İstatistiksel Öğrenme: Temel Kavramlar

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. İstatistiksel Modelleme
   - Girdi ve çıktı değişkenleri
   - Kestirim ve çıkarsama
2. Tahmin Konusu
   - Parametrik ve parametrik-dışı yöntemler
   - Kesinlik ve yorumlanabilirlik
   - Denetimli ve denetimsiz öğrenme
3. Kesinliğin Ölçülmesi
   - Yakışmanın iyiliği
   - Yanlılık-varyans ödünleşmesi
   - Sınıflandırmadaki durum
   - Bayes sınıflandırıcı
   - K-enyakın komşu sınıflandırıcı

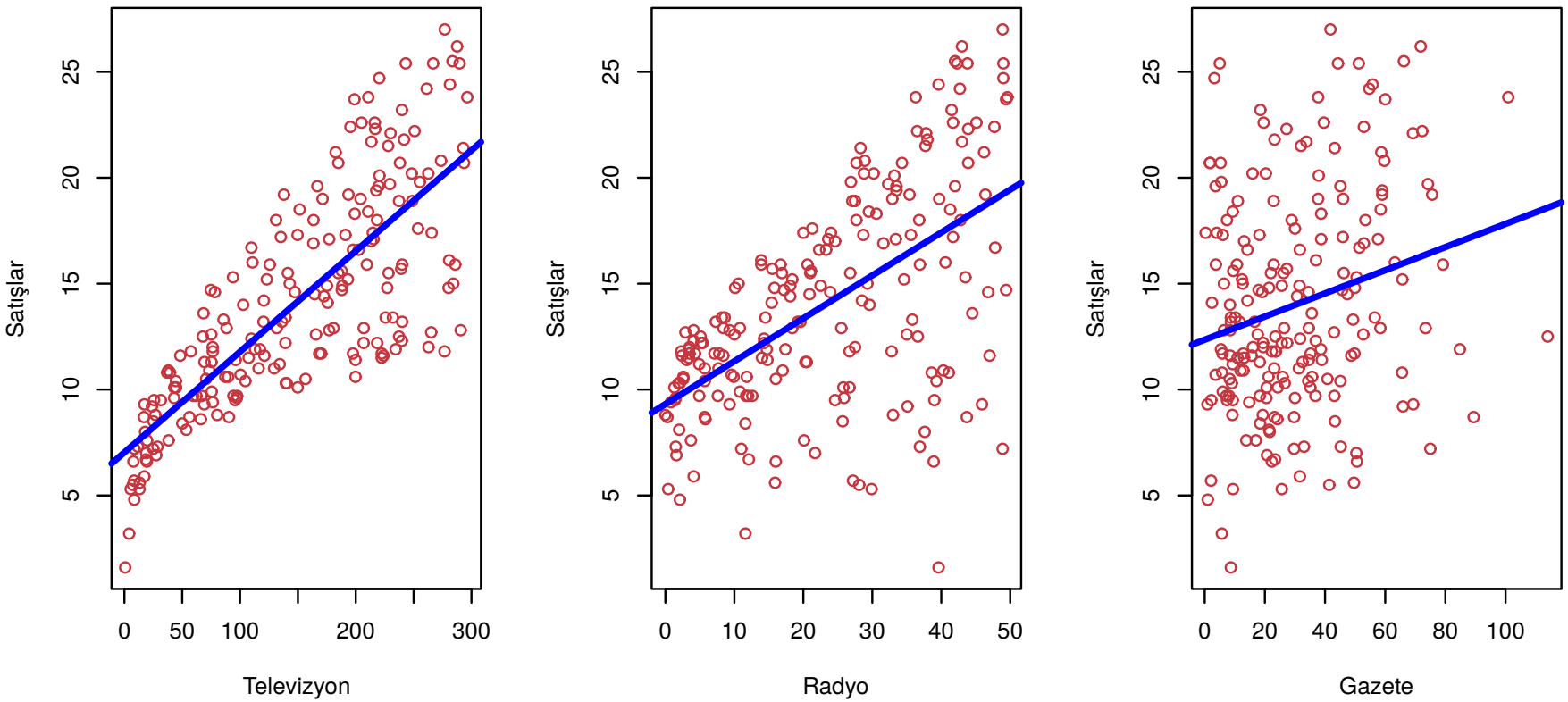
Ders Planı

1. İstatistiksel Modelleme
   - Girdi ve çıktı değişkenleri
   - Kestirim ve çıkarsama
2. Tahmin Konusu
   - Parametrik ve parametrik-dışı yöntemler
   - Kesinlik ve yorumlanabilirlik
   - Denetimli ve denetimsiz öğrenme
3. Kesinliğin Ölçülmesi
   - Yakışmanın iyiliği
   - Yanlılık-varyans ödünleşmesi
   - Sınıflandırmadaki durum
   - Bayes sınıflandırıcı
   - K-enyakın komşu sınıflandırıcı

Basit Bir Örnek

- İstatistiksel öğrenmenin amacı veri setlerinden sistematik bilgi elde etmektir. Gelin, bunu bir örnek üzerinde açıklayalım.
- Belli bir ürünün satışını artırmak istediğimizi düşünelim. Elimizde bu ürüne ait 200 farklı piyasadaki satışlar ile televizyon, radyo ve gazeteler için reklam harcamaları bilgisi olduğunu varsayalım.
- Kısaca, farklı medya türlerindeki reklam harcamasına bağlı olarak satışların nasıl değiştiğini anlamak istiyoruz.
- Burada öğrenmek istediğimiz hedef bilgi satışlardır. Buna çıktı (output) değişkeni diyelim.
- Bu durumda çıktı değişkenini açıklamak için kullanacağımız reklam harcamaları da girdi (input) değişkeni olur.
- Örneğimizdeki veriler Şekil 1'de gösterilmiştir.

Reklam Verileri



Şekil 1: Medya türüne göre reklamların satış üzerindeki etkisi

Girdi ve Çıktı Değişkenleri

- Şekilde çıktı değişkeninin y-ekseninde, girdi değişkenlerinin ise x-ekseninde bulunduğuna dikkat ediniz.
- Genel olarak, çıktı değişkenini Y harfi ile belirtiriz.
- Girdi değişkenlerini ise X_1, X_2, $X_3, \ldots$ harfleri ile gösteririz.
- Bu değişkenlere duruma göre farklı adlar da verebiliriz:

Çıktı Değişkeni (Y)		**Girdi Değişkeni (X)**	
Bağımlı değişken	(Dependent variable)	Bağımsız değişken	(Independent variable)
Kestirilen	(Predictand)	Kestirici	(Predictor)
Açıklanan değişken	(Explained variable)	Açıklayıcı değişken	(Explanatory variable)
Tepki değişkeni	(Response variable)	Denetim değişkeni	(Control variable)

İstatistiksel Model (1)

- Girdi ve çıktı değişkenleri arasındaki ilişkiyi incelemek için istatistiksel modellerden yararlanırız:

İstatistiksel model

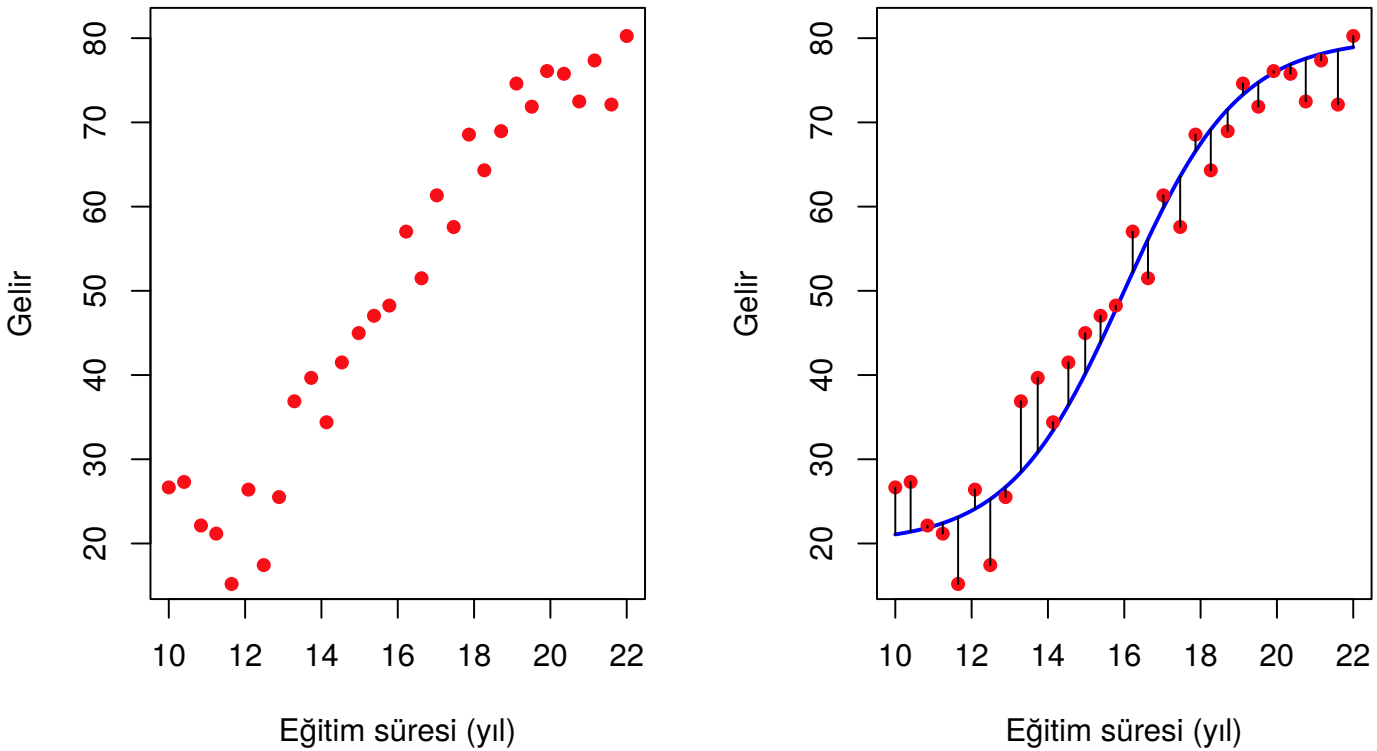
İstatistiksel model (statistical model), anakütleden gelen örneklem verilerinin nasıl oluştuğunu açıklayan matematiksel fonksiyondur.

- Bir istatistiksel modelin en genel gösterimi aşağıdaki gibidir:

$$Y = f(X) + \epsilon$$

- Burada f sabit ancak bilinmeyen bir matematiksel fonksiyondur. X'lerin Y hakkında sağladığı tüm sistematik bilgi budur.
- Soldaki ϵ (epsilon okunur) ise X'lerden bağımsız ve ortalaması sıfır olan rastsal hata terimi (random error term) olarak adlandırılır.
- f'yi ve ϵ'u anlamak için önceki bölümdeki ücret veri setine geri dönelim ve Şekil 2'yi inceleyelim.

