. Bölme işlemi rastsal ya da kuralsal olabilir.
- Eğitim seti kullanılarak model tahmin edilir ve geçerleme setine bakarak hata oranı ölçülür.
- Eğer Y değişkeni nicel ise hata ölçütü olarak genellikle hata kareleri ortalaması (mean squared error) ya da kısaca HKO (MSE) değeri kullanılır:

$$\text{HKO} = \frac{1}{n}\sum(y_i - \hat{f}(x_i))^2 = \frac{KKT}{n}$$

- Konuyu açıklayabilmek için biz şimdilik Y'nin nicel olduğunu varsayacağız.

Geçerleme-Seti (2)

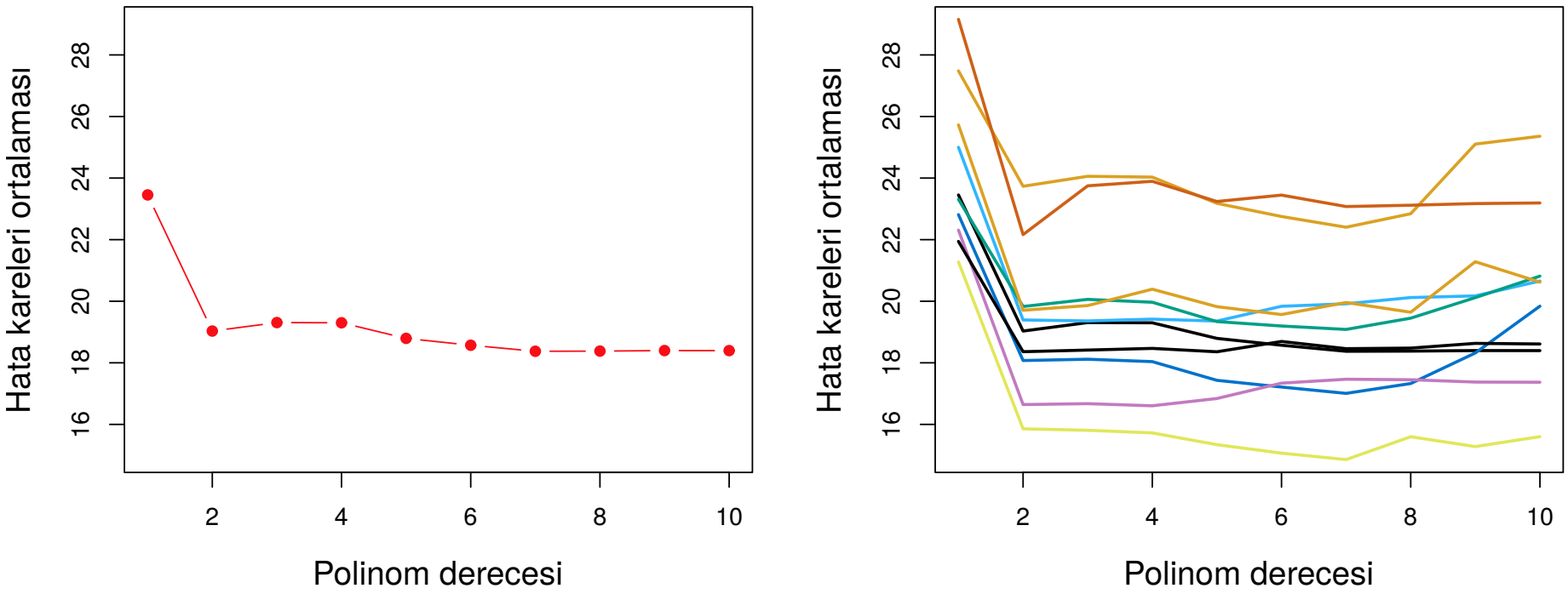
- Geçerleme-seti yaklaşımını örnek üzerinde görmek için 3. Bölümdeki otomobil verilerini kullanalım.
- Anımsayacağınız gibi bu veri setinde yakıt tüketimi ile motor gücü arasında doğrusal-dışı bir ilişki vardı. Bu yüzden 2. derece polinom regresyonu kullanmak görece iyi sonuçlar vermişti.
- Daha yüksek dereceli bir polinom daha da iyi sonuçlar verebilir.
- Geçerleme-seti kullanarak en uygun esneklik derecesini belirleyebiliriz. Bu işlemin adımları aşağıdaki gibidir:

1. Eldeki gözlemler iki parçaya bölünür.
2. Birinci parça eğitim seti olur. Bununla farklı derecelerde polinom regresyonları tahmin edilir.
3. İkinci parça geçerleme seti olarak kullanılarak her bir modele ait HKO değerleri hesaplanır ve en iyi model seçilir.

- Bu şekilde elde ettiğimiz sonuçlar Şekil 1'de verilmiştir.

Örnek Geçerleme-Seti Tahmini

- Sol panelde geçerleme-seti uygulaması 2. derece polinomu önermektedir. Sağda yöntem yeniden örnekleme yapılarak tekrar edilmiştir. Burada varyans yüksek olsa da genel sonuç aynıdır.



Şekil 1: Yakıt tüketimi ile motor gücü modeline ait geçerleme-seti tahmini

Geçerleme-Setinin Sakıncaları

- Bu örnekte geçerleme-seti uygulaması bize en iyi modelin 2. derece polinom olduğu bilgisini vermiştir. Daha yüksek polinomlar hata oranında fazla iyileşme sağlamamaktadır.
- Yöntem yararlı olduğu gibi uygulaması da oldukça kolaydır.
- Ancak iki noktaya dikkat etmek önemlidir:
  1. Şekilde sağ panelde de görüldüğü gibi sonuçlar fazlaca değişkenlik göstermektedir.
  2. Elimizdeki gözlemlerin yarısını kullanmadığımız için hesapladığımız hata oranı gerçek değerden yüksek olabilir.
- Sonuç olarak, geçerleme-seti hem yüksek varyans hem de yüksek yanlılığa sahiptir.
- Dolayısıyla bu sakıncaları azaltmaya yönelik almaşık çapraz geçerleme yöntemleri geliştirilmiştir.

Ders Planı

Bir-Eksiltmeli Çapraz-Geçerleme

- Geçerleme-setine benzer bir diğer yöntem bir-eksiltmeli çapraz-geçerleme (leave-one-out cross-validation) yaklaşımıdır.
- Bu yöntemin adımları aşağıdaki gibidir:

1. Örneklemden yalnızca bir gözlem çıkarılır.
2. Kalan $n-1$ gözlem eğitim seti olarak kullanılır.
3. Çıkartılan tek gözlem ile HKO hesaplanır.
4. İşlem tüm gözlemler için tekrar edilir
5. Tüm HKO'ların ortalaması alınır: $\text{ÇG}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{HKO}_i$

- Bir-eksiltmeli çapraz-geçerleme her türlü model tahmininde kullanılabilen genel bir yaklaşım sunar.
- Yöntemin önemli bir üstünlüğü eğitim verilerinin neredeyse tamamı kullanıldığı için yanlılığın neredeyse minimum olmasıdır.
- Ayrıca işlem tektip olduğu için sonuç her seferinde aynı çıkar.

