

Sınıflandırma Çözümlemesi

A. Talha Yalta

TOBB Ekonomi ve Teknoloji Üniversitesi

İKT-457 Ekonomi ve Finans İçin Yapay Zeka 1
Sürüm 0,93



Bu belge “Creative Commons Attribution-ShareAlike 3.0 Unported” (CC BY-SA 3.0) lisansı altında bir açık ders malzemesi olarak genel kullanıma sunulmuştur. Bazı şekiller “An Introduction to Statistical Learning, with applications in R” (Springer, 2017) kitabından yazarların izniyle alınmıştır. Tüm belge eserin ilk sahibinin belirtilmesi ve geçerli lisansın korunması koşuluyla özgürce kullanılabilir, çoğaltılabilir, ve değiştirilebilir. Creative Commons örgütü ve CC-BY-SA 3.0 lisansı ile ilgili ayrıntılı bilgi <https://creativecommons.org> Internet adresinde yer almaktadır. Ders notlarımın güncel sürümlerine <http://yalta.etu.edu.tr> adresinden ulaşabilirsiniz.

A. Talha Yalta

TOBB Ekonomi ve Teknoloji Üniversitesi

2020 – 2021 



1 Doğrusal Olasılık Modeli

- DOM tahminine yönelik eleştiriler

2 Lojistik Regresyon

- İkili lojistik model
- Lojistik tahmin ve kestirim
- Çoklu lojistik model

3 Diskriminant Çözümlemesi

- İkili doğrusal DÇ modeli
- Çoklu doğrusal DÇ modeli
- Almaşık DÇ modelleri

4 Diğer Konular

- Sınıflandırma kesinliğinin ölçülmesi
- Sınıflandırıcıların karşılaştırılması



Sınıflandırma Çözümlemesi (1)

- 3. Bölümde gördüğümüz regresyon yönteminde Y tepki değişkeni **nicel** (quantitative) değerlere sahipti.
- Ancak birçok durumda Y bir **nitel** (qualitative) değişken de olabilir.
- Bu bölümde nitel tepkiyi kestirmeye yönelik **sınıflandırma** (classification) çözümlemesini ele alacağız.
- Sınıflandırma uygulamada en az regresyon kadar yaygındır. Birkaç örnek olarak, aşağıdaki konu başlıklarını gösterebiliriz:
 - 1 Bir kredi başvurusunun ret edilip edilmeyeceği
 - 2 Bir şirketin ele geçirmeye hedef olup olmayacağı
 - 3 Şirketlerin hisse senedi çıkartıp çıkartmayacakları
 - 4 Belli bir ürünün iade edilip edilmeyeceği
 - 5 Kişilerin sigorta yaptırap yaptırmayacakları
 - 6 Bir kişinin işgücüne katılıp katılmayacağı
 - 7 Sendika üyesi olunup olunmayacağı



Sınıflandırma Çözümlemesi (2)

- Sınıflandırma çözümlemesinde kullanılan araçlar temelde Y 'nin belli bir sınıfa ait olma olasılığını tahmin etmeye dayanır.
- Dolayısıyla yöntem regresyona benzemektedir.
- Elimizde yine $(x_1, y_1), \dots, (x_n, y_n)$ şeklinde bir eğitim veri seti vardır. Amacımız bunları kullanarak bir sınıflandırıcı oluşturmaktır.
- Bu sınıflandırıcının yalnızca *eğitim* verilerinde değil, *test* verilerinde de başarı göstermesi önemlidir.
- Ayrıca burada Y değişkeni nicelik belirtmediği için sıradan enküçük kareler (SEK) dışında yöntemler de geliştirilmiştir.