İstatistiksel Model (2)



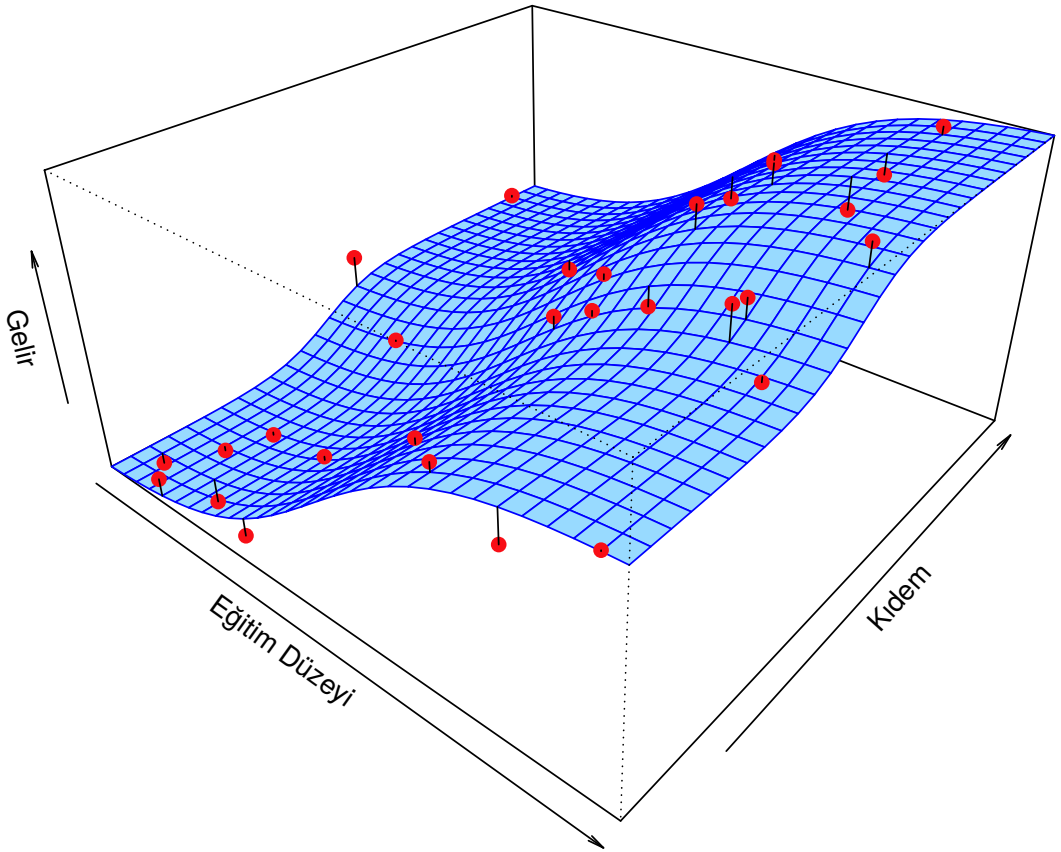
Şekil 2: Gelir ve eğitim ilişkisini gösteren istatistiksel model

İstatistiksel Model (3)

- Şekildeki veri seti simülasyon yoluyla oluşturulduğu için bu örnekte f bellidir ve sağ panelde mavi çizgi ile gösterilmiştir.
- Buradaki dikey çizgiler hata terimi olan ϵ'ları göstermektedir.
- Hataların bir bölümünün fonksiyon çizgisinin üstünde, diğerlerinin ise çizginin altında kaldığını ve ortalamalarının yaklaşık sıfır olacağına dikkat ediniz.
- Şekilde eğitim süresi arttıkça gelirin önce artarak arttığı ve bir noktadan sonra da azalarak artmaya başladığı anlaşılmaktadır.
- Ancak uygulamada X ve Y arasındaki ilişkiyi belirten f fonksiyonunu kesin olarak bilmek olanaksızdır.
- Gerçek hayatta elimizdeki tek bilgi sol panelde gösterilen verilerdir. Dolayısıyla f'yi tahmin (estimate) etmemiz gerekir.

İstatistiksel Model (4)

- İstatistiksel modellerde genellikle birden fazla girdi bulunur. Örnek olarak, bir yerine iki adet X değişkeni olduğu zaman f'yi aşağıdaki gibi üç boyutlu bir yüzey ile gösterebiliriz:



Şekil 3: Gelir, eğitim süresi ve kıdem arasındaki ilişki

Kestirim

- Bir istatistiksel modeli tahmin etmenin iki amacı vardır: kestirim (prediction) ve çıkarsama (inference).
- Kestirim, eldeki verili X değerlerini kullanarak buna karşılık gelen Y değerini tahmin etmek demektir. Bunu şöyle gösterebiliriz:

$$\hat{Y} = \hat{f}(X)$$

- Yukarıda görülen ^ (şapka) simgesi tahmin anlamındadır.
- Kestirim uygulamasında tahmin edilen f'yi bir kara kutu olarak görürüz. İçeriğine değil, doğru sonuç verip vermediğine bakarız.

Azaltılabilen Hata ve Azaltılamayan Hata (1)

- Kestirilen $\hat{Y}$ değerinin doğruluk derecesi iki büyüklüğe bağlıdır: Azaltılabilen hata ve azaltılamayan hata
- Azaltılabilen hata, daha iyi ve uygun bir istatistiksel öğrenme aracı kullanılarak kaçınılabilecek olan hatalardır.
- Azaltılamayan hata ise rastsal hata teriminden kaynaklanır. Y aynı zamanda ϵ'a da bağlı olduğu için bundan kaçınmak olanaksızdır.
- Peki, ϵ neden vardır? Hata terimi genel olarak (1) ölçemediğimiz, (2) hatalı ölçebildiğimiz, (3) göz ardı ettiğimiz ya da (4) hiç bilemediğimiz değişkenlerin ortak etkisini gösterir.
- Bunların Y ve X'lerden bağımsız ve birbirlerinin etkisini yok ettikleri için sıfır ortalamaya sahip oldukları kabul edilir.

Azaltılabilen Hata ve Azaltılamayan Hata (2)

- Azaltılabilen hata ile azaltılamayan hatayı açıklamak için $\hat{f}$ ve X'lerin sabit olduğunu varsayalım. Bu durumda şunu yazabiliriz:

$$\begin{aligned} \mathrm{E}(Y - \hat{Y})^2 &= \mathrm{E}[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) + \epsilon - \hat{f}(X)]^2}_{\text{Azaltılabilen}} + \underbrace{\mathrm{var}(\epsilon)}_{\text{Azaltılamayan}} \end{aligned}$$

- Burada $\mathrm{E}(Y - \hat{Y})^2$, kestirim hatasının karesinin beklenen değeridir. Kısaca ortalama hatanın karesidir.
- $\mathrm{var}(\epsilon)$ ise hata terimine ait varyans (variance) değeridir.
- Görüldüğü gibi, istatistiksel öğrenmedeki amacımız azaltılabilen hatayı minimize etmektedir.
- Ancak uygulamada ϵ'dan kaynaklanan bir kestirim hatası her zaman olacaktır.

Çıkarsama

- Veri çözümlemesinde belli bir Y değerini kestirmek dışında çoğu zaman çıkarsama da yapılır.
- Burada amaç Y'nin X'lerden nasıl etkilendiğini anlamaktır. Bu durumda f fonksiyonunu bir kara kutu olarak görmeyiz.
- Çıkarsama yaparak aşağıdaki sorulara yanıt ararız:
  Hangi X değişkenleri Y üzerinde etkilidir?
  Y ile her bir X değişkeni arasındaki ilişki nedir?
  Y'nin X'lerle ilişkisi doğrusal mı yoksa daha karmaşık mıdır?
- Örnek olarak, ev satışı yapan bir firmayı ele alalım.
- Burada daire genişliği bir m^2 artarsa fiyatın ne kadar artacağı, dairenin ön cephede olmasının ya da çocuk odasında banyo bulunmasının fiyatı nasıl etkileceği çıkarsamanın konusudur.
- Belli bir dairenin yalnızca fiyatını tahmin etmek isteseydik bu kestirim konusu olurdu.

Ders Planı

1. İstatistiksel Modelleme
   - Girdi ve çıktı değişkenleri
   - Kestirim ve çıkarsama
2. Tahmin Konusu
   - Parametrik ve parametrik-dışı yöntemler
   - Kesinlik ve yorumlanabilirlik
   - Denetimli ve denetimsiz öğrenme
3. Kesinliğin Ölçülmesi
   - Yakışmanın iyiliği
   - Yanlılık-varyans ödünleşmesi
   - Sınıflandırmadaki durum
   - Bayes sınıflandırıcı
   - K-enyakın komşu sınıflandırıcı

Matematiksel Gösterim

- Tahmin konusuna geçmeden önce kullanacağımız matematiksel gösterimden kısaca söz edelim.
- Bu derste ele alacağımız birçok farklı yöntemi açıklarken ya da bunları karşılaştırırken ortak bir gösterimden yararlanacağız.
- Örnek olarak, veri setimizin uzunluğunu n olarak belirleyeceğiz.
- Diğer bir deyişle elimizde n adet gözlem olduğunu düşünecek ve bunları i olarak adlandıracağız: $i = 1, 2, \ldots, n$.
- Açıklayıcı değişken sayımız ise p olacak ve bunları da j harfi ile göstereceğiz: $j = 1, 2, \ldots, p$.
- Değişkenleri Y ve X'ler şeklinde büyük harflerle yazacağız.
- Tekil gözlemler ise x_{ip} şeklinde küçük harf olacak.
- Örnek olarak, x_{11} dediğimiz zaman bu, birinci gözlemdeki birinci X değişkeni anlamına gelecek.
- Tüm X değişkenlerine birlikte eğitim verileri (training data) diyeceğiz: $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.
- Amacımız $Y \approx \hat{f}(X)$ olacak şekilde $\hat{f}$ tahminini bulmak olacak.

Parametrik ve Parametrik-dışı Yöntemler

- Matematiksel gösterimi açıkladıktan sonra artık f fonksiyonunun nasıl tahmin edileceği konusuna geçebiliriz.
- Veri çözümlemesinde amacımızın kestirim mi yoksa çıkarsama mı olduğuna bağlı olarak farklı yöntemler kullanılabiliriz.
- Örnek olarak, doğrusal (linear) modeller görece basit ve anlaşılabilir yapıları nedeniyle yorumlama kolaylığı sağlar.
- Dolayısıyla bunlar çıkarsama amacı için daha uygundur.
- Doğrusal olmayan (non-linear) modelleri ise yorumlamak güçtür ama bunlar da kestirim konusunda çoğu zaman daha başarılıdır.
- Bu derste f fonksiyonunu tahmin etmek için çok sayıda doğrusal ve doğrusal-dışı yöntem göreceğiz.
- Bu yöntemleri genel olarak parametrik (parametric) ve parametrik-dışı (non-parametric) şeklinde iki gruba ayırabiliriz.

Parametrik Yöntemler (1)

Parametrik yöntemlerde iki adımlı bir yaklaşım izlenir.

1. İlk adımda f'nin fonksiyon yapısına karar verilir. Örnek olarak, şöyle bir doğrusal model (linear model) kullanabiliriz:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

   Doğrusallık varsayımı model tahminini kolaylaştırır. Çok karmaşık olabilecek bir fonksiyonla uğraşmak yerine $p + 1$ adet katsayıyı tahmin etmek burada yeterli olur.

2. İkinci olarak, elimizdeki verileri modele yakıştırmak (fitting), diğer bir deyişle modeli eğitmek (training) isteriz. Böylece, modeldeki katsayıları tahmin ederiz:

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

   Bu işlemin birçok farklı yolu vardır. Bunlar içinde en yaygını ise sıradan enküçük kareler (ordinary least squares) yöntemidir.

Parametrik Yöntemler (2)

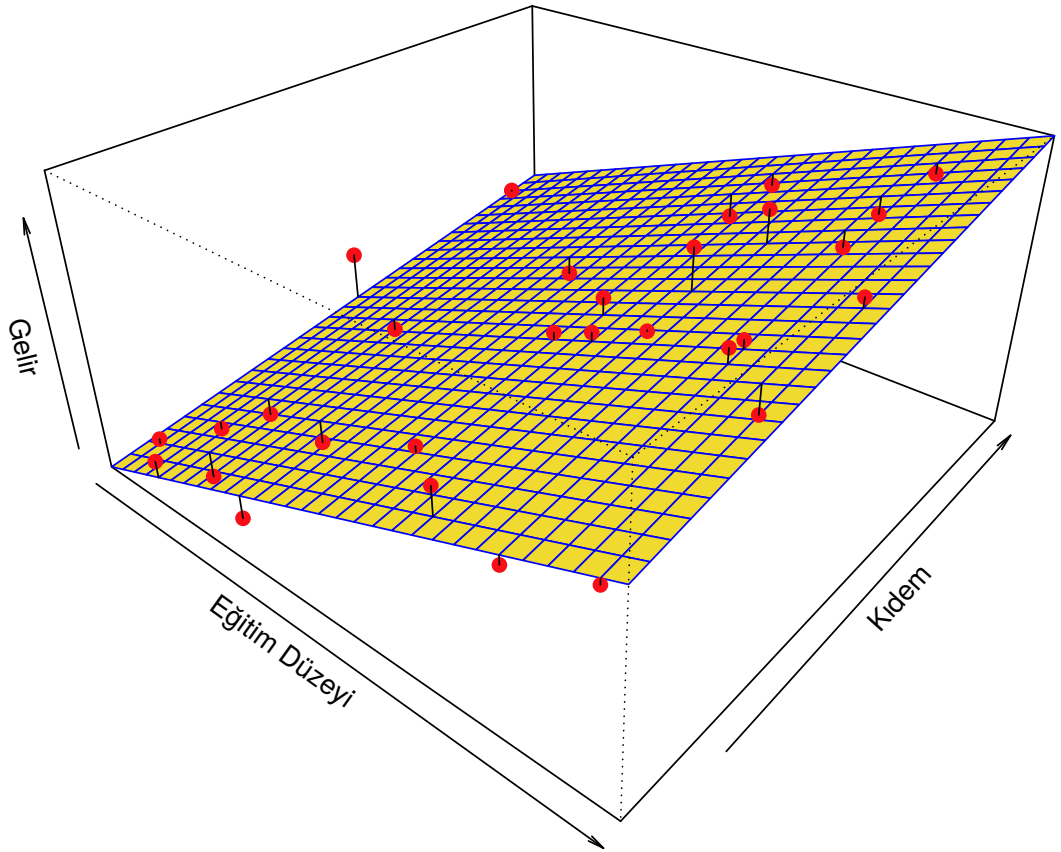
- Parametrik yöntemde fonksiyon yapısını biz belirlediğimiz için tahmin süreci oldukça kolaylaşır.
- Ancak gerçek hayatta değişkenler arasındaki karmaşık ilişkileri önceden bilmek zordur.
- Dolayısıyla parametrik yaklaşımın olası sakıncası aşırı basit bir model kullanmaktır.
- Örnek olarak, başta söz ettiğimiz eğitim süresi, kıdem ve gelir arasındaki ilişkiyi parametrik yöntemle tahmin etmek için aşağıdaki doğrusal modeli kullanalım:

$$\text{Gelir} \approx \beta_0 + \beta_1 \text{ Eğitim} + \beta_2 \text{ Kıdem}.$$

- Elimizdeki verileri sıradan enküçük kareler yöntemi ile modele yakıştıracak olursak Şekil 4'teki tahmin yüzeyini elde ederiz.