Bir-Eksiltmeli ÇG'nin Sakıncaları

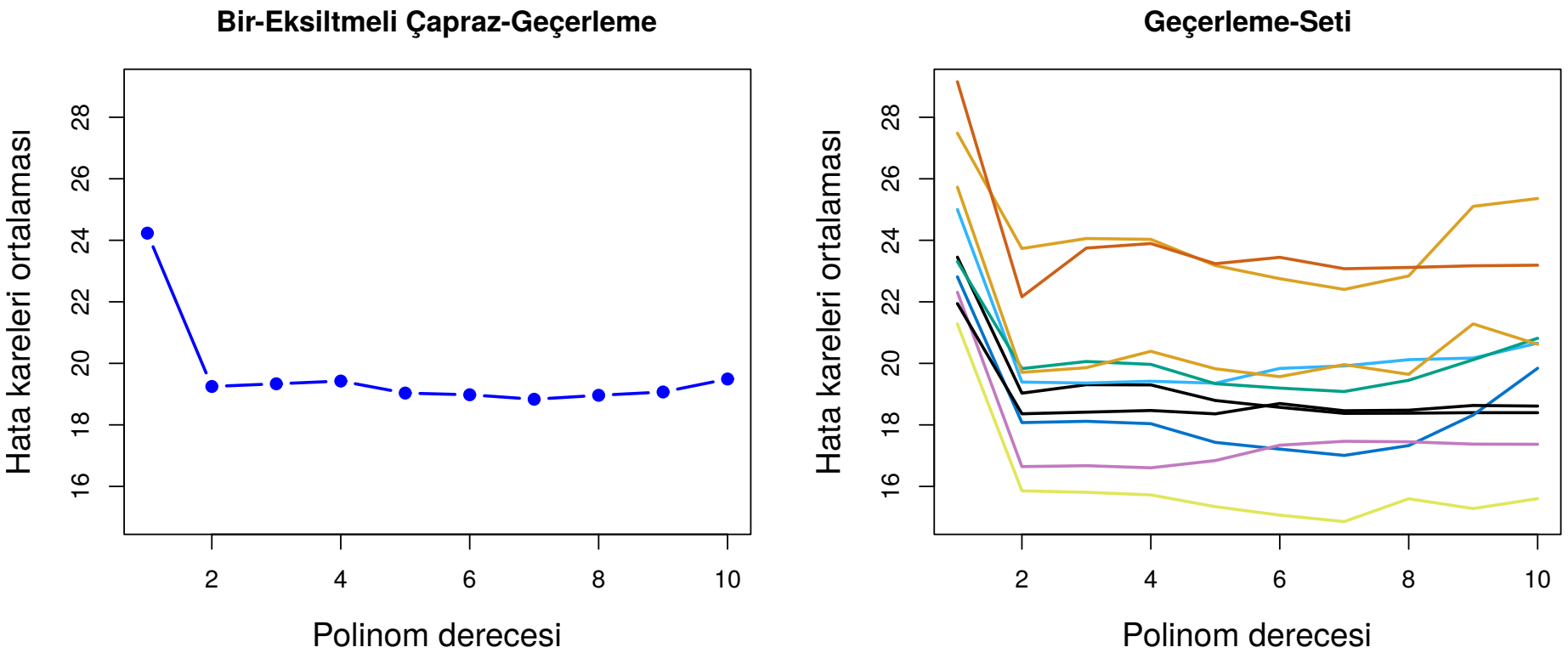
- Bir-eksiltmeli çapraz-geçerlemede yanlılık düşük olmakla birlikte varyans yine de yüksektir.
- Bunun nedeni eğitim setlerinin bir gözlem hariç birbiriyle aynı olmasıdır. Geçerleme için kullanılan tek gözlem ise sürekli değiştiği için yanlılık yüksek çıkar.
- Ayrıca n adet model tahmin edildiği için hesaplama yükü fazladır. Bu durum doğrusal-dışı ileri yöntemlerde sorun yaratabilir.
- Öte yandan doğrusal ve polinom modellerde yöntemi tek seferde hesaplamayı sağlayan aşağıdaki kısayol formülü bulunmaktadır:

$$\text{ÇG}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

- Yukarıda h_i, 3. Bölümde gördüğümüz kaldıraç istatistiği değeridir.
- Her zaman $[0, 1]$ aralığında olan kaldıraç etkisi 1'e yaklaştıkça payda da 0'a gider. Böylece, dışadüşen gözlemler ÇG hata oranını yükseltir.

Örnek Bir-Eksiltmeli ÇG Tahmini

- Yakıt tüketimi ile motor gücü modeline ait bir-eksiltmeli ÇG tahminleri Şekil 2'de sağ panelde verilmiştir. Sol panelde ise yüksek varyanslı geçerleme-seti tahminleri görülmektedir.



Şekil 2: Bir-eksiltmeli ÇG ile geçerleme-seti yönteminin karşılaştırılması

Ders Planı

K-Kat Çapraz-Geçerleme

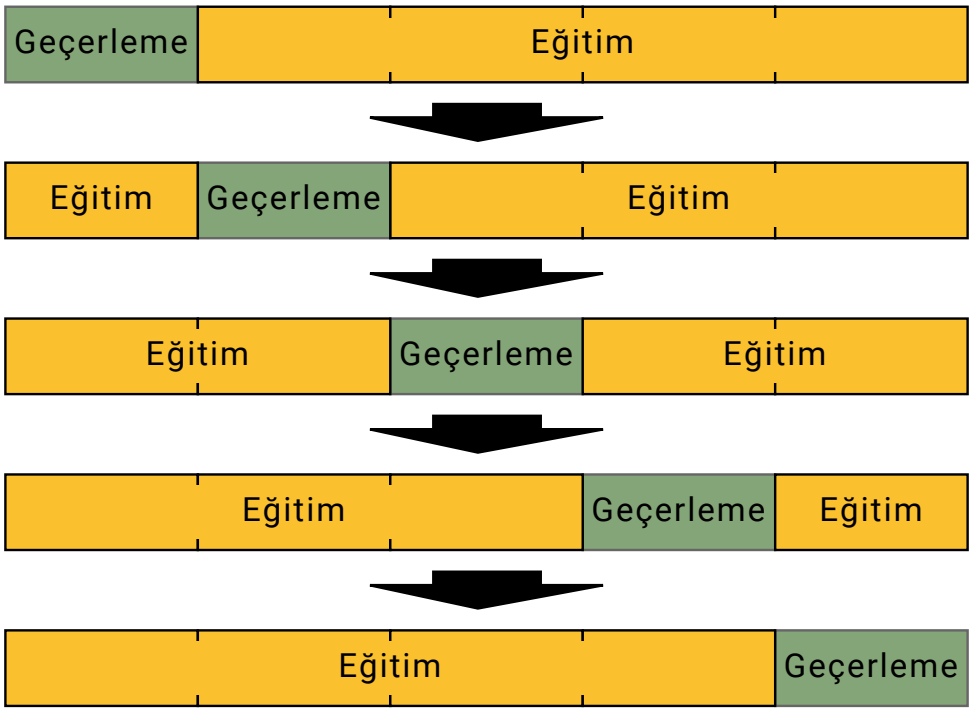
- Varyans-yanlılık ödünleşmesi dikkate alınınca en uygun ve en çok tercih edilen geçerleme yaklaşımı k-kat çapraz-geçerleme (k-fold cross-validation) olarak görülmektedir.
- Bu yöntemin adımları aşağıdaki gibidir:

1. Örneklem eşit büyüklükte k adet parçaya bölünür.
2. Parçalardan biri geçerleme için ayrılır.
3. Kalan $k-1$ parça topluca eğitim için kullanılır.
4. Parçalar değiştirilerek işlem k kez tekrar edilir.
5. HKO'ların ortalaması alınır: $\text{ÇG}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{HKO}_i$

- Eğer $k = n$ olursa yöntem bir-eksiltmeli ÇG'ye dönüşür. Diğer bir deyişle bir-eksiltmeli ÇG aslında k-kat ÇG'nin özel durumudur.
- Uygulamada k için genellikle 5 ya da 10 değeri kullanılır.
- K-kat çapraz geçerlemenin adımları Şekil 3'te gösterilmiştir.

K-Kat Çapraz Geçerleme Yönteminin Uygulanması

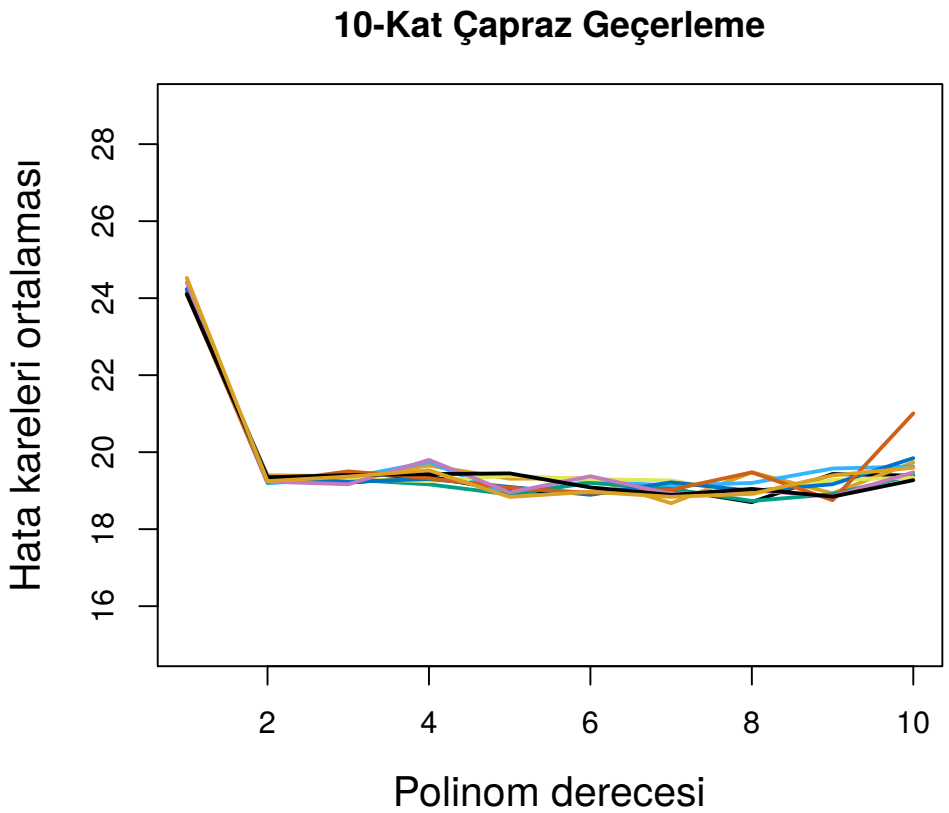
- $K = 5$ için k-kat ÇG yönteminde veri seti önce 5 eşit parçaya bölünür. Parçaların 4'ü ile model tahmin edilip diğer parça ile hata oranı hesaplanır. Bu işlem farklı parçalar ile 5 kez tekrar edilir.



Şekil 3: K-kat çapraz geçerleme yönteminin uygulanması

Örnek *K*-Kat Çapraz Geçerleme Tahmini

- Yakıt ve motor gücü modeline ait 10-kat ÇG tahminleri Şekil 4'te verilmiştir. Rastsal yeniden örneklemeler sonucunda elde edilen farklı tahmin sonuçlarının birbirine yakın olduğuna dikkat ediniz.



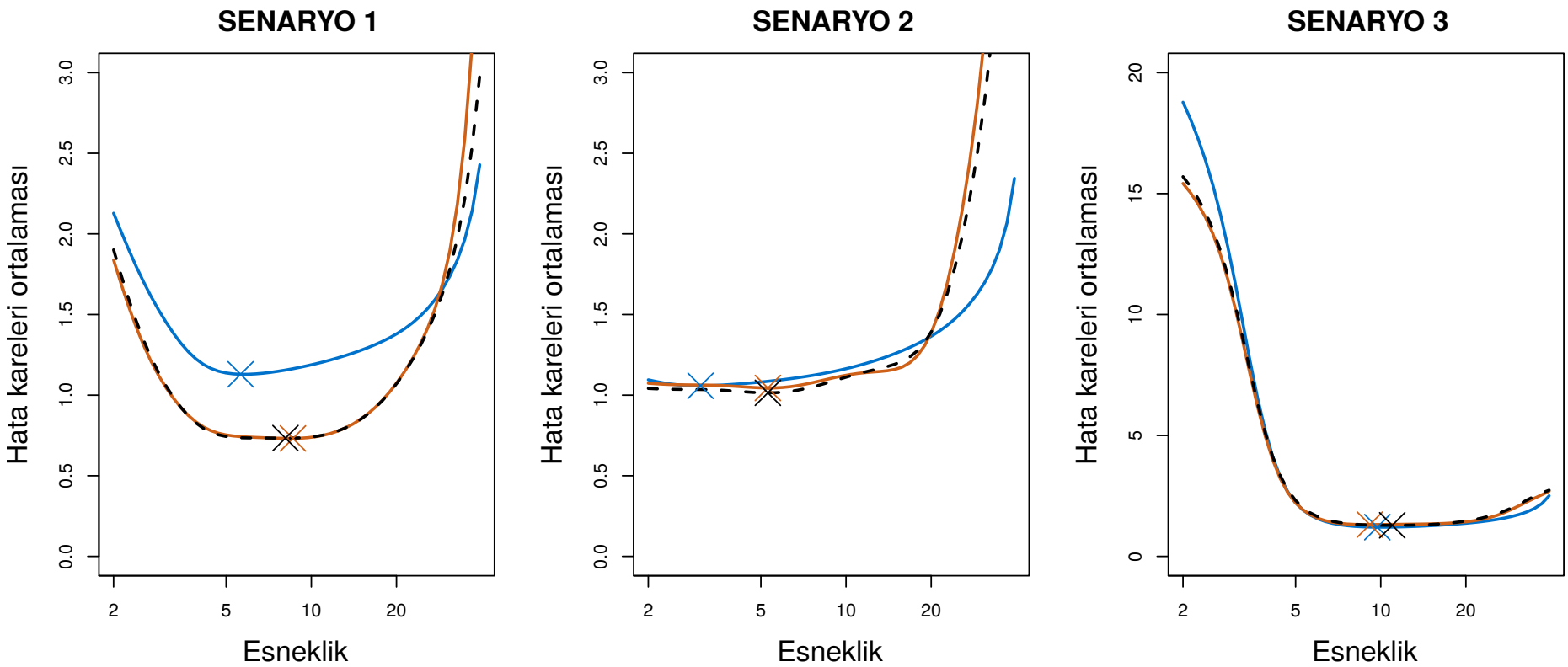
Şekil 4: Yakıt tüketimi ile motor gücü modeline ait *k*-kat ÇG tahminleri