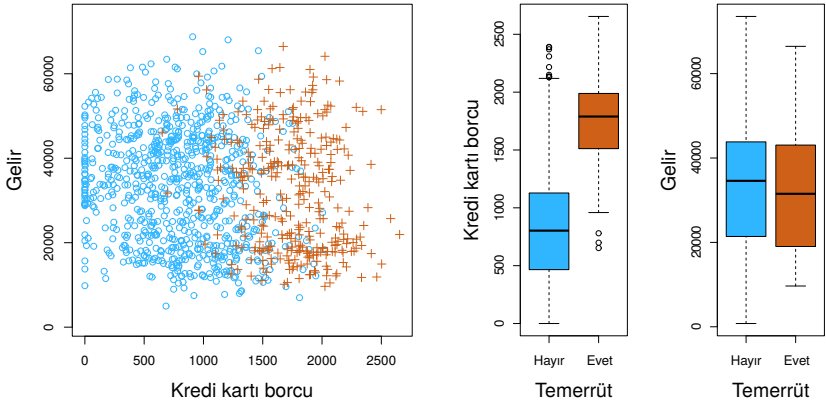
Temel Sınıflandırma Yöntemleri

- Bu bölümde aşağıdaki beş temel sınıflandırma yöntemini sırayla ele alacağız:
 - 1 Doğrusal olasılık modeli
 - 2 Lojistik regresyon
 - 3 Doğrusal diskriminant çözümlemesi
 - 4 İkinci derece diskriminant çözümlemesi
 - 5 Sade Bayes sınıflandırıcı
- Bunlara ek olarak, 2. Bölümde açıkladığımız K-enyakın komşu yöntemini de bölüm sonunda diğerleriyle karşılaştıracacağız.
- Örnek uygulama olarak ise önceki bölümlerde de gördüğümüz kredi veri setini kullanacağız.
- Farklı kişilere ait gelir, kredi kartı borcu ve borç ödeyip ödememe bilgileri Şekil 1'de gösterilmiştir.



Kredi Kartı Verileri

- Sol paneldeki serpilim çiziminde farklı kişilerin kredi kartı borçları ile gelirleri verilmiştir. Sağ panelde ise borcu yüksek olanların temerrüte düşme olasılığının daha yüksek olduğu anlaşılmaktadır.



Şekil 1: Gelir, kredi kartı borcu ve borç ödeyip ödememe verileri



Ders Planı

1 Doğrusal Olasılık Modeli

- DOM tahminine yönelik eleştiriler

2 Lojistik Regresyon

- İkili lojistik model
- Lojistik tahmin ve kestirim
- Çoklu lojistik model

3 Diskriminant Çözümlemesi

- İkili doğrusal DÇ modeli
- Çoklu doğrusal DÇ modeli
- Almaşık DÇ modelleri

4 Diğer Konular

- Sınıflandırma kesinliğinin ölçülmesi
- Sınıflandırıcıların karşılaştırılması



Doğrusal Olasılık Modeli (1)

- Şekildeki ilişkiyi sıradan enküçük kareler (SEK) yöntemi ile tahmin etmek istersek aşağıdaki gibi bir model belirtebiliriz:

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- Burada X kredi kartı borcudur. $Y = 1$ temerrüte düşüldüğünü, $Y = 0$ ise borcun ödendiğini göstermektedir.
- Bir kukla değişken olan Y 'yi X 'in doğrusal fonksiyonu olarak belirten yukarıdaki modele “doğrusal olasılık modeli” (linear probability model), ya da kısaca “DOM” (LPM) adı verilir.
- Bu modelde $\hat{Y}$ tahminleri X veriliyken $Y = 1$ olayının gerçekleşmesine yönelik koşullu olasılık (conditional probability) olarak yorumlanır:

$$\Pr(Y_i = 1 | X_i)$$



Doğrusal Olasılık Modeli (2)

- Elimizdeki modele neden doğrusal olasılık denildiğini görmek için hata terimi ϵ 'un ortalamasının 0 olduğunu anımsayalım. Bu durumda şunu yazabiliriz:

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i$$

- $Y_i = 1$ olduğunda temerrüt gerçekleşmektedir. Bunun olasılık değerine P_i dersek Y 'nin olasılık dağılımı aşağıdaki gibi olur:

Y	Olasılık
1	P_i
0	$1 - P_i$
Toplam	1

- Beklenen değer tanımından yararlanarak şunu görebiliriz:

$$E(Y_i|X_i) = 0(1 - P_i) + 1(P_i) = P_i$$

- Görüldüğü gibi $E(Y_i|X_i)$ bir olasılıktır. Dolayısıyla $0 \leq E(Y_i) \leq 1$ şeklinde bir sınırlama söz konusudur.



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



DOM Tahminindeki Sorunlar

- Eğer amacımız çıkarsama ise DOM tercih edilen yaklaşım değildir. Bunun temelde dört nedeni vardır:
 - 1 Sınıflar arası ilişkinin sayısal olmaması
 - 2 Hata teriminin normal-dışılığı
 - 3 Hatalarda farklı serpilimsellik görülmesi
 - 4 $0 \leq E(Y_i|X_i) \leq 1$ koşulunun sağlanamaması
- Öte yandan, günümüzde büyük veri setleri ve yoğun hesaplama gerektiren yapay zeka uygulamalarında DOM yöntemi sınıflandırma amacı için uygun bir seçenek olabilir.
- Ayrıca DOM kullanılarak elde edilen sınıflandırmalar bu bölümde göreceğimiz diğer bir yöntem olan doğrusal diskriminant çözümlemesi ile aynı sonuçları vermektedir.
- Şimdi DOM yönteminin eleştirilen yanlarından kısaca söz edelim.
- Böylece, lojistik model gibi farklı araçlara neden gerek duyulduğunu da daha iyi görebiliriz.