Doğrusal Modelin SEK Yöntemi ile Tahmin Edilmesi

- Doğrusal modelin Şekil 3'te gördüğümüz gerçek *f*'deki eğri yüzeyi yakalayamadığı anlaşılıyor. Ancak küçük bir veri seti ile yapılabilecek en iyi tahmin belki de bu olabilir.



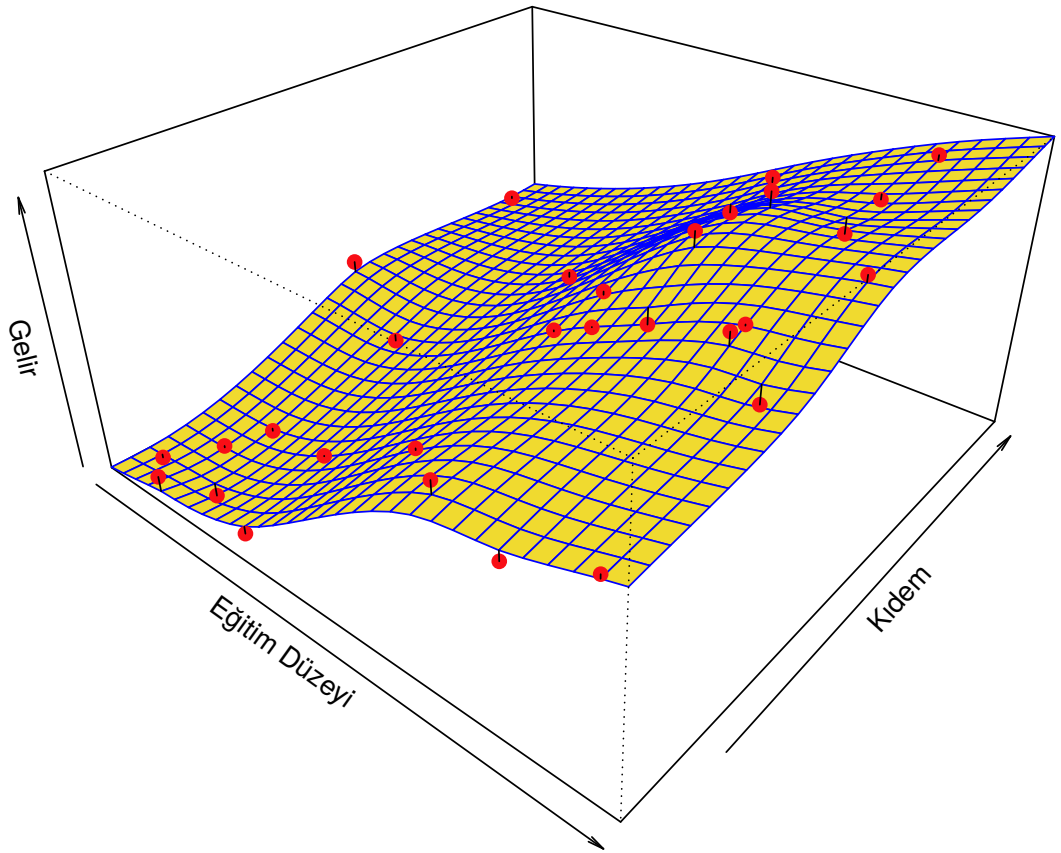
Şekil 4: Gelir, eğitim, kıdem ilişkisinin doğrusal model ile tahmini

Parametrik-dışı Yöntemler (1)

- Parametrik-dışı yöntemler f'nin fonksiyon yapısı konusunda bir varsayımda bulunmaz. Bunun yerine eldeki verilere en iyi yakışan fonksiyonu bulmaya çalışır.
- Bu yöntemde tahmin sonuçlarının aşırı düz ya da aşırı eğri olmaması önemlidir. Bunun için uygun bir düzleştirme (smoothing) derecesi seçmek gereklidir.
- Parametrik-dışı yaklaşımı kullanarak değişkenler arasındaki çok karmaşık ilişkileri dikkate alabiliriz.
- Ancak bu yaklaşımın sakıncası da bu iş için çok daha fazla veriye gereksinim duymalarıdır.
- Parametrik-dışı yönteme örnek olarak, şimdi de gelir modelimizi ince-katman spline (thin-plate spline) yöntemi ile tahmin edelim. Buradan elde edilen sonuçlar Şekil 5'te gösterilmiştir.

Doğrusal Modelin SEK Yöntemi ile Tahmin Edilmesi

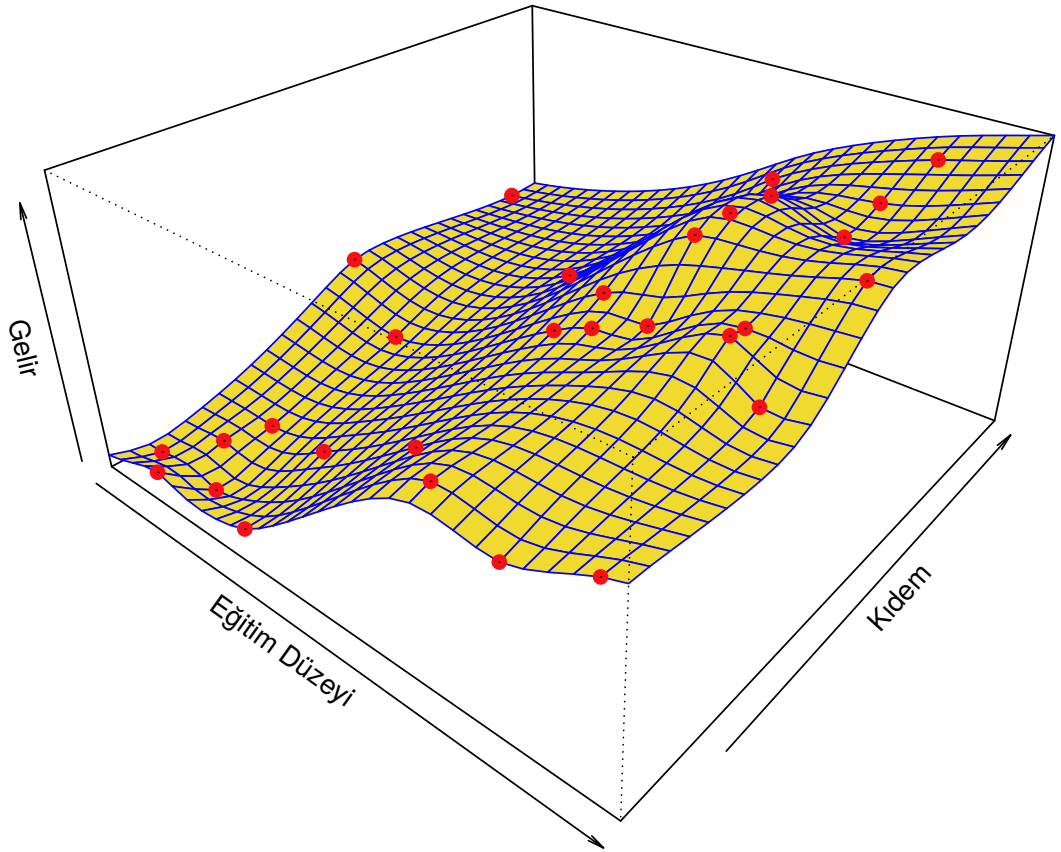
- Bu şekilde spline yöntemi için yüksek bir düzleştirme değeri kullanılmıştır. Tahmin edilen yüzeyin daha önce Şekil 3'te gösterilen gerçek duruma çok yakın olduğu görülüyor.



Şekil 5: Gelir, eğitim, kıdem ilişkisinin düzgün ince-katman spline ile tahmini

Doğrusal Modelin SEK Yöntemi ile Tahmin Edilmesi

- Burada ise spline için düşük bir düzleştirme uygulanmıştır. Burada aşırı yakışma (over fitting) söz konusudur. Elde edilen sonuç gerçek durumu tam yansıtmamaktadır.

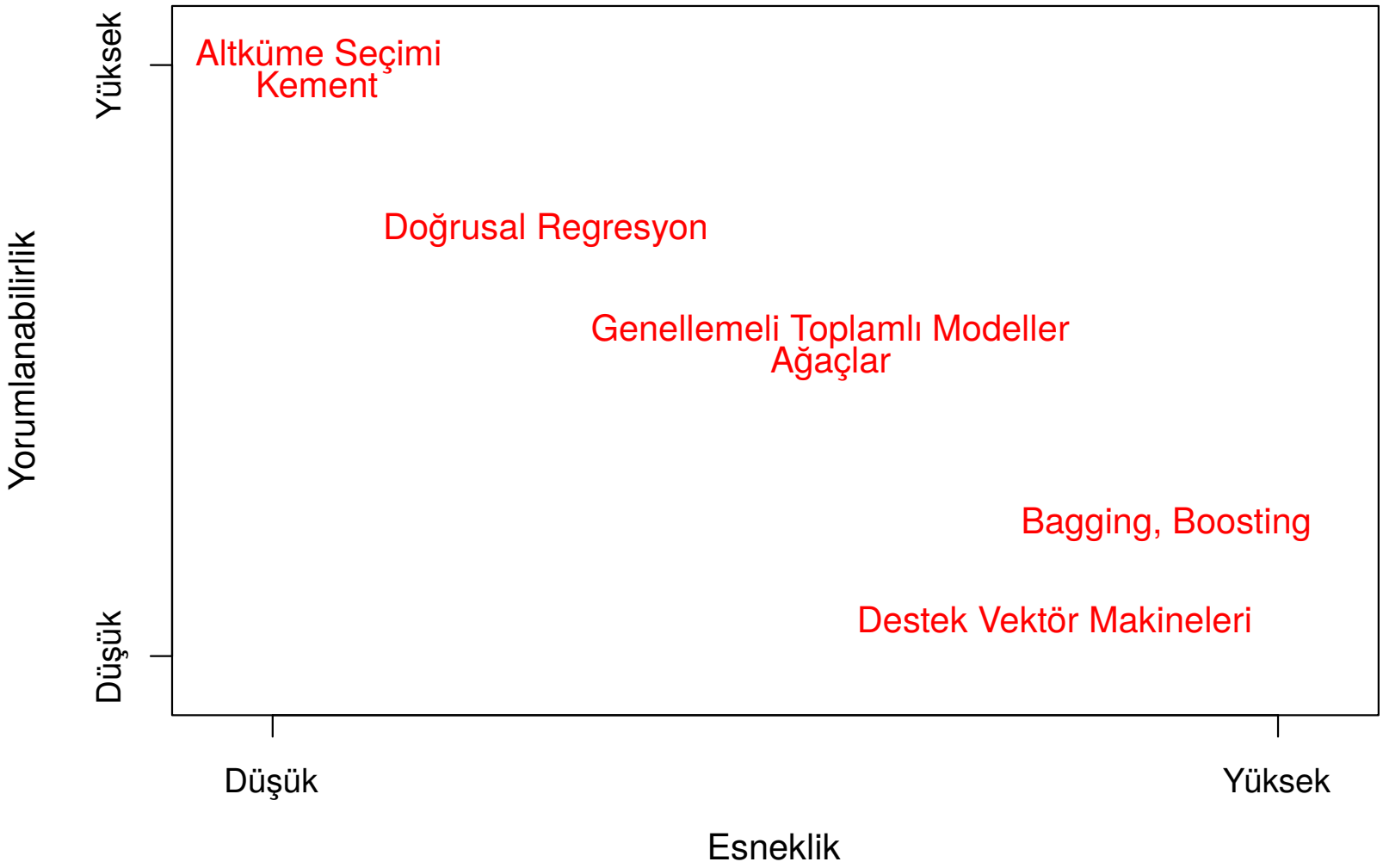


Şekil 6: Gelir, eğitim, kıdem ilişkisinin engebeli ince-katman spline ile tahmini