K-Kat ÇG ve Varyans-Yanlılık Ödünleşmesi

- Bir-eksiltmeli ÇG yöntemi toplam n adet model tahmini içerirken k-kat ÇG yalnızca k adet tahmin işlemine gerek duyar.
- Ancak hesaplamasal kolaylık bir yana, k-kat ÇG'nin asıl üstünlüğü varyans-yanlılık ödünleşmesi ile ilgilidir.
- Bir-eksiltmeli ÇG yönteminde bir gözlem hariç tüm eğitim verileri kullanıldığı için yanlılık çok düşüktür. Ancak bu durumda geçerleme işlemine tek bir gözlem kaldığı için varyans yüksek çıkar.
- K-kat ÇG ise test seti ile geçerleme seti büyüklükleri arasında bir denge kurarak uygulamada daha iyi sonuçlar üretir.
- Bu konudaki çeşitli çalışmalar varyans-yanlılık ödünleşmesi açısından $k = 5$ ve $k = 10$ değerlerini kullanmayı önermektedir.
- Bir-eksiltmeli ÇG ile k-kat ÇG yöntemlerinin test hata oranını tahmin etme konusundaki başarısı Şekil 5'te görülmektedir.

K-Kat ÇG ile Bir-Eksiltmeli ÇG'nin Karşılaştırılması

- Şekilde mavi çizgi gerçek test HKO değeri, kırmızı ve siyah çizgiler ise bir-eksiltmeli ÇG ile 10-kat ÇG tahminleridir. İki yöntemin de genel olarak gerçeğe yakın tahminler ürettiğine dikkat ediniz.



Şekil 5: *K*-Kat ÇG ile bir-eksiltmeli ÇG yöntemlerinin kesinliği

Ders Planı

Sınıflandırma İçin Çapraz-Geçerleme (1)

- Şimdiye kadar çapraz-geçerleme yöntemini Y değişkeninin nicel olduğu regresyon bağlamında inceledik.
- Bu doğrultuda test hata oranlarını da HKO ile ölçtük.
- Öte yandan Y'nin nitel olduğu sınıflandırma çözümlemesinde de çapraz-geçerleme son derece kullanışlıdır.
- Burada da yöntem yukarıda gösterdiğimiz şekillerde uygulanabilir. Ancak burada hata oranını ölçmek için yanlış sınıflandırılan gözlem sayısı (number of misclassified observations) kullanılır:

$$\text{CV}_n = \frac{1}{n}\sum_{i=1}^{n} \text{Err}_i$$

- Yukarıda $\text{Err}_i = I(y_i \neq \hat{y}_i)$ şeklinde tanımlı bir $\{0,1\}$ değişkenidir.

Sınıflandırma İçin Çapraz-Geçerleme (2)

- Sınıflandırma çözümlemesinde çapraz-geçerlemeyi açıklamak için farklı derecelerdeki polinom lojistik modellerden yararlanabiliriz.
- Örnek olarak, iki X değişkenli 2. derece lojistik model şöyledir:

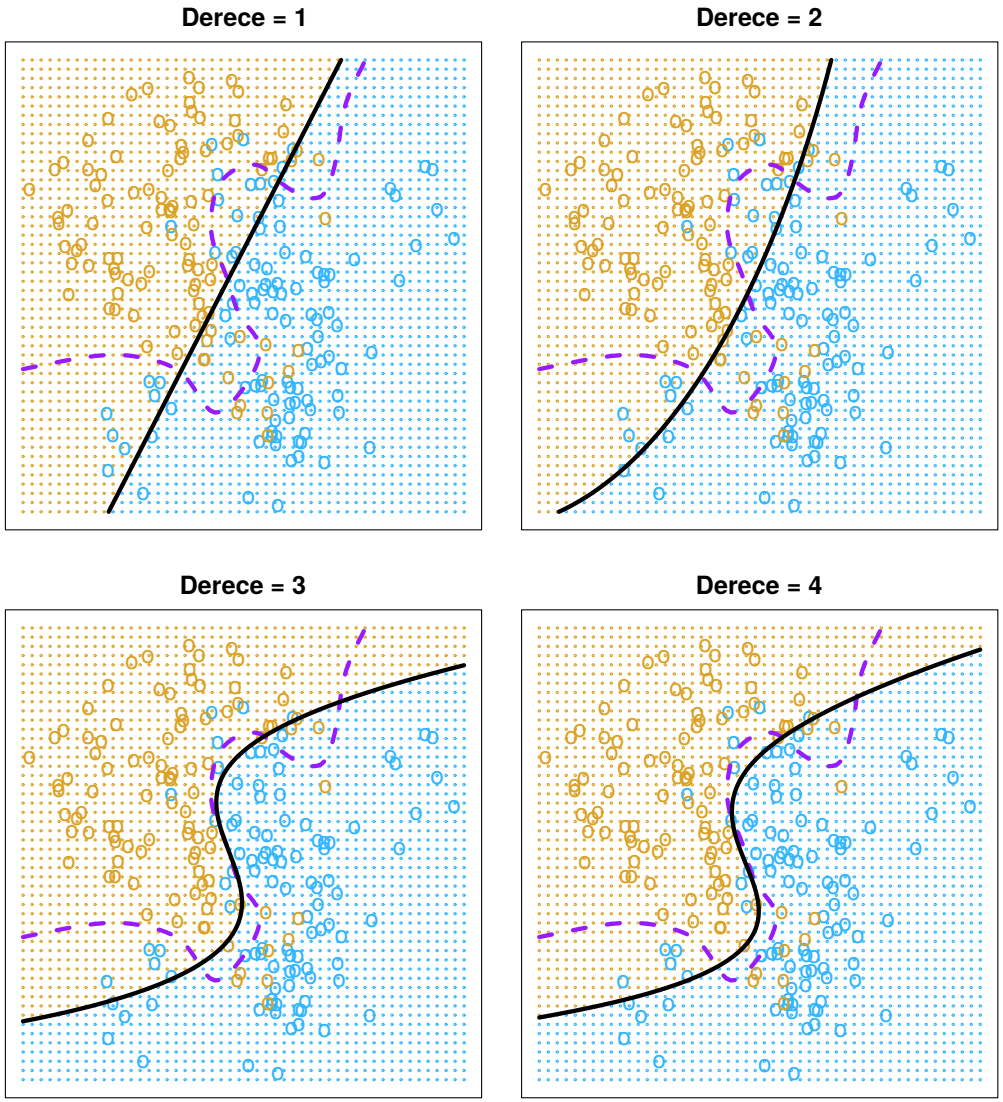
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2$$