Sınıflar Arası İlişkinin Sayısal Olmaması

- DOM'a yönelik ilk eleştiri, sınıflandırma için Y 'ye sayısal değerler versek de ilişkinin aslında sayısal olmamasıyla ilgilidir.
- Ayrıca çoğu zaman ikiden fazla sınıf söz konusudur.
- Örnek olarak, İnternet üzerindeki satışların hangi ilden gerçekleşeceğini kestirmek istiyor olalım:

$$Y = \begin{cases} 1 & \text{eğer İstanbul ise} \\ 2 & \text{eğer Ankara ise} \\ 3 & \text{eğer İzmir ise} \end{cases}$$

- Yukarıda tepki değişkeninin aldığı değerlerin belli bir önem ya da oluş sırasına sahip olmadığına dikkat ediniz.
- Diğer bir deyişle, Ankara İstanbul'un iki katıdır ya da İzmir Ankara'dan yüzde 50 daha fazladır diyemeyiz.
- Dahası bunların sırasını değiştirirsek tahminler de değişir.
- Dolayısıyla ikiden fazla sınıf olması durumunda DOM tahmini uygun bir sınıflandırma yöntemi değildir.



Hata Teriminin Normal-dışılığı

- İkinci olarak, SEK yönteminde çıkarsama için hata terimi ϵ_i 'nin normal dağıldığı varsayımının önemli olduğunu anımsayalım.
- Ancak DOM tahmininde ϵ_i 'nin normal dağılması olanaksızdır.
- Aşağıda da görüldüğü gibi, Y_i 'ler yalnızca iki değer aldığı için iki farklı ϵ_i kümesi ortaya çıkar:

$$\epsilon_i = Y_i - \beta_1 - \beta_2 X_i$$

$$Y_i = 1 \quad \text{ise} \quad \epsilon_i = 1 - \beta_1 - \beta_2 X_i$$

$$Y_i = 0 \quad \text{ise} \quad \epsilon_i = -\beta_1 - \beta_2 X_i$$

- Burada ϵ_i 'ler normal dağılıma değil, Bernoulli dağılımına uyar.
- Öte yandan nokta tahminleri yansız olmayı sürdürür.
- Ayrıca merkezi limit kanıtına göre örneklem büyüklüğü artarken kalıntıların normale yaklaşacağı da unutulmamalıdır.
- Dolayısıyla büyük örneklemelerde ϵ 'nin normal-dışılığı kestirim ve çıkarsama açısından bir sorun yaratmayabilir.



Hatalarda Farklıserpilimsellik

- Üçüncü olarak, hata teriminin Bernoulli dağılımına uyduğu bilindiğine göre, DOM tahmininde ϵ_i 'lerin aynıserpilimsel olduğu varsayımını korumak da olanaksızdır.
- İkiterimli dağılımın genel biçimi olan kesikli Bernoulli dağılımının ortalaması p , varyansı $p(1 - p)$ 'dir.
- Buna göre doğrusal olasılık modelinin varyansı da şu olur:

$$\text{var}(\epsilon_i) = P_i(1 - P_i)$$

- $P_i = E(Y_i|X_i) = \beta_1 + \beta_2 X_i$ olduğuna göre, ϵ_i sonuçta X_i değerlerine bağlıdır ve bu nedenle aynıserpilimsel olamaz.
- Farklıserpilimsellik altında SEK yansız olmayı sürdürmekle birlikte artık en iyi (enaz varyanslı) yöntem değildir.
- Ancak büyük örneklemelerde bu da DOM için sorun olmayabilir. Farklıserpilimsellik, ağırlıklı enküçük kareler ya da daha gelişmiş sağlam ölçünlü hata tahminleri kullanarak aşılmaya çalışılabilir.



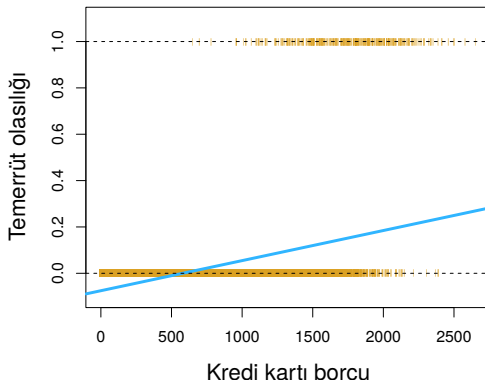
$0 \leq E(Y_i|X_i) \leq 1$ Koşulunun Sağlanmaması