Kesinlik ve Yorumlanabilirlik (1)

- Bu derste göreceğimiz çok sayıda yöntemden bazıları uygulama konusunda esnek, bazıları ise kısıtlayıcıdır.
- Örnek olarak doğrusal regresyon görece kısıtlayıcıdır çünkü yalnızca doğrusal fonksiyonlara izin verir.
- Öte yandan ince-katman spline yöntemi f fonksiyonu için oldukça ayrıntılı şekiller üretebildiği için esnektir.
- Peki, esnek yöntemler varken neden kısıtlayıcı bir yöntem kullanalım? Bunun yanıtı ikisi arasındaki kesinlik (accuracy) ve yorumlanabilirlik (interpretability) farkıdır.
- Esnek yöntemler kullanarak değişkenler arasındaki çok karmaşık ilişkileri dikkate alabiliriz. Dolayısıyla bunlar kestirim konusunda genellikle daha başarılıdır.
- Kısıtlayıcı yöntemler ise görece basit ve anlaşılabilir yapıları nedeniyle anlaması ve yorumlaması kolay sonuçlar üretirler.
- Kesinlik ve yorumlanabilirlik ödünleşmesi Şekil 7'deki gibidir.

Kesinlik ve Yorumlanabilirlik (2)



Şekil 7: Kesinlik ve yorumlanabilirlik arasındaki ödünleşme

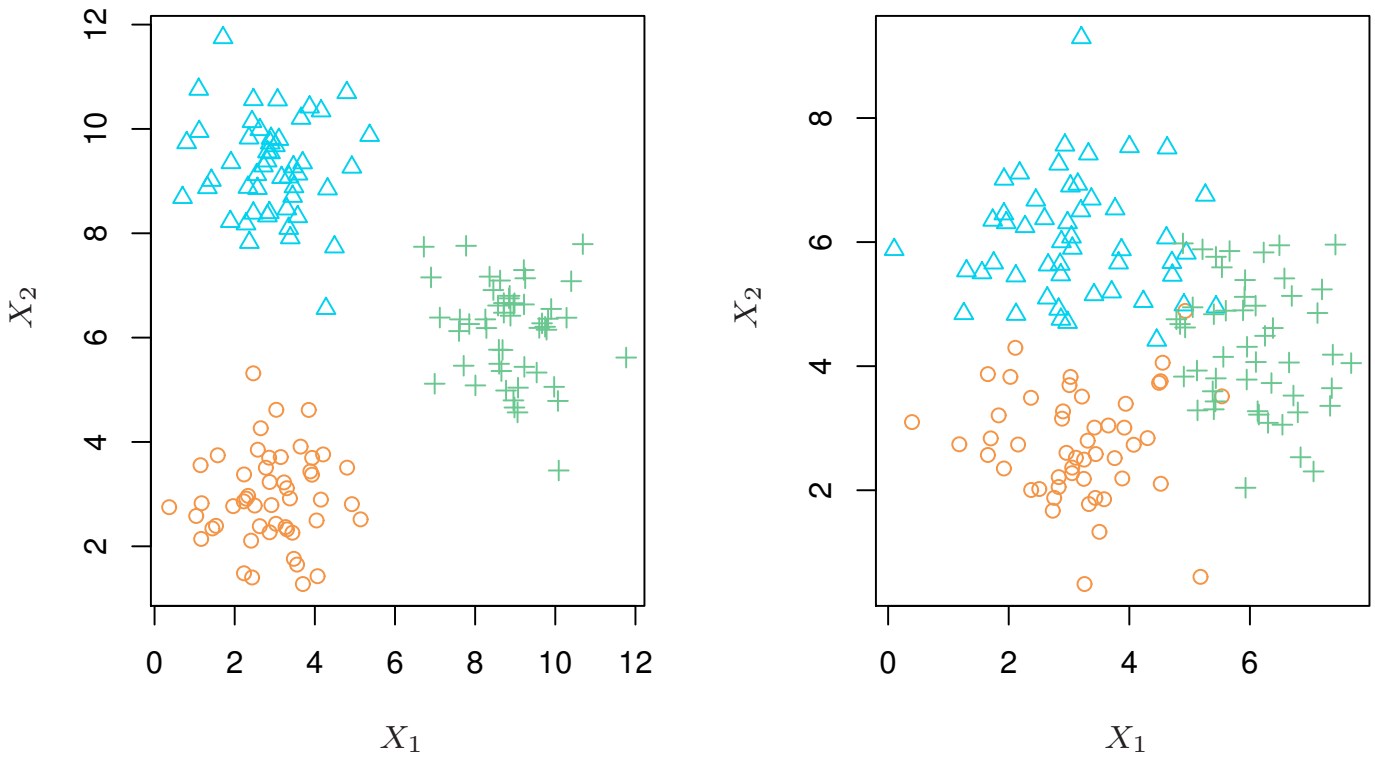
Kesinlik ve Yorumlanabilirlik (3)

- Şekilde çeşitli istatistiksel öğrenme yöntemlerinin esnekliği arttıkça yorumlanabilirliğinin düştüğü görülmektedir.
- Örnek olarak, Bölüm 8'de ele alacağımız boosting ile Bölüm 9'da göreceğimiz destek vektör makineleri oldukça esnek araçlardır.
- Ancak bunların ürettiği f fonksiyonu tahminleri son derece karmaşık olabildiği için her bir X'in Y üzerindeki etkisini anlamak zordur.
- 7. Bölümde tartışacağımız genellemeli toplamlı modeller ise 3 Bölümde göreceğimiz doğrusal regresyon yanında daha esnektir.
- Yine, 6. Bölümde inceleyeceğimiz kement (lasso) yöntemi de bazı parametreleri sıfıra eşitlediği için doğrusal regresyona göre daha katıdır ancak bu durum yorumlamada kolaylık sağlar.
- Peki, yorum yapmakla ilgilenmiyorsak ne olacak? Amacımız yalnızca kestirim yapmak ise en esnek yöntem en iyisi midir?
- Hayır! Esnek yöntemler eğer doğru kullanılmazsa aşırı yakışma (overfitting) sorununa neden olurlar. Bu durumda yorumlanabilirlik pahasına kazanılan kesinlik kolayca kaybedilir.

Denetimli ve Denetimsiz Öğrenme

- İstatistiksel öğrenme süreçleri parametrik ve parametrik-dışı ayrımına ek olarak, denetimli (supervised) ve denetimsiz (unsupervised) olarak da ikiye ayrılır.
- Denetimli öğrenmede $x_i, i = 1, 2, \ldots, n$ şeklinde n adet gözlem içeren X değişkenleri ve bunlara karşılık y_i değerleri vardır.
- Doğrusal regresyon ve lojistik regresyon gibi klasik yöntemlerin yanında GAM, boosting, destek vektör makineleri gibi birçok modern yöntem denetimli öğrenmeye örnektir.
- Denetimsiz öğrenmede ise $i = 1, 2, \ldots, n$ şeklinde gözlemler vardır ancak veri seti yalnızca X değişkenlerinden oluşur.
- Tepki değeri olarak y_i'lerin olmadığı böyle durumlarda değişkenler arası ilişkileri anlamak için küme çözümlemesi (cluster analysis), diğer bir deyişle kümeleme (clustering) yapılabilir.
- Kümeleme yöntemine basit bir örnek Şekil 8'de verilmiştir.

Küme Çözümlemesi (1)



Şekil 8: İki farklı veri seti için küme çözümlemesi

Küme Çözümlemesi (2)

- Şekilde iki farklı veri seti için küme çözümlemesi yapılmıştır.
- İki örnekte de üç veri kümesi bulunmaktadır. Bunlar burada farklı renklerle gösterilmiştir ancak gerçekte kümeler bilinmemektedir.
- Sol paneldeki kümeleri ayrıştırmak daha kolaydır. Sağda ise kümeler örtüştüğü için hatasız bir sonuç elde etmek beklenemez.
- Küme çözümlemesi günümüzde sık kullanılan bir yaklaşımdır.
- Örnek olarak, bir firma bu yöntemle potansiyel müşterileri arasında çok ya da az harcama yapacakları ayrıştırmak isteyebilir.
- Eğer elimizde harcama verileri bulunsaydı denetimli bir çözümleme yapılabilirdi. Ancak gerçekleşecek harcama genellikle önceden bilinmediği için en uygulanabilir yöntem budur.
- Son olarak, çoğu durumda ikiden fazla değişken olacağına dikkat ediniz. Eğer elimizde p adet değişken varsa her bir değişken çifti için toplam $p(p-1)/2$ farklı serpilim çizimi oluşturulabilir.
- Bunları insanların yorumlaması zor olduğu için otomatik sınıflandırma yapan gelişmiş yöntemler giderek önem kazanmaktadır.

Ders Planı

1. İstatistiksel Modelleme
   - Girdi ve çıktı değişkenleri
   - Kestirim ve çıkarsama
2. Tahmin Konusu
   - Parametrik ve parametrik-dışı yöntemler
   - Kesinlik ve yorumlanabilirlik
   - Denetimli ve denetimsiz öğrenme
3. Kesinliğin Ölçülmesi
   - Yakışmanın iyiliği
   - Yanlılık-varyans ödünleşmesi
   - Sınıflandırmadaki durum
   - Bayes sınıflandırıcı
   - K-enyakın komşu sınıflandırıcı

Yakışmanın İyiliği

- Bu derste birçok farklı veri çözümleme tekniğini açıklayacağımızı söylemiştik. Peki, neden yalnızca en yeni ve en gelişmiş yöntemi öğrenmiyoruz?
- Çünkü tüm bu yöntemler içinde diğerlerine her veri setinde üstün gelebilen tek bir yöntem yoktur.
- Dolayısıyla istatistiksel öğrenmedeki en önemli aşamalardan biri belli bir durumda en iyi sonucu verecek yöntemi belirlemektir.
- Bu amaçla, hesapladığımız kestirimlerin gerçekleşen değerlere ne kadar yakın olduğunu ölçmek isteriz.
- Bunun için yakışmanın iyiliği (goodness-of-fit) ölçütleri kullanırız.