- Benzer şekilde 3. derece polinom model de aşağıdaki gibi olur:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1^3 + \beta_6 X_2^3$$

- Bunlar ve daha yüksek derece modeller için test hata oranını tahmin etmek ve en iyi modeli seçmek için çapraz-geçerleme yöntemini kolayca uygulayabiliriz.
- Bunun için 2. Bölümde kullandığımız simülasyon veri setine geri dönelim. Buradaki Bayes karar sınırı ve 4. dereceye kadar lojistik model ÇG karar sınırları Şekil 6'da verilmiştir.

Sınıflandırma İçin Örnek ÇG Tahmini (1)



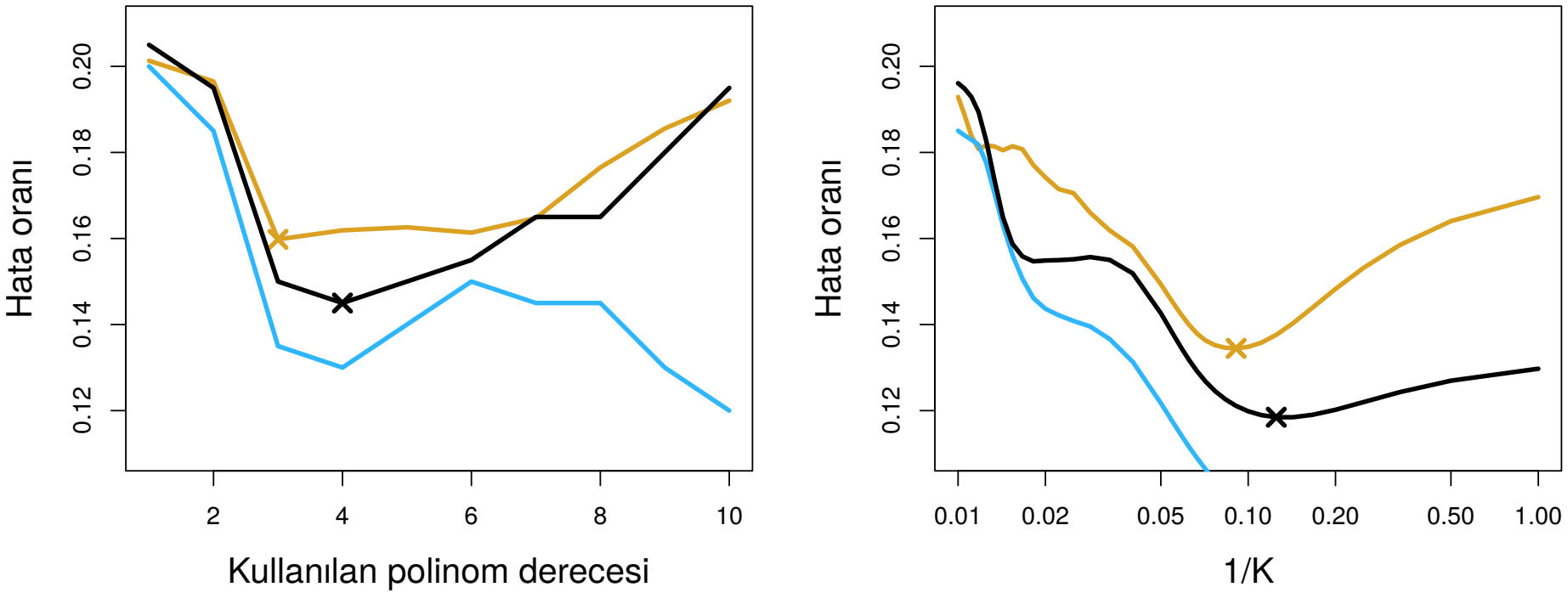
Şekil 6: Farklı derece lojistik model ÇG karar sınırları ve gerçek sınır

Sınıflandırma İçin Örnek ÇG Tahmini (2)

- Yukarıdaki şekilde sol üst paneldeki doğrusal lojistik modelin gerçek karar sınırını iyi tahmin edemediği görülmektedir.
- Sağ üst panelde 2. derece polinom modeli daha başarılı olmakla birlikte yeterince esnek değildir.
- Sol alt köşede ise 3. derece polinom çok daha iyi bir karar sınırı vermektedir. Öte yandan sağ alt panelde görülen 4. derece polinom hata oranını daha da fazla düşürememiştir.
- Şekillere bakıldığında en iyi model 3. derece polinomdur.
- Ancak uygulamada Bayes karar sınırını ve gerçek test hata oranını bilmek olanaksızdır. Bu durumda en iyi modeli seçmek için çapraz geçerleme yönteminden yararlanabiliriz.
- Bu örnekteki dört farklı modele ait gerçek ve eğitim hata oranlarının yanı sıra 10-kat ÇG hata oranı tahminleri Şekil 7'deki gibidir.

K-Kat ÇG ile Bir-Eksiltmeli ÇG'nin Karşılaştırılması

- Şekilde gerçek hata oranı turuncu, eğitim hata oranı mavi ve 10-kat ÇG hata oranı siyah renkle gösterilmiştir. Her iki panelde de ÇG yönteminin genel olarak iyi tahminler ürettiği görülmektedir.