- Son olarak, DOM tahmini için bir diğer sorun $0 \leq E(Y_i|X_i) \leq 1$ koşulunun sağlanamamasıdır.
- Bu modeller X veriliyken Y olayının gerçekleşme koşullu olasılığını ölçtüğü için, $E(Y_i|X_i)$ değerinin 0 ile 1 arasında yer alması önemlidir.
- SEK yöntemi böyle bir matematiksel sınırlama içermediği için DOM tahmini sonrasında yakıştırılan değerlerin 0'dan küçük ya da 1'den büyük çıkmasına sıkça rastlanır.
- Öte yandan böyle durumlarda eksi değerli $\hat{Y}_i$ 'leri sıfır, 1'den büyük $\hat{Y}_i$ 'leri ise 1 varsaymak yoluna da gidilebilir.
- Sonuç olarak, DOM aslında gözden düşmüş bir yöntem gibi görünse de günümüzde İnternet ölçeğinde veri setleri ve hesaplama yoğun tekniklerle birlikte kullanmak için potansiyel vadetmektedir.
- Varsayımsal örneğimize ilişkin DOM tahmini Şekil 2'de verilmiştir.



DOM Yakıştırması

- Aşağıda görülen DOM tahminine göre temerrüte düşme olasılığı borçla birlikte doğrusal olarak artmaktadır. 500'den küçük borçlar için negatif olasılık tahmin edildiğine dikkat ediniz.



Şekil 2: Temerrüt olasılığı ile kredi kartı borcu DOM tahmini



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Lojistik Model

- DOM tahminine alması olarak, kestirilen olasılıkların 0 ve 1 arasında olmasını sağlayan sınıflandırma yöntemleri de mevcuttur.
- Bunlar içinde **lojistik regresyon** (logistik regression) yöntemi, Y tepki değişkenini doğrudan tahmin etmek yerine Y 'nin belli bir sınıfa ait olma *olasılığını* bulmayı amaçlar:

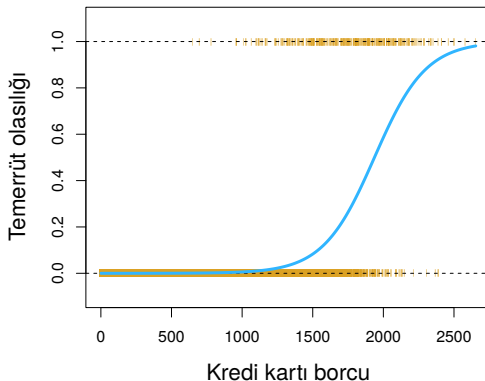
$$p(X_i) = \beta_1 + \beta_2 X_i + \epsilon_i$$

- Yukarıda $p(X_i)$, $\Pr(Y_i = 1 | X_i)$ koşullu olasılığıdır.
- Bu modelde kestilen $\hat{Y}_i$ değerleri her zaman $[0,1]$ aralığında olur.
- Dolayısıyla temerrüte düşüp düşmeme sınıflandırmasını $\Pr(\hat{Y}_i) = 0,5$ eşik değerine göre yapabiliriz.
- Öte yandan söz gelimi bir banka eğer kredi verme konusunda daha seçici davranmak istiyorsa temerrüt riski olan kişileri belirlemek için 0,4 gibi daha düşük bir düzey de seçebilir.
- Örneğimize ilişkin lojistik model tahmini Şekil 3'teki gibidir.



Lojistik Model Yakıştırması

- Lojistik regresyon sonuçlarına göre borç arttıkça temerrüte düşme olasılığı önce artarak artmakta, daha sonra artış hızı düşmektedir. Böylece, olasılık her zaman $[0,1]$ aralığında tahmin edilmektedir.



Şekil 3: Temerrüt olasılığı ile kredi kartı borcu lojistik model tahmini



Lojistik Fonksiyon (1)

- Lojistik regresyon, $0 \leq E(Y_i) \leq 1$ sınırlamasını sağlamak için aşağıdaki lojistik fonksiyonu kullanır:

$$p(X_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

- Bu fonksiyonu şu şekilde de yazabiliriz:

$$\frac{p(X_i)}{1 - p(X_i)} = e^{\beta_0 + \beta_1 X_i}$$

- Yukarıdaki orana **bahis oranı** (odds ratio) denir. Her zaman $(0, \infty)$ aralığında olan bu oran at yarışı bahislerinde kullanılmaktadır.
- Örnek olarak, $p(X) = 0,2$ olursa bahis oranı $\frac{0,2}{1-0,2} = 1/4$ olur. Bu oran ortalama 5 kişiden 1'inin temerrüte düşeceğini söyler.
- Benzer şekilde $p(X) = 0,9$ için bahis oranı $\frac{0,9}{1-0,9} = 9$ 'dur. Bu da 9'a 1 anlamına gelir. Diğer bir deyişle 10 kişiden dokuzu demektir.