Hata Kareleri Ortalaması

- En temel yakışmanın iyiliği ölçütlerinden biri hata kareleri ortalaması (mean squared error), ya da kısaca HKO (MSE) değeridir:

$$\text{HKO} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 = \text{Ortalama}(y_i - \hat{f}(x_i))^2$$

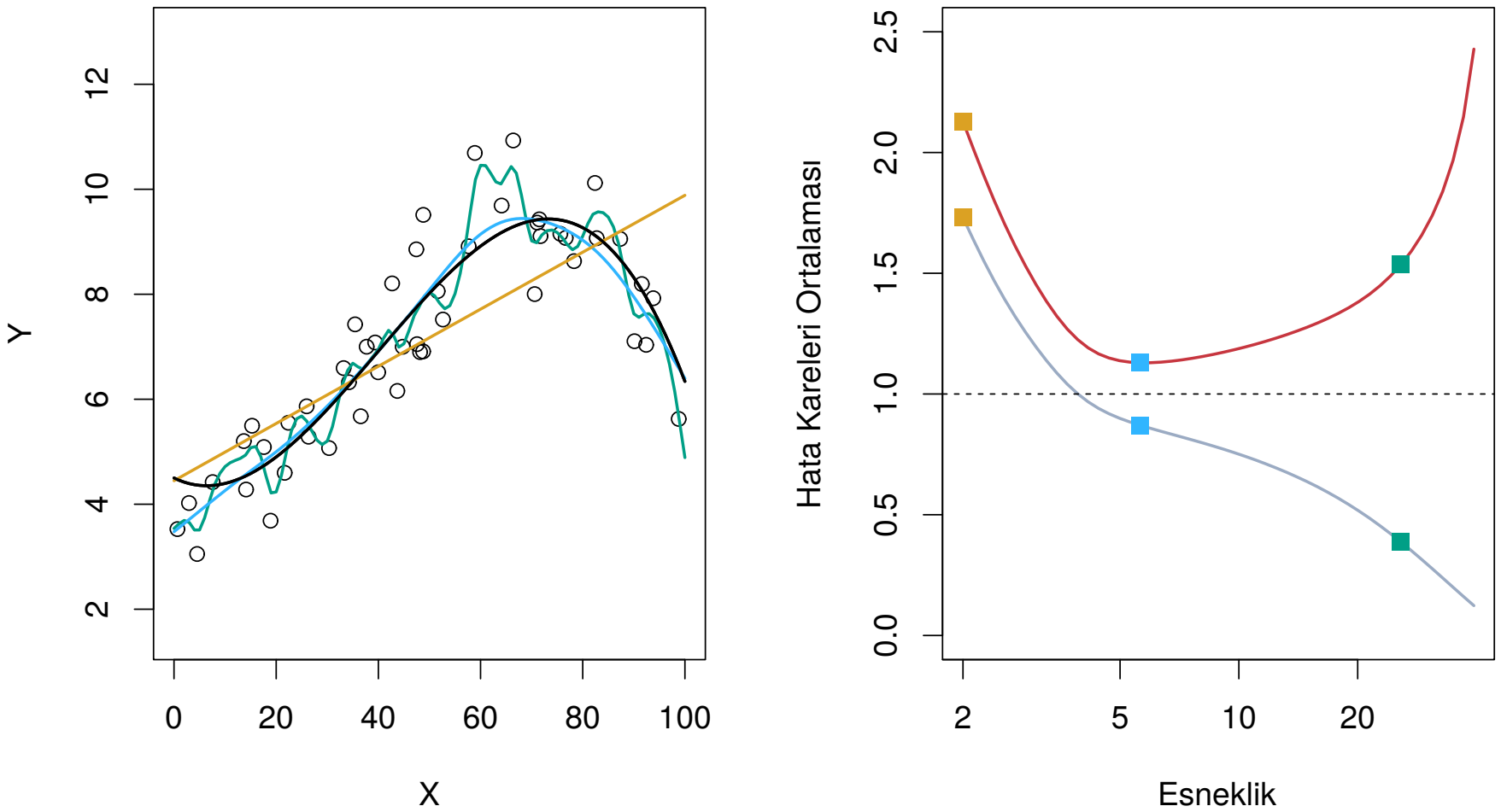
- Görüldüğü gibi HKO, elimizde bulunan x_i ve y_i'leri kullanarak yaptığımız $\hat{f}(x_i) \approx y_i$ şeklinde tahminlerin hata kareleri ortalamasıdır.
- Eğer $\hat{f}(x_i) = y_i$ olursa HKO'nun da sıfır olacağına dikkat ediniz.
- Yukarıdaki formülü elimizde var olan verilerle hesaplarız. Dolayısıyla buna eğitim HKO (training MSE) demek daha doğru olur.
- Ancak bizi asıl ilgilendiren şey elimizde bulunmayan test verileri (test data) kullanırsak tahmin başarısının ne olacağıdır.
- Elimizde olmayan test verilerine x_0 ve y_0 diyelim. Dolayısıyla biz aslında test HKO (test MSE) değerini bilmek istiyoruz:

$$\text{HKO}^{Test} = \text{Ortalama}(y_0 - \hat{f}(x_0))^2$$

Eğitim HKO ile Test HKO İlişkisi (1)

- Genellikle elimizde test verileri yoktur. Bu durumda elde olanla yetinip eğitim HKO'yu minimum yapan yöntemi seçmek isteyebiliriz.
- Bu mantıklı bir yaklaşım gibi görünür. Sonuçta eğitim verileri ile test verileri birbirine yakın olmak zorundadır.
- Ancak burada temel bir sorun ortaya çıkar: Eğitim verilerini minimum yapan yöntem, test verilerinde de aynı başarıyı göstermek zorunda dağildir.
- Uygulamada genellikle eğitim HKO oldukça düşüktür ama test HKO çok daha yüksek çıkar.
- Bu olguyu anlayabilmek için Şekil 9'u inceleyelim.

Eğitim HKO ile Test HKO İlişkisi (2)



Şekil 9: Yöntem esnekliği ile HKO arasındaki ilişki

Eğitim HKO ile Test HKO İlişkisi (3)

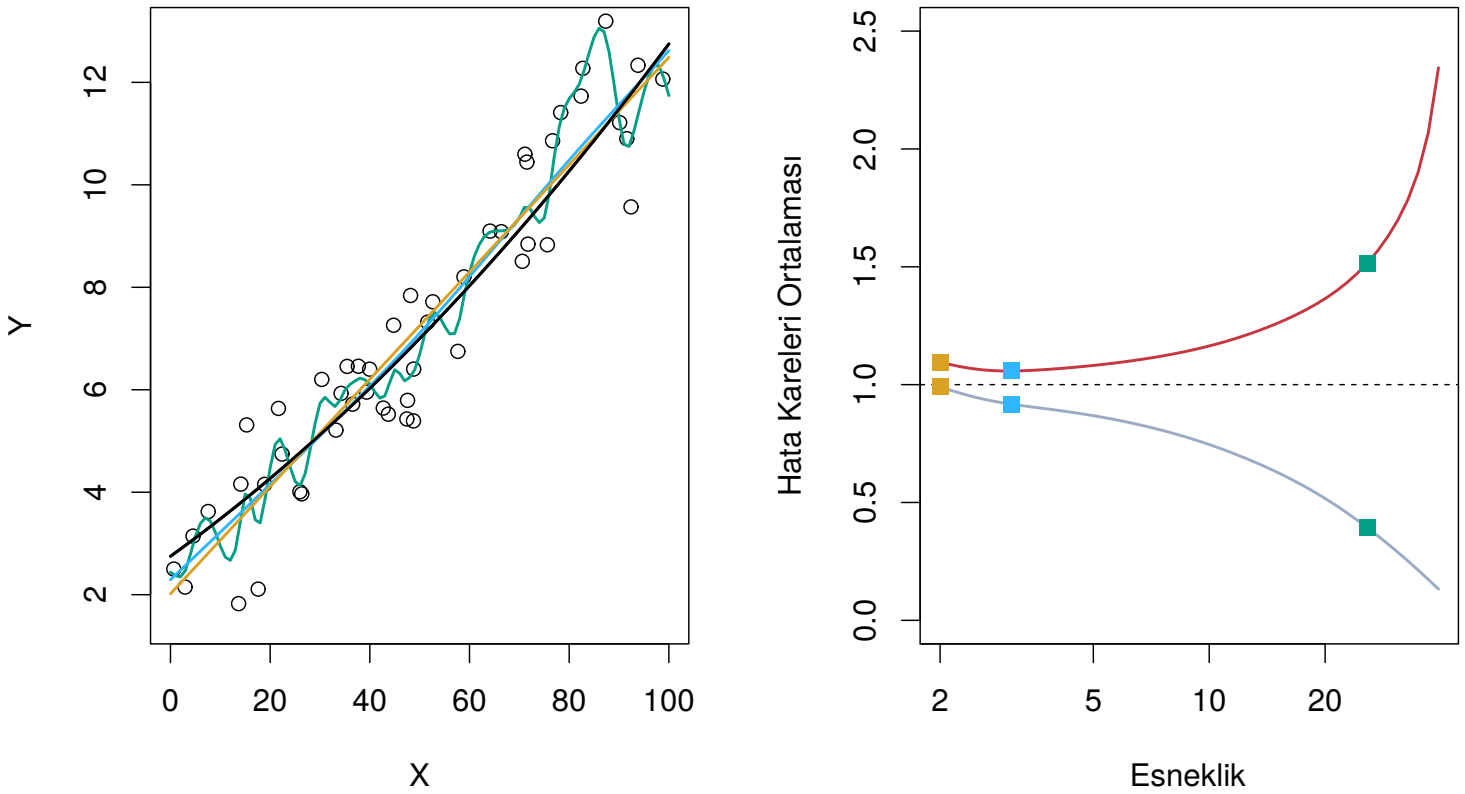
- Şekilde sol paneldeki küçük yuvarlaklar verilerdir. Siyah eğri ise bu verilerin geldiği gerçek f fonksiyonudur.
- Turuncu, mavi ve yeşil çizgiler ise esnekliği giderek artan üç farklı yöntemi göstermektedir.
- Turuncu çizgi doğrusal regresyon, mavi çizgi düzleştirme derecesi yüksek bir spline, yeşil çizgi ise düzleştirme derecesi düşük bir spline tahminidir.
- Sağ panelde bu üç yöntemin esneklikleri ve HKO değerleri görülmektedir. Bir çizginin esnekliğini ya da kıvrımlılığını serbestlik derecesi (degree of freedom) belirler. Burada bunlar 2, 6 ve 23'tür.
- Sağdaki gri renk eğri her bir yönteme ait eğitim HKO değerleridir.
- Bu veri seti belli bir formüle göre yapay olarak üretildiği için test verileri kolayca yaratılabilir. Dolayısıyla kırmızı eğri de buna göre hesaplanan test HKO değerleridir.
- Son olarak, ortadaki yatay çizgi ise hata teriminin varyansı olup, azaltılamayan minimum hata düzeyini belirtmektedir.

Eğitim HKO ile Test HKO İlişkisi (4)

- Şekili incelediğimizde kullanılan yöntemin esnekliği arttıkça eğitim HKO değerinin sürekli azaldığı görülmektedir.
- Öte yandan, test HKO değeri ise önce azalmakta ancak bir noktadan sonra artmaya başlamaktadır.
- Sürekli azalan eğitim HKO'suna karşılık U-şeklindeki test HKO'su istatistiksel öğrenmedeki temel olgulardan biridir.
- Bu durum her veri seti ve her yöntemde karşımıza çıkar.
- Bunun nedeni ise aşırı yakıştırma (overfitting) sorunudur. Kısaca, esneklik arttıkça yöntemin aşırı detaylı çalışmaya başlamasıdır.
- Böylece, bilinmeyen gerçek f fonksiyonunda gerçekte olmayan, rastlantısal oluşmuş değişiklikler içinde örüntü yakalamaya çalışır.
- Eğitim HKO ise sürekli düşer çünkü yöntem bunu minimize eder.
- Eğitim HKO'su ile test HKO'su arasındaki bu ilişki Şekil 10 ve Şekil 11'de farklı veri setleri için gösterilmiştir.