Şekil 7: Sınıflandırma çözümlemesinde ÇG yönteminin kesinliği

Ders Planı

1. Çapraz-Geçerleme
   - Geçerleme-seti yaklaşımı
   - Bir-eksiltmeli çapraz-geçerleme
   - K-kat çapraz-geçerleme
   - Sınıflandırma için çapraz-geçerleme
2. Özyetinim
   - Örnek uygulama
   - Maksimum entropi özyetinimi

Özyetinim

- Günümüzde son derece yaygın ve güçlü bir diğer yeniden örnekleme yaklaşımı özyetinim (bootstrap) yöntemidir.
- Özyetinim farklı şekillerde yapılabilmekle birlikte temel olarak aşağıdaki adımları içerir:

1. Eldeki veri seti yeniden örnekleme yoluyla çoğaltılır.
2. Üretilen örneklemlerden çok sayıda tahmin hesaplanır.
3. Bu tahminler birleştirilerek ilgi duyulan değere ilişkin yüksek kesinlikli güven aralıkları ve kestirimler elde edilir.

- Görüldüğü gibi özyetim yoğun hesaplamaya dayalıdır. Ancak modern bilgisayarlar bu yöntemi uygulamayı kolaylaştırmıştır.
- Özyetinimin önemli bir üstünlüğü başka türlü hesaplanamayan ya da hesaplaması zor olan tahmin ve istatistiklere kolayca uyarlanabilen son derece esnek bir yöntem olmasıdır.

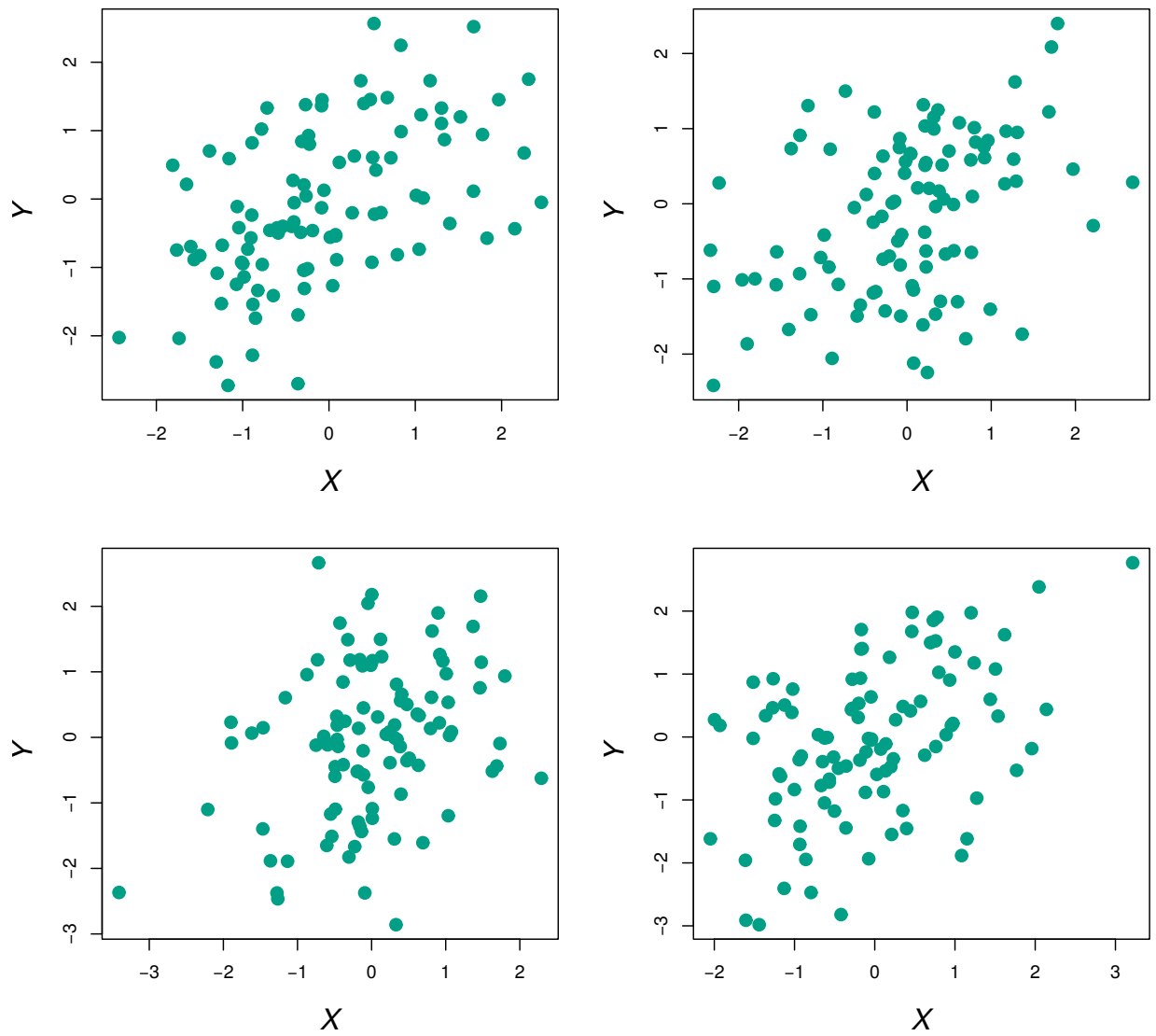
Portföy Seçimi Uygulaması (1)

- Gelin, özyetinimi basit bir örnek yardımıyla açıklayalım.
- Elimizdeki bir miktar parayı α oranında X ve $(1-\alpha)$ oranında Y finansal araçlarında değerlendirmek istediğimizi varsayalım.
- Yatırımımızı planlarken de toplam risk ya da varyansı minimize edecek şekilde α'yı belirleyeceğimizi düşünelim.
- Burada $\text{var}(\alpha X + (1-\alpha)Y)$ değerini minimize eden α şudur:

$$\alpha = \frac{\text{var}(Y) - \text{cov}(X,Y)}{\text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y)}$$

- Yukarıdaki hesaplamayı yapabilmek için çok sayıda bağımsız örnekleme ihtiyaç duyarız. Ancak getiriler zamana bağlı olduğu için yeterince gözlem elde etmek zordur.
- Özyetinimin sağlayacağı yararı görmek için önce simülasyon verileri kullanalım. X ve Y için kendimiz varyans ve kovaryans formülleri belirleyerek istediğimiz kadar örneklem üretebiliriz.
- Bu şekilde üretilen 4 adet örnek veri seti Şekil 8'de gösterilmiştir.

Portföy Seçimi Uygulaması (2)



Şekil 8: Simülasyon yoluyla oluşturulan örnek veri setleri

Portföy Seçimi Uygulaması (3)

- Bu simülasyon örneğinde şekildeki gibi veri setlerini 1000 kez üreterek 1000 farklı $\hat{\alpha}$ tahmini elde edebiliriz.
- Daha sonra elimizdeki çok sayıda yansız tahmini kullanarak $\hat{\alpha}$'ların ortalama değerini buluruz:

$$\bar{\hat{\alpha}} = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\alpha}_i = 0{,}5996$$

- Burada α için başta belirlenen gerçek değer 0,6 idi. Tahmin edilen ortalamanın buna ne kadar yakın olduğuna dikkat ediniz.
- Şimdi, elimizdeki 1000 adet veri setini kullanarak $\hat{\alpha}$'ya ait ölçünlü hatayı da aşağıdaki gibi hesaplayabiliriz:

$$\text{ÖH}(\hat{\alpha}) = \sqrt{\frac{1}{1000-1} \sum_{i=1}^{1000} (\hat{\alpha}_i - \bar{\hat{\alpha}})^2} = 0{,}083$$

- Böylece, $\hat{\alpha}$'nın kesinliği konusunda bilgi sahibi oluruz.