Lojistik Fonksiyon (1)

- Bahis oranının her iki tarafının logaritmasını alırsak aşağıdaki denklemini elde ederiz:

$$\log \left(\frac{p(X_i)}{1 - p(X_i)} \right) = \beta_0 + \beta_1 X_i$$

- Yukarıda sol yanda görülen ifadeye **logaritmik birim** (logarithmic unit) ya da **logbirim** (logit) adı verilir.
- Görüldüğü gibi, lojistik regresyon modeli X 'e göre doğrusal bir logbirim fonksiyonuna sahiptir.
- Doğrusal regresyon modelinde X 'teki 1 birim değişikliğin Y 'yi β_1 kadar etkilediğini anımsayınız.
- Bundan farklı olarak, lojistik modelde ise X , 1 birim arttığı zaman $Y = 1$ olayının gerçekleşme “log-bahis oranı” β_1 kadar artar.
- Aradaki ilişki doğrusal olmadığı için X 'in 1 birim artmasının yarattığı etki X 'in baştaki büyüklüğüne bağlıdır. Dolayısıyla lojistik modeli yorumlamak görece daha zordur.



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - **Lojistik tahmin ve kestirim**
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Ençok Olabilirlik Yöntemi

- Lojistik modelde β_0 ve β_1 parametrelerini tahmin etmek için **ençok olabilirlik** (maximum likelihood) yöntemi kullanılır.
- Yöntemin mantığı basittir: Eldeki eğitim verilerinin gerçekleşme olasılığı maksimum olacak şekilde $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri tahmin edilir.
- Diğer bir deyişle, farklı gözlemler için X_i değerini denkleme koyunca temerrüte düşenler için $p(X_i) \approx 1$ ve düşmeyenler için de $p(X_i) \approx 0$ sonucunu vermeye en yakın parametre seti hesaplanır.
- Bu amaçla aşağıdaki **olabilirlik fonksiyonu** (likelihood function) maksimize edilir:

$$l(\beta_0, \beta_1) = \prod_{i: y_i=1} p(x_i) \prod_{i': y_{i'}=0} (1 - p(x_{i'}))$$

- Doğrusal regresyonda gördüğümüz sıradan enküçük kareler yaklaşımı da aslında ençok olabilirliğin özel bir durumudur.
- Ençok olabilirlik tahmini bilgisayarlar tarafından kolayca yapılmaktadır. Dolayısıyla burada yöntemin ayrıntılarına girmeyeceğiz.



İkili Lojistik Regresyon Bilgisayar Çıktısı

- Örneğimizdeki lojistik model tahmininin bilgisayar çıktısı şöyledir:

Değişken	Katsayı	Ölçünlü hata	Z-istatistiği	p-değeri
Sabit terim	-10,6513	0,3612	-25,5	< 0,0001
Kart borcu	0,0055	0,0002	24,9	< 0,0001

Çizelge 1: Temerrüt olasılığı ile kredi kartı borcu lojistik regresyonu

- Çizelgenin SEK tahmin sonuçlarına benzediğine dikkat ediniz.
- İkinci satırda kart borcunun katsayısı pozitif bulunmuştur. Buna göre borç miktarı arttıkça temerrüte düşme olasılığı artmaktadır.
- Z-istatistiği, daha önce gördüğümüz t -istatistiği gibi $H_0 : \beta = 0$ önsavını sınar. Kabaca 2'den büyük değerler %95 güvenle β 'nin 0'dan anlamlı derecede uzak olduğunu gösterir.
- Burada her iki istatistiğe ait p -değeri on binde birden küçük bulunmuştur. Dolayısıyla β_0 ve β_1 'in anlamlı olduğu sonucuna ulaşırız.
- Sabit terim bu modelde önemli değildir. Tahmin edilen olasılıkların ortalamasını veri setine ayarlama işlevi görmektedir.



Lojistik Modelde Kestirim

- Lojistik modeli tahmin ettikten sonra farklı kredi kartı borç durumlarına ait temerrüt olasılığını kolayca hesaplayabiliriz.
- Örnek olarak, 1000 dolar borç düzeyi için temerrüt olasılığı şudur:

$$p(\hat{X}) = \frac{e^{-10,6513+0,0055 \times 1000}}{1 + e^{-10,6513+0,0055 \times 1000}} = 0,006$$

- Görüldüğü gibi bu borç düzeyinde temerrüt olasılığı yüzde 1'den küçük bulunmuştur.
- Öte yandan borç 2000 dolara çıktığı zaman olasılık çok daha yüksek tahmin edilir:

$$p(\hat{X}) = \frac{e^{-10,6513+0,0055 \times 2000}}{1 + e^{-10,6513+0,0055 \times 2000}} = 0,586$$

- Demek ki borç düzeyi 1000 dolar iken 1000 dolar daha borçlanması durumunda temerrüt olasılığı $58,6 - 0,6 =$ yüzde 58,0 artmaktadır.