Yüksek Doğrusallık Durumunda Eğitim ve Test HKO

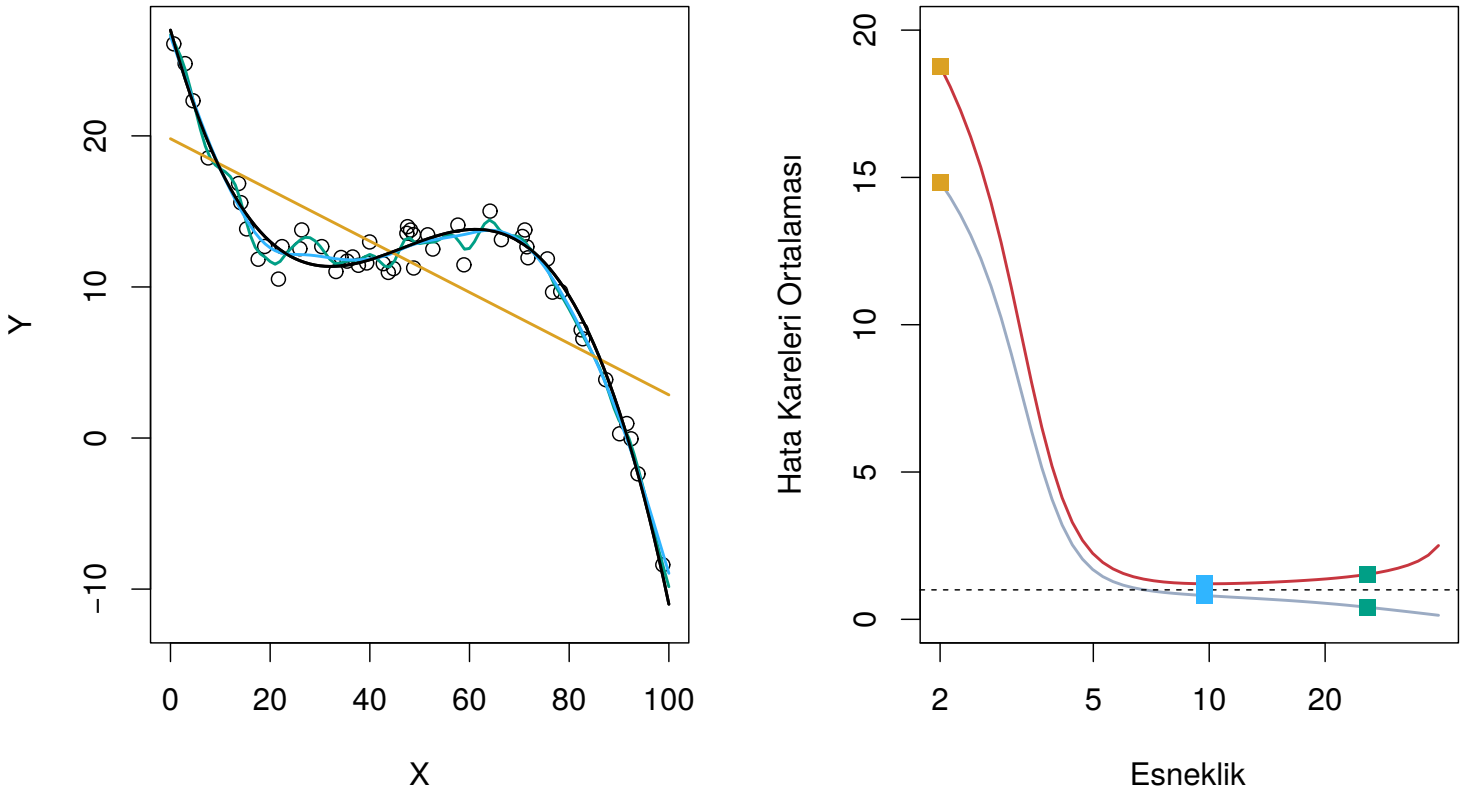
- Bu örnekte gerçek f doğrusala yakın olduğu için test HKO bir miktar azaldıktan sonra artmaya başlamaktadır. Dolayısıyla uygun yöntem doğrusal regresyondur.



Şekil 10: Yöntem esnekliği ile HKO arasındaki ilişki (yüksek doğrusallık)

Düşük Doğrusallık Durumunda Eğitim ve Test HKO

- Bu örnekte ise gerçek *f* düşük doğrusallık gösterdiği için test HKO 10 serbestlik derecesine kadar azalıp daha sonra artmaktadır. Burada uygun yöntem düzleştirme derecesi yüksek spline'dır.



Şekil 11: Yöntem esnekliği ile HKO arasındaki ilişki (düşük doğrusallık)

Yanlılık-Varyans Ödünleşmesi (1)

- Yukarıda gördüğümüz U-şeklindeki Test HKO'lar istatistiksel öğrenme yöntemlerindeki iki farklı özelliğin sonucudur.
- Bunu göstermek için beklenen test HKO formülünü şöyle yazalım:

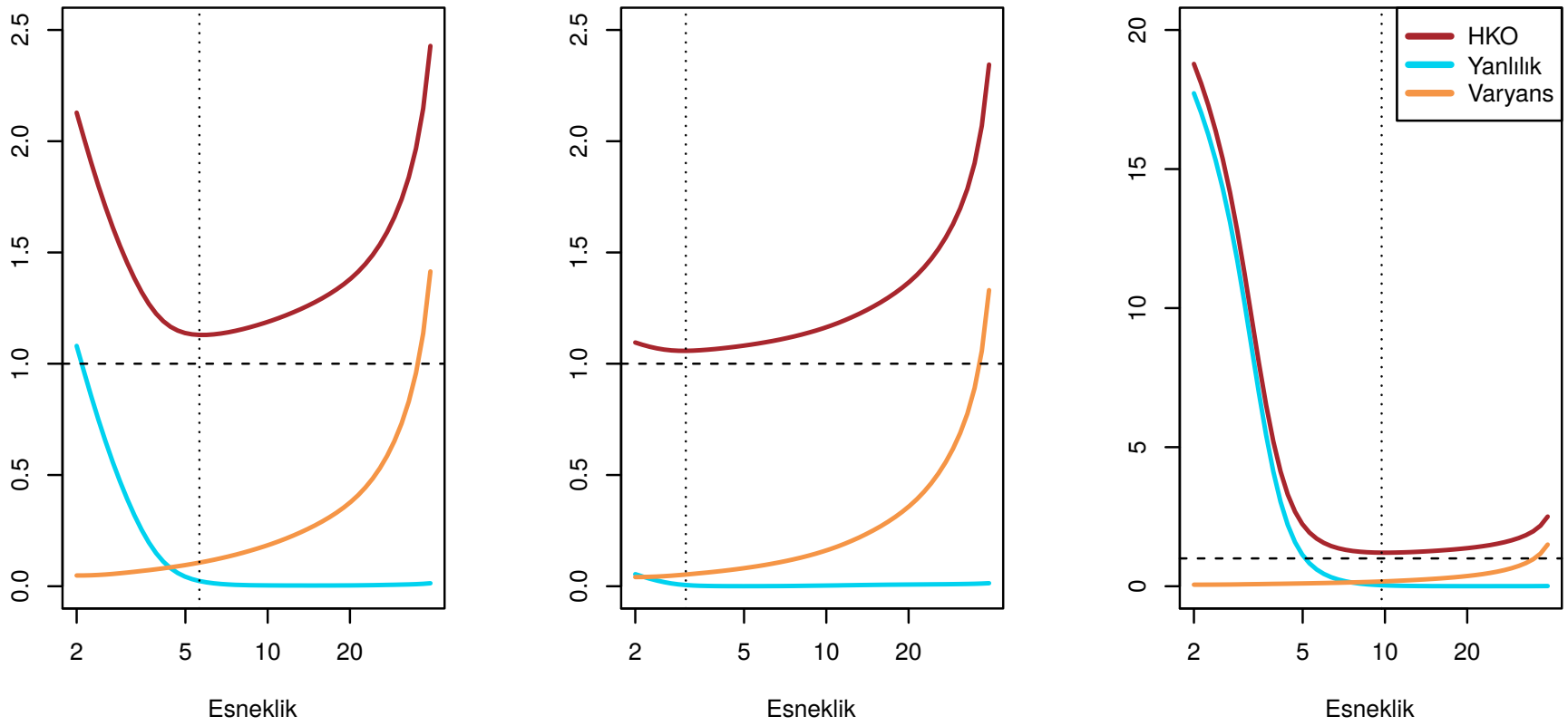
$$\mathrm{E}\left((y_0 - \hat{f}(x_0)\right)^2 = \mathrm{var}\left(\hat{f}(x_0)\right) + \text{Yanlılık}\left[\left(\hat{f}(x_0)\right)\right]^2 + \mathrm{var}(\epsilon)$$

- Yukarıdaki $\mathrm{E}\left((y_0 - \hat{f}(x_0)\right)^2$ ifadesi, test HKO'nun beklenen değeri (ortalama değer) anlamındadır.
- Sağdaki $\mathrm{var}(\epsilon)$ ise "azaltılamayan hata" değeridir. Diğer terimler negatif olamayacağı için test HKO da $\mathrm{var}(\epsilon)$'dan düşük olamaz.
- Formüle göre, düşük HKO için aynı anda hem düşük varyans hem de düşük yanlılık sağlayacak yöntemi kullanmamız gereklidir.
- Varyans, elimizde farklı bir eğitim veri seti olsaydı $\hat{f}$'nın ne kadar değişeceğini gösterir. Esnek yöntemlerde varyans daha yüksektir.
- Yanlılık ise gerçek hayatı görece basit bir modele indirgemekten kaynaklanır. Esnek yöntemlerde yanlılık genellikle düşüktür.

Yanlılık-Varyans Ödünleşmesi (2)

- Genel kural olarak, daha esnek yöntemler kullandıkça varyans artarken yanlılık da düşer. Test HKO değerinin artması ya da azalması bu ikisinin değişim hızına bağlıdır.
- Daha esnek bir yöntem kullandıkça ilk başlarda yanlılık hızla düşerken varyans ise çok artmaz.
- Ancak bir noktadan sonra esnekliği daha fazla artırmak yanlılığı çok etkilemeyip varyansın hızla artırmasına yol açar. Böylece, test HKO yükselmeye başlar.
- İşte, bu ilişkiye yanlılık-varyans ödünleşmesi (bias-variance trade-off) diyoruz.
- Bu ödünleşmeyi daha iyi anlamak için önceki örnekte gördüğümüz üç farklı eğitim-test HKO grafiklerini birlikte inceleyebiliriz. Bunlar Şekil 12'de verilmiştir.

Yanlılık-Varyans Ödünleşmesi (3)



Şekil 12: Yanlılık-varyans ödünleşmesi

Yanlılık-Varyans Ödünleşmesi (4)

- Şekildeki her üç panelde de kulanılan yöntemin esnekliği arttıkça yanlılık azalırken varyansın da arttığı görülüyor.
- Öte yandan, en düşük test HKO için gerekli esneklik derecesi üç örnekte de farklılık gösteriyor.
- Dik çizgiler ise test HKO'yu minimize eden esneklikleri veriyor.
- Buradan istatistiksel öğrenmedeki asıl zorluğun bu dik çizgilerdeki test HKO düzeyini sağlayan yöntemi bulmak olduğunu anlıyoruz.
- Bu derste göreceğimiz bazı yöntemler o kadar esnektir ki varyansı tümüyle yok edebilir. Ancak farklı uygulamalarda bunların basit yöntemlerden daha başarılı sonuç vereceğinin garantisi de yoktur.
- Sonuç olarak, veri çözümlemesinde yanlılık-varyans ödünleşmesini her zaman göz önünde bulundurmalıyız.

Sınıflandırma Çözümlemesi

- Model kesinliğinin ölçülmesine yönelik yukarıdaki tartışmamızda regresyon örneğini kullandık. Öte yandan, regresyon için vurguladığımız noktalar diğer yöntemler için de geçerlidir.
- Örnek olarak, istatistiksel öğrenmede sıkça kullandığımız bir diğer yaklaşım sınıflandırma (classification) çözümlemesidir.
- Sınıflandırmada da amacımız aynıdır. $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ şeklindeki eğitim veri setini kullanarak f fonksiyonunu tahmin ederiz.
- Ancak burada $y_1,\ldots,y_n$ tepki değişkeni nicel (quantitative) değil, nitel değerlerden oluşur.
- Nitel değişkenler, üniversite mezunu olup olmama ya da kadın ve erkek gibi farklı sınıflandırmaları gösterir.
- Bunlar farklı kategorileri belirten 0, 1, 2 gibi sabit ve kısıtlı değerler alır. Bu yüzden bunlara kategorik (categorical) değişken de denir.
- Bir veri setinde X'ler de Y'ler de kategorik olabilir. Ancak Y değişkeni eğer nitel ise bu durumda sınıflandırma çözümlemesi olur.

Hata Oranı

- Sınıflandırma çözümlemesinde eğitim HKO yerine eğitim hata oranı (training error rate) ölçütünden yararlanırız:

$$\text{Hata Oranı} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) = \text{Ortalama}\big(I(y_i \neq \hat{y}_i)\big)$$

- Yukarıda $\hat{y}_i$, $\hat{f}$'yı kullanarak i gözlemi için tahmin ettiğimiz sınıftır.
- $I(y_i \neq \hat{y}_i)$ ifadesine ise gösterge değişkeni (indicator variable) deriz. Belli bir i gözlemi için $y_i \neq \hat{y}_i$ olduğu zaman hatalı tahmin var demektir ve $I = 1$ olur. Aksi durumda $I = 0$ değerini alır.
- Yukarıdaki formüle eğitim hata oranı deriz çünkü hesaplarken elimizdeki eğitim verilerini kullanırız. Ancak, aslında ilgilendiğimiz şey test hata oranı (test error rate) değeridir:

$$\text{Hata Oranı}^{Test} = \text{Ortalama}\big(I(y_0 \neq \hat{y}_0)\big)$$

- Burada y_0, elimizde olmayan test verilerinden gelecek y'lerdir.
- En iyi sınıflandırma ise en düşük test hata oranını verendir.