Portföy Seçimi Uygulaması (4)

- Gerçek yaşam kontrolümüzdeki bir simülasyon olmadığı için yukarıda açıkladığımız işlemi uygulamada yapmak olanaksızdır.
- Bu durumda özyetinim bize ideale yakın bir tahmin yöntemi sunar.
- Bunun için yukarıdaki gibi yeni veri setleri oluşturulur. Ancak bu sefer veriler formülden hesaplanmaz. Bunun yerine eldeki tek eğitim veri setinden *rastsal yeniden örnekleme* yoluyla üretilir.
- Örnek olarak, n büyüklüğündeki veri setinden n adet gözlem çekilir. Ancak her seferinde çekilen değer yerine geri koyulur.
- Dolayısıyla oluşturulan özyetinim veri setinde bazı gözlemler birden fazla kez yer alırken bazı gözlemler de hiç yer almayabilir.
- Yukarıdaki işlem yinelenerek topluluk (ensemble) adı verilen Ω adet özyetinim çoğaltması üretilir.
- Uygulamada Ω için genellikle 999 ya da 99 değerleri kullanılır.

Portföy Seçimi Uygulaması (5)

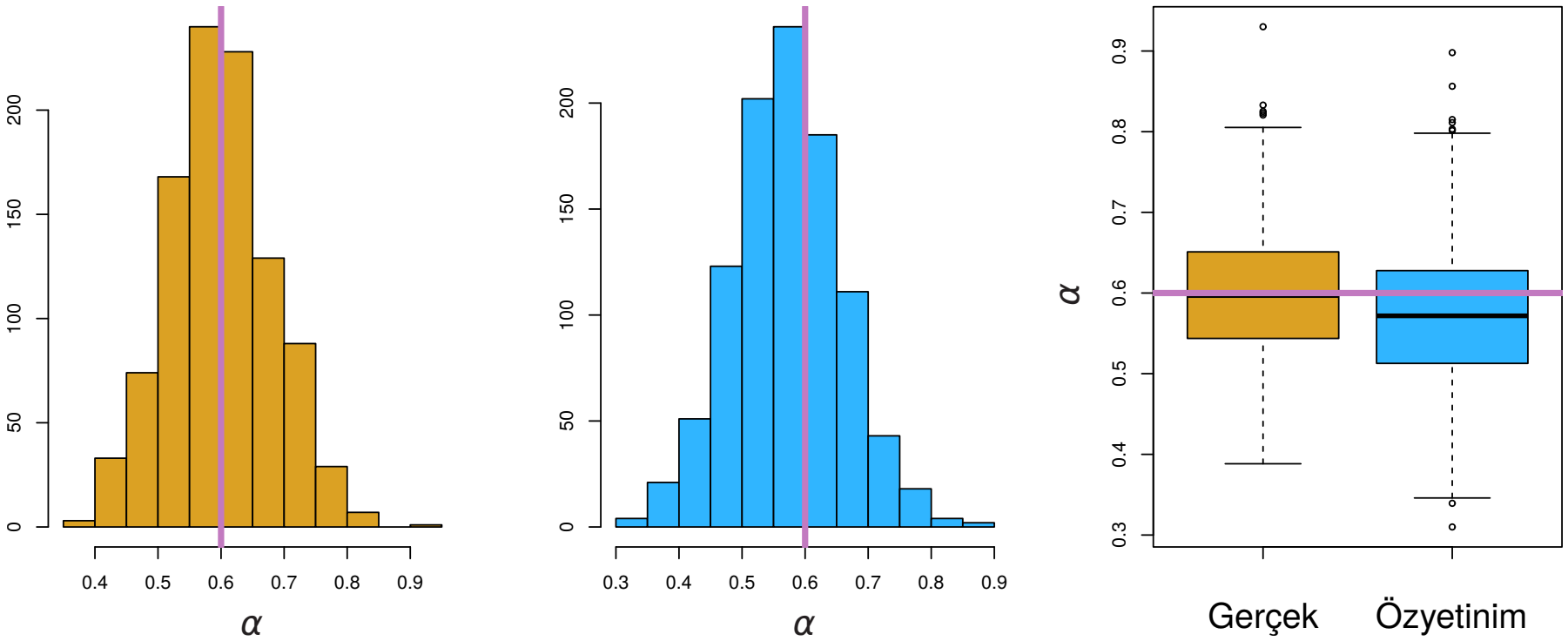
- Özyetinim yöntemiyle üretilen çoğaltmalardan $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \ldots, \hat{\alpha}_\Omega^*$ şeklinde gösterilen özyetinim tahminleri hesaplanabilir.
- Daha sonra bunlara ait $\bar{\hat{\alpha}}$ ve ÖH$(\hat{\alpha})$ değerleri de yukarıdakine benzer şekilde elde edilir:

$$\text{ÖH}_{\ddot{o}}(\hat{\alpha}) = \sqrt{\frac{1}{\Omega}\sum_{i=1}^{\Omega}(\hat{\alpha}_i^* - \bar{\hat{\alpha}}^*)^2} = 0{,}087$$

- Burada ÖH$_{\ddot{o}}$ özyetinim yoluyla bulunan ölçünlü hata anlamındadır.
- Simülasyon veri setlerini kullanarak α için ölçünlü hatayı 0,083 tahmin etmiştik. Tek bir veri setinden ürettiğimiz özyetinim ölçünlü hata tahmininin buna ne kadar yakın çıktığına dikkat ediniz.
- Portföy örneğimizde simülasyondan gelen $\hat{\alpha}$ ideal tahminleri ile $\hat{\alpha}^*$ özyetinim tahminlerinin karşılaştırılması Şekil 9'daki gibidir.

Portföy Seçimi Uygulaması (6)

- Şekilde solda simülasyona dayalı $\hat{\alpha}$ tahminleri, ortada ise özyetinime dayalı $\hat{\alpha}^*$ tahminleri görülmektedir. En sağdaki kutu çiziminden de anlaşıldığı gibi özyetinim ideale çok yakın sonuç vermiştir.