Lojistik Modelde Kukla Değişkenler

- Lojistik modelde kukla değişkenleri de kullanabiliriz.
- Örnek olarak, X değişkeni borç yerine öğrenci olup olmamayı gösteren bir kukla olsun. Bu durumda şu sonuçları elde ederiz:

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	-3,5041	0,0707	-49,55	< 0,0001
Öğrenci (=1)	0,4049	0,1150	3,52	0,0004

Çizelge 2: Temerrüt olasılığı ile öğrenci olma durumu lojistik regresyonu

- Burada öğrenci olan ve olmayanlar için aşağıdaki temerrüt olasılıkları kestirilir:

$$\Pr(\text{Temerrüt} = 1 | \text{öğrenci} = 1) = \frac{e^{-3,5041+0,4049 \times 1}}{1 + e^{-3,5041+0,4049 \times 1}} = 0,0431$$

$$\Pr(\text{Temerrüt} = 1 | \text{öğrenci} = 0) = \frac{e^{-3,5041+0,4049 \times 0}}{1 + e^{-3,5041+0,4049 \times 0}} = 0,0292$$

- Buna göre öğrenciler için borç ödememe riski daha yüksektir.



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Çoklu Lojistik Model

- Doğrusal regresyonda olduğu gibi lojistik regresyonda da genellikle birden fazla X değişkeni yer alır:

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- Yukarıdaki çoklu model de yine, en çok olabilirlik ile tahmin edilir.
- Kredi kartı borcu, gelir ve öğrenci olma durumunu açıklayıcı değişken olarak alan logbirim tahmini aşağıdaki gibidir:

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	-10,8690	0,4923	-22,08	< 0,0001
Kart borcu	0,0057	0,0002	24,74	< 0,0001
Gelir	0,0030	0,0082	0,37	0,7115
Öğrenci (=1)	-0,6468	0,2362	-2,74	0,0062

Çizelge 3: Temerrüt olasılığına ilişkin çoklu lojistik regresyon

- Sonuçlar kart borcu ile öğrenci olma durumunun temerrüt olasılığını anlamlı derecede etkilediğini göstermektedir.



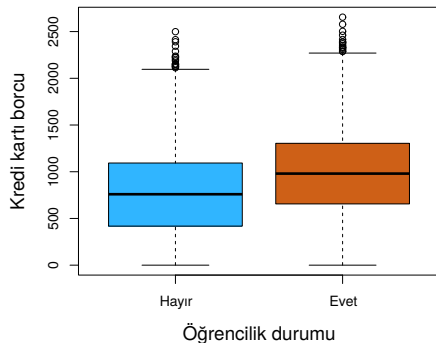
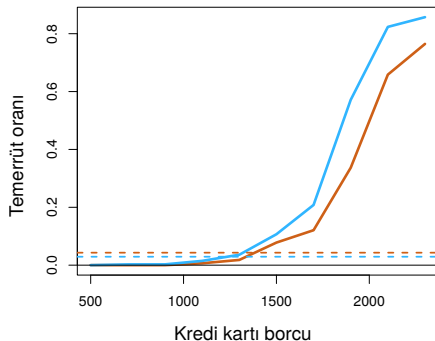
Karıştırıcı Değişken Etkisi (1)

- Çoklu modelde öğrenci olma katsayısının negatif olduğuna dikkat ediniz. Baştaki ikili modelde ise öğrencilerin temerrüte düşme olasılığını daha yüksek bulmuştuk. Bu bir çelişki değil mi?
- Hayır! Bazı değişkenlerin modele koyulmadığı zaman yanıltıcı sonuçlara yol açtığını ve bunlara **karıştırıcı değişken** (confounding variable) dediğimizi önceki bölümden anımsayınız.
- İşte burada da kredi kartı borcu bir karıştırıcı değişkendir.
- Öğrencilerin daha fazla temerrüte düşmesinin nedeni bu grubun borç düzeyinin daha yüksek olmasıdır. Diğer bir deyişle, borç düzeyi sabitken bir öğrencinin temerrüt riski aslında daha düşüktür.
- Karıştırıcı değişken olan kredi borcu eğer modelde yer almaz ise öğrencilerin riskli olduğu gibi yanıltıcı bir sonuca ulaşılmaktadır.
- Bu da bir banka yönetimi ya da yapay zeka uygulaması için dikkat edilmesi gereken önemli bir farktır.
- Bu örnekte kredi kartı borcunun neden karıştırıcı değişken olduğu Şekil 4'te görülmektedir.



Karıştırıcı Değişken Etkisi (2)

- Sol panelde kesikli turuncu çizgiyle gösterilen öğrencilerin genel temerrüt oranı daha yüksektir. Ancak her bir borç düzeyinde öğrencilerin temerrüt oranı aslında daha düşüktür. Sağ panelde ise öğrencilerin daha borçlu olduğu anlaşılmaktadır.