Bayes Sınıflandırıcı (1)

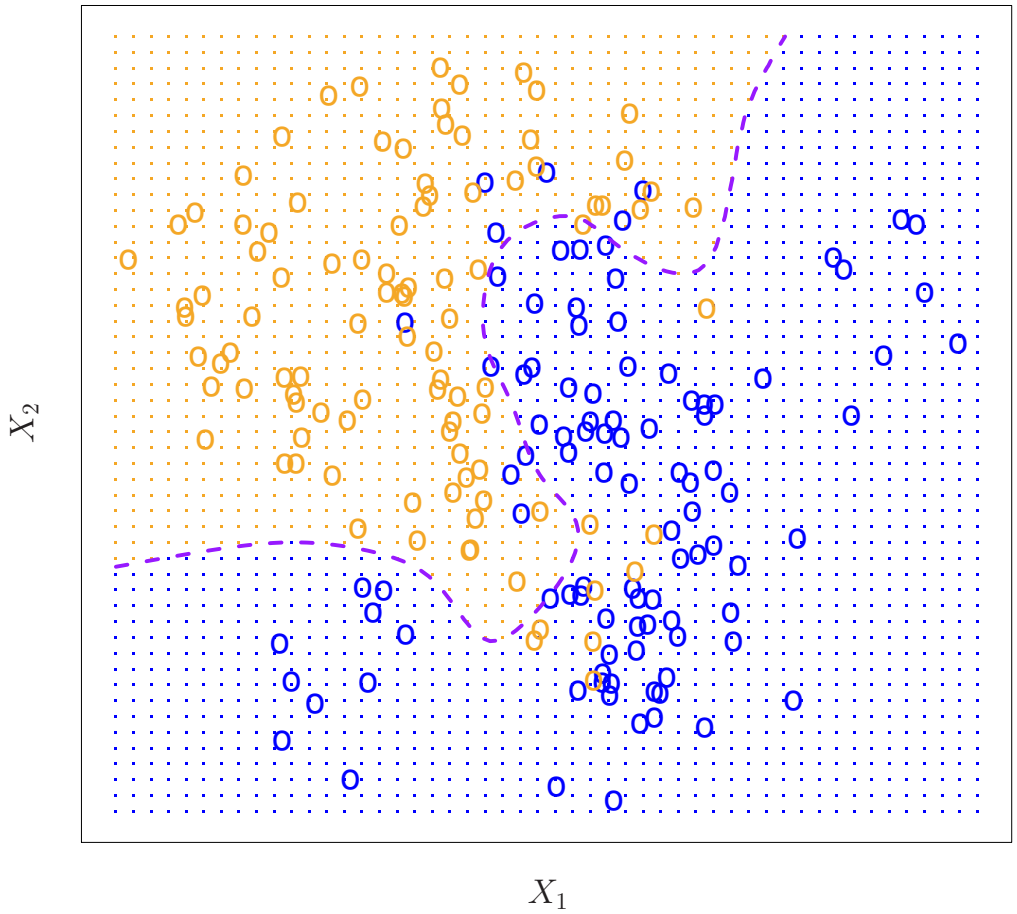
- Test hata oranını minimum yapan en ideal yöntem Bayes sınıflandırıcı (Bayes classifier) adı verilen olasılık hesaplamasıdır.
- Bu yöntemde her bir x_0 gözlemi için

$$\Pr(Y = j | X = x_0)$$

olasılığı maksimum olacak şekilde j sınıfı seçilir.

- Burada Pr, olasılık (probability) demektir. Parantezin içindeki | işareti ise "verili" şeklinde okunur.
- Dolayısıyla yukarıdaki ifade $X = x_0$ durumu verili iken $Y = j$ olma koşullu olasılığı (conditional probability) anlamına gelir.
- Görüldüğü gibi, Bayes sınıflandırıcının yaptığı şey her bir gözlem için en yüksek olasılığa sahip sınıfı seçmektir.
- İki sınıftan oluşan bir çözümlemede gerçekleşme olasılığı yüzde 50'den yüksek olan sınıf seçilir.
- Bu basit duruma yönelik bir örnek Şekil 13'te verilmiştir.

Bayes Sınıflandırıcı (2)



Şekil 13: Bayes sınıflandırıcıya göre yapılmış sınıflandırma örneği

Bayes Sınıflandırıcı (3)

- Şekilde X_1 ve X_2 değişkenlerinden oluşan simülasyon verileriyle ikili bir sınıflandırma çözümlemesi yapılmıştır.
- Turuncu ve mavi daireler iki farklı sınıfa ait gözlemlerdir. Farklı X_1 ve X_2 değerlerine bağlı olarak her bir gözlemin turuncu ya da mavi olma olasılığı farklıdır.
- Bu örnekte veriler belli bir formül kullanılarak yapay olarak oluşturulduğu için çok sayıda test verisi oluşturmak mümkündür. Bunu yaparak X_1 ve X_2 için koşullu olasılıkları hesaplayabiliriz.
- Bu işlem sonucunda turuncu olma olasılığı %50'den yüksek olan bölge turuncu noktalarla taranmıştır. $\Pr(Y = \text{mavi}|X_1, X_2) > \%50$ olan bölge de benzer şekilde mavi renkle taranmıştır.
- Ortadan geçen kesikli çizgiye ise Bayes karar sınırı (Bayes decision boundary) adı verilir.

Bayes Hata Oranı

- Bayes sınıflandırıcı en düşük test hata oranını veren yöntemdir.
- Ancak burada da eldeki verilerden kaynaklı olarak bir hata oranı söz konusudur. Buna Bayes hata oranı (Bayes error rate) denir:

$$\text{Bayes Hata Oranı} = 1 - \text{E}\left(\max_j \Pr(Y = j|X_1, X_2)\right)$$

- Formüle göre Bayes hata oranı, yukarıda açıkladığımız kural uygulanarak her sınıf için maksimum yapılan oranın 1'den farkıdır.
- Bu oran daha önce tartıştığımız "azaltılamayan hata" kavramı ile yakından ilişkilidir.

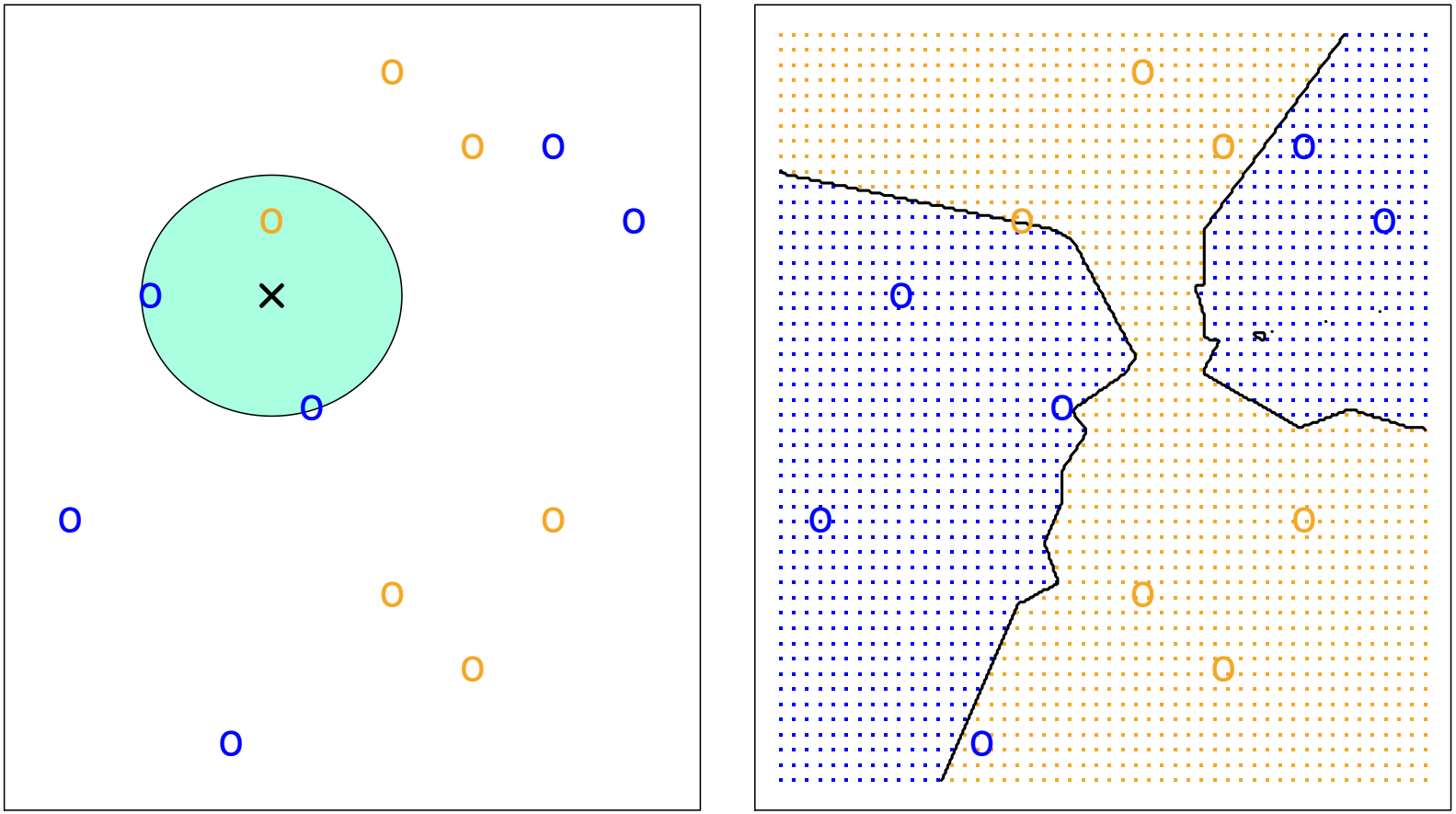
K-Enyakın Komşu (1)

- Uygulamada tüm sınıflandırmalarımızı Bayes hata oranı minimum olacak şekilde yapmak isteriz.
- Ancak gerçek hayatta Y'nin X'e bağlı koşullu olasılıklarını bilemediğimiz için Bayes sınıflandırıcıyı kullanmak olanaksızdır.
- Bunun yerine koşullu olasılıkları tahmin etme yoluna gideriz.
- Bu amaçla kullanılan en yaygın yöntemlerden biri K-enyakın komşu (K-nearest neighbor) ya da kısaca K-EK (K-NN) sınıflandırıcıdır.
- Bu yöntemde ilk önce pozitif tam sayı olan bir K değeri belirlenir.
- Daha sonra her bir x_0 gözlemi için bu gözleme en yakın diğer K adet gözlem seçilir. Böylece, N_0 adı verilen bu set içinden

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

oranı bulunarak her bir j'ye ait koşullu olasılık tahmin edilir.
- Son olarak, Bayes kuralı uygulanarak her bir gözlem için en yüksek olasılıklı sınıf belirlenir. Yöntem Şekil 14'te gösterilmiştir.