Şekil 9: İdeal tahminler ile özyetinim tahminlerinin karşılaştırılması

Ders Planı

Maksimum Entropi Özyetinimi (1)

- Yukarıda açıkladığımız özyetinim işlemi birçok durumda geleneksel yöntemlerden çok daha başarılı tahminler üretebilmektedir.
- Ancak zaman serilerinde genellikle iyi sonuç verememektedir.
- Bunun nedeni zaman serilerinde gözlemlerin belli bir sıra izlemesidir. Gözlemleri rastsal bir şekilde yeniden düzenlemek bu mantıksal sırayı yok eder.
- Bu sorun aşmak amacıyla geçmişte zaman serilerine özel özyetinim teknikleri geliştirilmiştir.
- Örnek olarak, blok özyetinimi (block bootstrap) adı verilen yöntemde gözlemler sabit büyüklükte bloklar şeklinde alınarak yeniden örnekleme yapılır.
- Ancak bu ve bunun gibi yaklaşımların başarı düzeyi düşüktür. Yeterli sonuçlar ancak durağanlık, düşük özilinti ve büyük örneklem gibi belli koşullar altında elde edilebilmektedir.
- Sonuç olarak, özyetinim tekniği yukarıdaki zorluklar nedeniyle makroekonomi ve finans gibi alanlarda yaygın kullanım bulamamıştır.

Maksimum Entropi Özyetinimi (2)

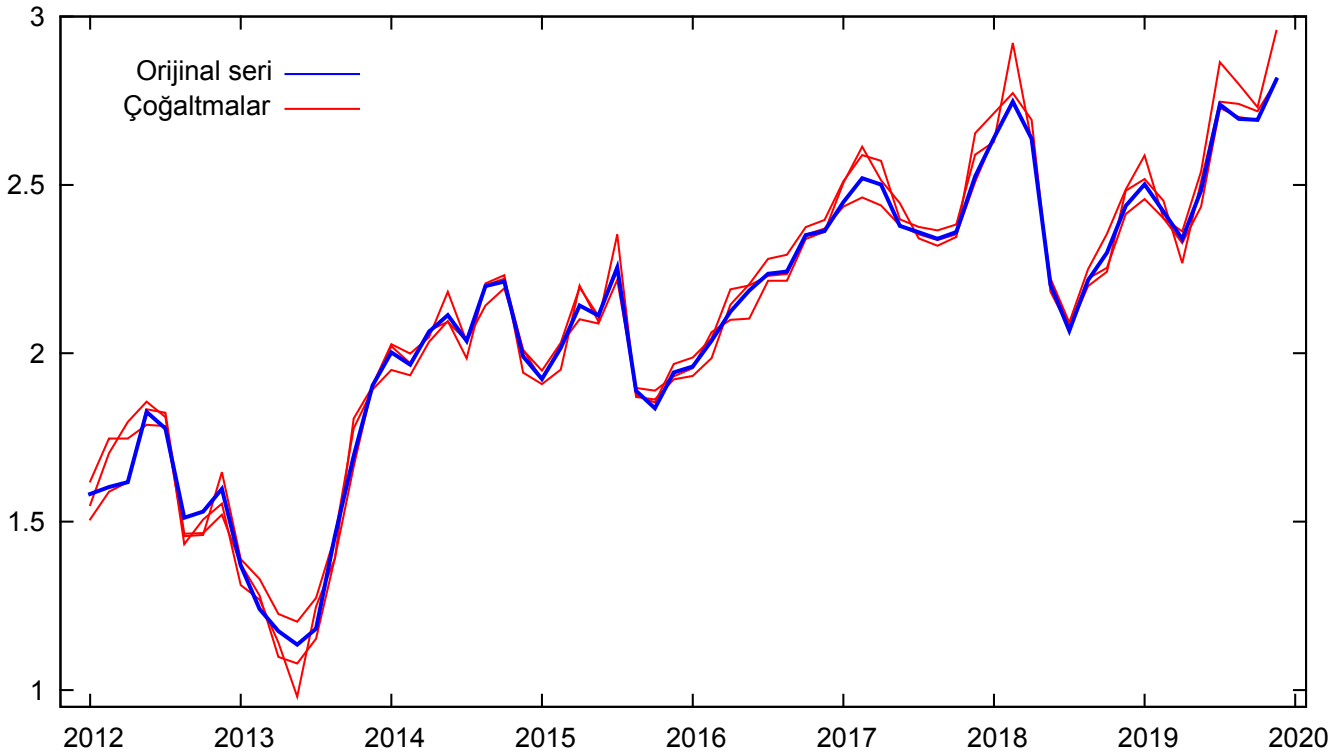
- Özellikle zaman serileri için geliştirilmiş olan en yeni ve güncel özyetinim tekniği maksimum entropi özyetinimi (maximum entropy bootstrap) ya da kısaca meboot denilen yöntemdir.
- Entropi (entropy) kavramı, Bayesçi olasılık kuramında yaygın olarak kullanılan ve yetersiz bilgi kısıtı altında gereksiz varsayımlardan kaçınmaya yarayan güçlü bir araçtır.
- Bu kavramı temel alan maksimum entropi özyetinimi yedi adımlı bir algoritma kullanarak Shannon bilgi ölçütünü maksimize eder:

$$H = \mathrm{E}(-\ln f(x))$$

- Böylece, zaman serisi bir bütün olarak çoğaltılır.
- Bu şekilde üretilen seriler baştaki asıl seriye ait iniş, çıkış, eğilim gibi bağımlılık bilgilerini koruma özelliğine sahiptir.
- Örnek meboot çoğaltmaları Şekil 10'da gösterilmiştir.

Örnek Meboot Çoğaltmaları

- Şekilde orijinal seri mavi renkle, üç adet örnek meboot çoğaltması ise kırmızı renkle gösterilmiştir. Çoğaltılmış serilerin asıl seriye ait iniş, çıkış, eğilim gibi özellikleri koruduğuna dikkat ediniz.



Şekil 10: Örnek meboot çoğaltmaları

Maksimum Entropi Özyetiniminin Uygulanması

- Meboot özyetiniminin en önemli üstünlüğü her türlü yapısal kırılma, durağan-dışılık ve eştümleşim altında herhangi bir dönüştürmeye gerek olmadan kullanılabilmesidir.
- Yöntemi uygulamak başta açıkladığımız özyetinim gibidir:

1. Eldeki eğitim veri seti meboot algoritması ile Ω kez çoğaltılır.
2. Çoğaltılan setler ististiksel öğrenme yöntemine uygulanır.
3. Elde edilen çok sayıda tahmin toplulaştırılarak sağlam kestirimler yapılır ya da güven aralıkları inşa edilir.

- Yukarıdaki işlemin blok özyetinimi gibi almaşık tekniklere göre daha üstün tahminler ürettiği kapsamlı simülasyon çalışmaları ile gösterilmiştir (Vinod, 2015; Yalta, 2016; Singvejsakul et al. 2018).
- Meboot tahminine yönelik R dilinde yazılmış açık kaynaklı paket ve örnek kodlar bulunmaktadır.

Önümüzdeki Dersin Konusu ve Ödev

Ödev

Kitaptan Bölüm 5 “Yeniden Örnekleme” okunacak.

Önümüzdeki Ders

Model Seçimi