Şekil 4: Temerrüt, kredi kartı borcu ve öğrencilik durumu ilişkisi



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Diskriminant Çözümlemesi (1)

- Sınıflandırma uygulamalarında lojistik regresyon dışında yaygın olarak kullanılan araçlardan biri de **diskriminant çözümlemesi** (discriminant analysis) yöntemidir.
- Lojistik regresyonda $\Pr(Y|X)$ koşullu olasılığını modellemiştik.
- Diğer bir deyişle, X verili iken Y tepkisinin dağılımını bulmuştuk.
- Diskriminant çözümlemesinde ise öncelikle her bir X 'in dağılımı modellenir. Daha sonra bu tersine çevirilerek $\Pr(Y = k|X)$ tahminine dönüştürülür.
- Bu işlem için ünlü **Bayes kanıtı** (Bayes' theorem) kullanılır:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$



Diskriminant Çözümlemesi (2)

- Diskriminant çözümlemesi özellikle temerrüt tahmininde yaygındır. Ayrıca boyut azaltma yöntemi olarak nesne ve yüz tanıma uygulamalarında da sıkça kullanılır.
- Lojistik model yerine bu yaklaşımı kullanmanın dört nedeni vardır:
 - 1 Sınıflar birbirinden iyi ayrılmışsa bu yöntem daha karardır.
 - 2 Eğitim veri seti küçük olduđu zaman daha iyi sonuç verir.
 - 3 Teoride en düşük test hata oranı Bayes sınıflandırıcıya aittir.
 - 4 İkiden fazla sınıf olduđuunda bu yöntem daha kullanışlıdır.
- Sınıf sayısı ikiden fazla ise lojistik regresyon yine de kullanılabilir, ancak sınıflandırma için çok kullanışlı değildir. **Çokterimli logbirim** (multinomial logit) denilen bu modeli burada tartışmayacağız.
- Ayrıca logbirime almasıık **probirim** (probit) adlı yaygın bir nitel tepki regresyon modeli daha vardır. Burada buna da değinmiyoruz.



Bayes Kanıtsavı ile Sınıflandırma

- Diskriminant çözümlemesini açıklamak için elimizdeki eğitim verilerini $K > 2$ adet sınıfa ayırmak istediğimizi düşünelim.
- Bunun için ilk olarak, rastsal bir gözlemin k no'lu sınıfa ait olma olasılığını π_k ile gösterelim. Buna **öncül** (prior) olasılık da denir.
- Ayrıca $f_k(x) \equiv \Pr(X = x|Y = k)$ şeklinde bir **olasılık yoğunluk fonksiyonu** (probability density function) tanımlayalım. Bu, $Y = k$ sınıfına ait bir gözlemin $X = x$ değerini alma olasılığını ölçer.
- Bu durumda Bayes kanıtsavı aşağıdaki gibi olur:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

- Kolaylık açısından $p_k(X) = \Pr(Y = k|X = x)$ diyelim.
- Yukarıda tanımladığımız $p_k(x)$, belli bir $X = x$ gözleminin k no'lu sınıfa ait olma **ardıl** (posterior) olasılık değeridir.
- 2. Bölümde Bayes sınıflandırıcının $p_k(x)$ değeri en yüksek sınıfı seçtiğini görmüştük. Öyleyse amacımız $p_k(x)$ 'i tahmin etmektir.



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



İkili Doğrusal Diskriminant Çözümlemesi

- Bayes kanıtsavını kullanarak sınıflandırmanın nasıl yapıldığını açıkladıktan sonra tahmin aşamasına geçebiliriz.
- En basit durum olarak ikili doğrusal modeli ele alalım.
- Burada tek bir X için π_k ve $f_k(x)$ fonksiyonlarını tahmin etmek istiyoruz. Bunları yukarıdaki formüle yerleştirerek $p_k(x)$ 'i bulabiliriz.
- Öncül olasılık olan π_k 'yi tahmin etmek basittir. Bunu, eğitim veri setinde k sınıfına ait gözlemlerin oranı olarak belirleriz.
- Ancak $f_k(x)$ olasılık yoğunluk fonksiyonu için bazı varsayımlar yapmak gereklidir. Bunun için $f_k(x)$ 'in normal dağıldığını varsayalım:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

- Formülde yer alan μ_k ve σ_k parametreleri k no'lu sınıf için normal dağılıma ait ortalama ve ölçünlü sapma değerleridir.
- Doğrusal diskriminant çözümlemesinde her bir sınıfa ait varyansın sabit olduğu ($\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$) varsayılır.