K-Enyakın Komşu (2)

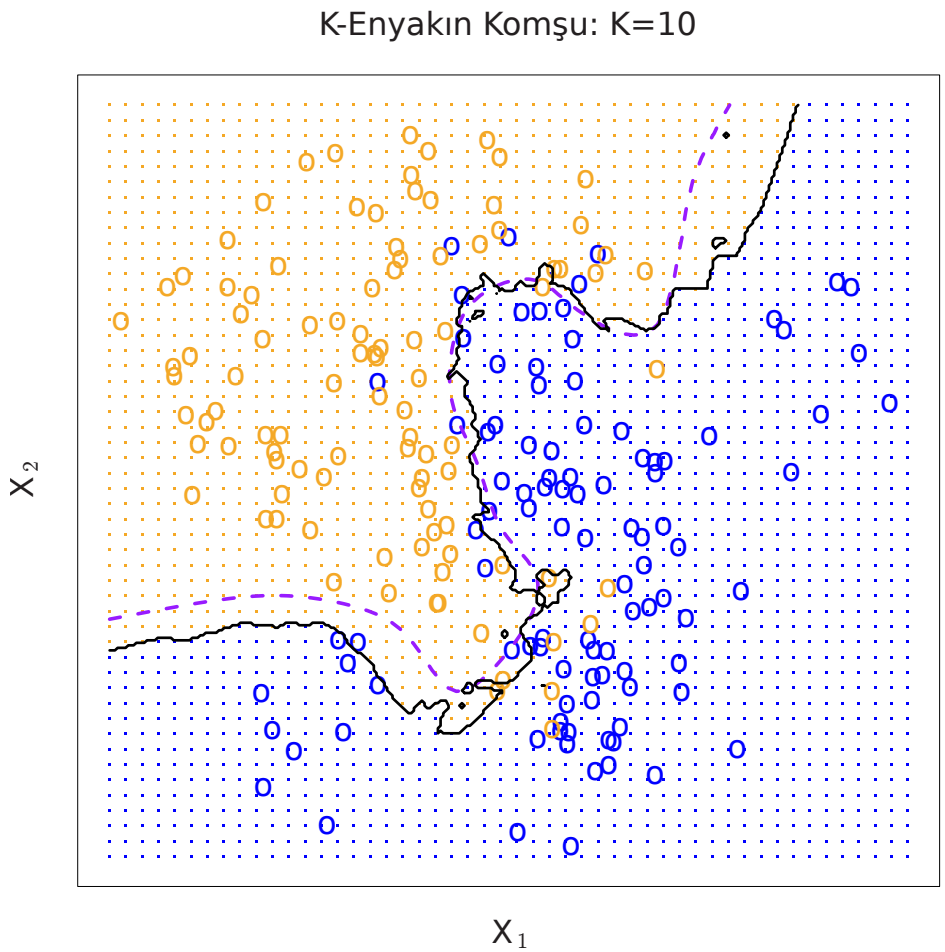


Şekil 14: K-enyakın komşu yöntemine göre yapılan sınıflandırma örneği

K-Enyakın Komşu (3)

- Şekildeki sol panelde K-EK yönteminin nasıl uygulandığı açıklanmıştır. Bu örnekte $K = 3$ olarak belirlenmiştir.
- Çarpı işareti ile gösterilen noktanın sınıfını tahmin etmek istediğimiz düşünelim. Bunun için en yakın 3 gözlemin sınıfına bakarız.
- Daire ile gösterilen alan içinde x noktasına en yakın 2 adet mavi ve 1 adet turuncu gözlem bulunmaktadır.
- Bu durumda mavi olasılığı yüzde 67, turuncu olasılığı ise yüzde 33'tür. Dolayısıyla çarpı noktası için tahminimiz de mavi olur.
- Bu işlemi şekildeki tüm noktalara uygulayarak sağ panelde gösterilen mavi ve turuncu bölgeleri hesaplayabiliriz.
- Böylece, bölgeleri ayıran K-EK karar sınırını da bulmuş oluruz.
- Simülasyon verileriyle yaptığımız yukarıdaki örnek için K-EK ile Bayes sınıflandırıcılarının karşılaştırması Şekil 15'te verilmiştir.

K-Enyakın Komşu (4)



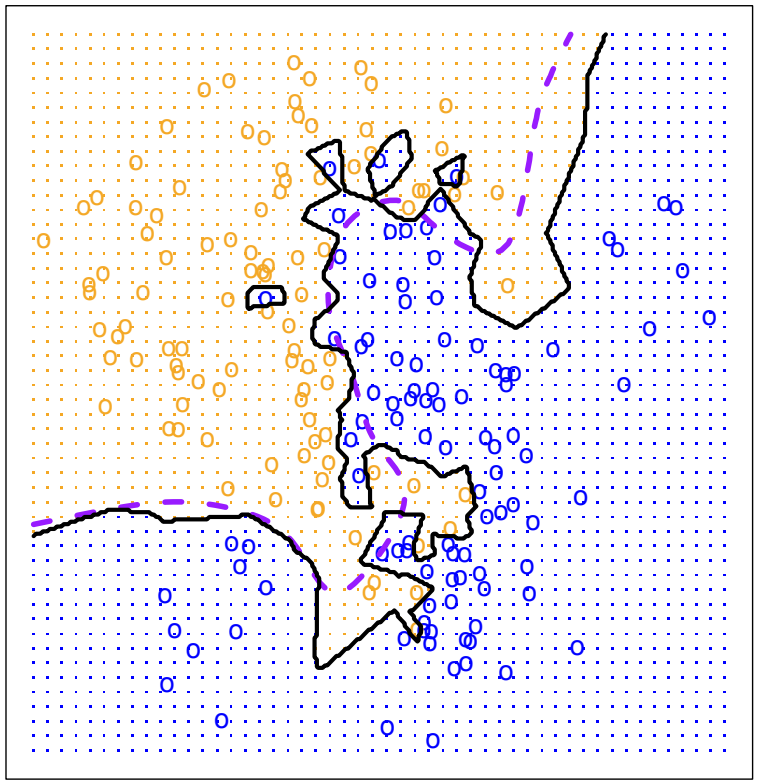
Şekil 15: K-enyakın komşu ile Bayes sınıflandırıcıların karşılaştırması

K-Enyakın Komşu (5)

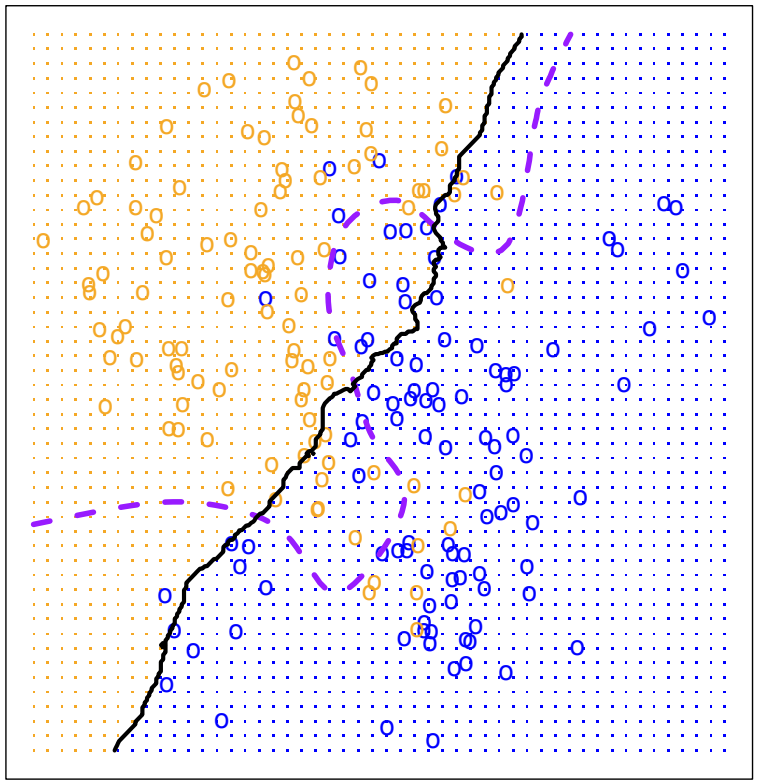
- Şekilde görüldüğü gibi, K-EK sınıflandırıcı uygulamada Bayes'e oldukça yakın sonuçlar üretebilmektedir.
- Ancak başta tartıştığımız yanlılık-varyans ödünleşmesi burada da geçerlidir. Bu da K-EK'in esnekliğini belirleyen K değeri ile yakından ilişkilidir.
- $K = 1$ ve $K = 100$ için elde edilen tahminler Şekil 16'da verilmiştir.
- Bu şekili incelediğimizde K eğer çok küçük olursa yöntemin aşırı esnek sonuçlar ürettiği görülmektedir. K çok büyük olduğunda ise doğrusala yakın, aşırı katı bir tahmin ortaya çıkmaktadır.
- Esneklik arttıkça eğitim ve test hata oranlarının nasıl değiştiği ise Şekil 17'de verilmiştir. Daha önce tartıştığımız U-şeklindeki test hata oranının burada da geçerli olduğuna dikkat ediniz.
- Sonuç olarak, tüm istatistiksel öğrenme yöntemleri için doğru esneklik düzeyini seçmek son derece önemlidir. En iyi esnekliği belirlemeye yarayan yöntemleri Bölüm 5'te göreceğiz.

K-Enyakın Komşu (6)

K-Enyakın Komşu: K=1

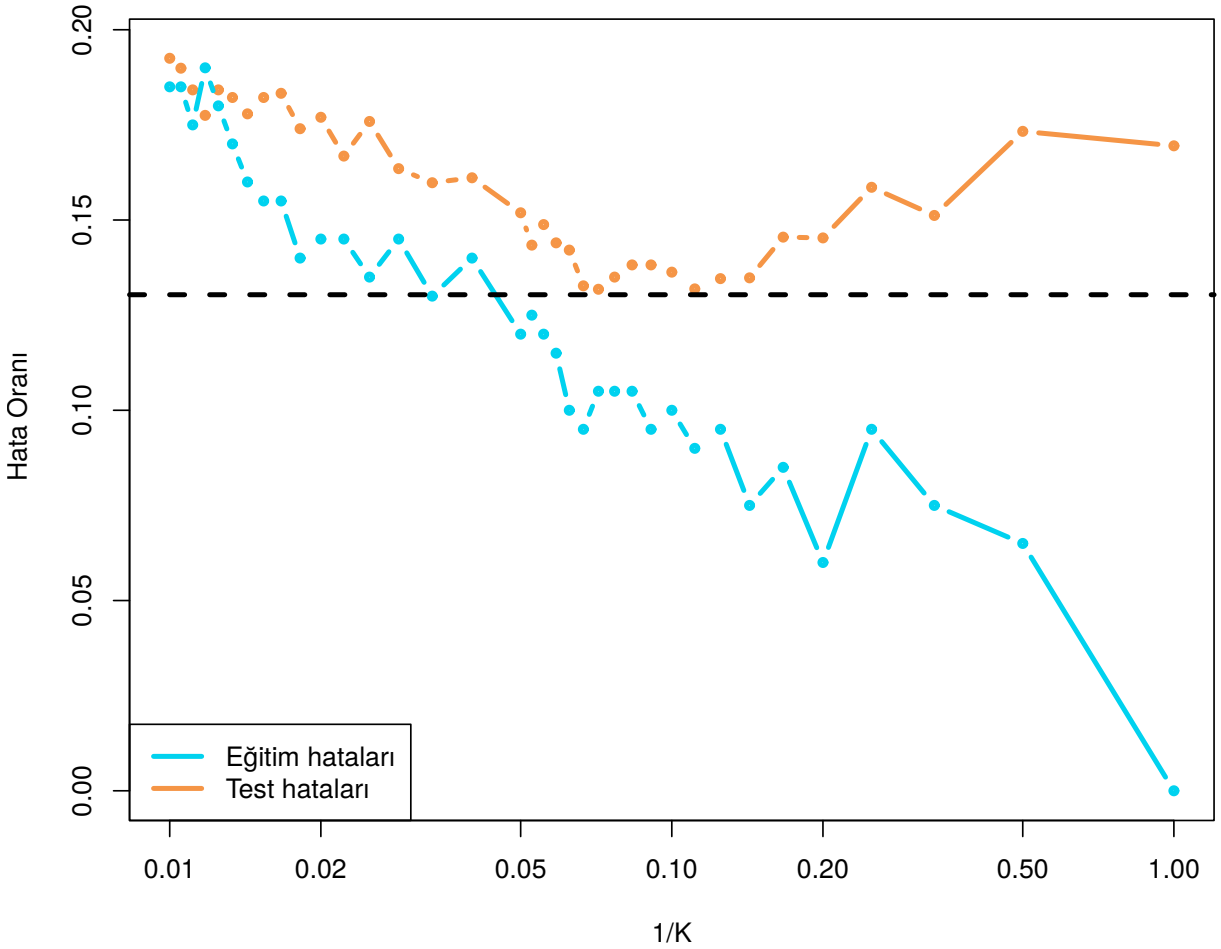


K-Enyakın Komşu: K=100



Şekil 16: $K = 1$ ve $K = 100$ için K-enyakın komşu karar sınırları

K-Enyakın Komşu (7)



Şekil 17: K-enyakın komşu yönteminde eğitim hata oranı ile test hata oranı

Önümüzdeki Dersin Konusu ve Ödev

Ödev

Kitaptan Bölüm 2 “İstatistiksel Öğrenme” okunacak.

Önümüzdeki Ders

Doğrusal Regresyon