Diskriminant Fonksiyonu

- Yukarıda belirlediğimiz π_k ve $f_k(x)$ fonksiyonlarını $p_k(x)$ formülü içine yerleştirince, sadeleştirmelerden sonra aşağıdaki **diskriminant fonksiyonu** (discriminant function) elde edilir:

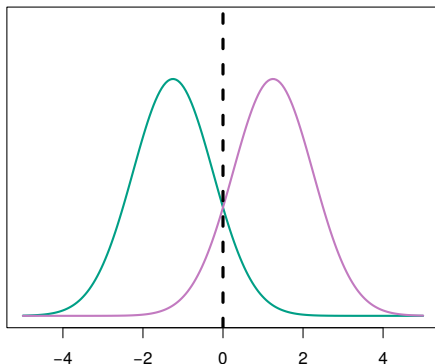
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

- Herhangi bir x değerini sınıflandırmak için yukarıdaki formül önce her sınıf için ayrı ayrı hesaplanır ve **diskriminant skoru** (discriminant score) en yüksek olan sınıf seçilir.
- Ayrıca formülü $\delta_k(x) = \delta_l(x)$, $k \neq l$ şeklinde farklı sınıflar için çözerek **Bayes karar sınırı** (Bayes decision boundary) da bulunabilir.
- Örnek olarak, iki sınıf olsun ve $\mu_1 = -1,25$ ve $\mu_2 = 1,25$ diyelim. Burada karar sınırı $(-1,25 + 1,25)/2 = 0$ olur. Böylece, $x < 0$ için sınıf 1 seçilirken $x > 0$ olduğunda ise sınıf 2 seçilir.
- Bu örnekteki $\delta_k(x)$ 'ler ve Bayes karar sınırı Şekil 5'teki gibidir.



Bayes Karar Sınırı

- Şekilde iki ayrı sınıfı belirleyen $f_1(x)$ ve $f_2(x)$ dağılımlarının ortalamaları sırasıyla $-1,25$ ve $1,25$ 'tir. Ortadaki çizgi Bayes karar sınırıdır. $x < 0$ olduğunda yeşil renkle gösterilen $f_1(x)$ daha yüksek olduğu için sınıf 1 seçilir. Dolayısıyla $x > 0$ için de sınıf 2 seçilir.



Şekil 5: Tek X ve iki sınıf için örnek Bayes karar sınırı



Diskriminant Fonksiyonunun Tahmin Edilmesi

- Uygulamada X 'in her sınıf için normal dağıldığı biliniyor olsa bile $\mu_1, \dots, \mu_K, \sigma^2$ ve $\pi_1, \dots, \pi_K$ değerlerini tahmin etmek gerekir.
- Doğrusal çözümlemede bu işlem için şu formülleri kullanırız:

$$\hat{\pi}_k = n_k/n$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}_k^2 = \frac{1}{n-K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

- Böylece, *tahmin edilmiş* diskriminant fonksiyonunu elde ederiz:

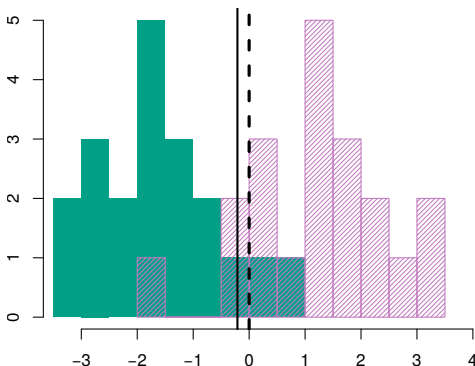
$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}_k^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}_k^2} + \log(\hat{\pi}_k)$$

- Baştaki örneğimize ait doğrusal diskriminant çözümlemesi tahmin sonuçları Şekil 6'da görülmektedir.



İkili Doğrusal Diskriminant Çözümlemesi Tahmini

- Şekilde siyah düz çizgiyle gösterilen doğrusal DÇ karar sınırı, kesikli çizgiyle gösterilen Bayes karar sınırına oldukça yakın tahmin edilmiştir. İki yöntemin hata oranları sırasıyla %11,1 ve %10,6'dır.



Şekil 6: Tek X ve iki sınıf için örnek doğrusal diskriminant çözümlemesi



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 **Diskriminant Çözümlemesi**
 - İkili doğrusal DÇ modeli
 - **Çoklu doğrusal DÇ modeli**
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Çoklu Doğrusal DÇ Modeli

- Doğrusal diskriminant çözümlemesi genellikle birden daha çok ($p > 1$) sayıda X değişkeni ile yapılır.
- Bu durumda normal dağılım yerine **çokdeğişkenli** (multinomial) normal dağılım kullanılır:

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- Yukarıda μ , k sınıfına ait ortalamadır. Σ ise farklı X 'lerin birbiriyle kovaryanslarını veren sınıflararası ortak bir covaryans matrisidir.
- Bayes denklemine bu ikisini yerleştirdiğimiz zaman başta tartıştığımız diskriminant fonksiyonunun matris sürümünü elde ederiz:

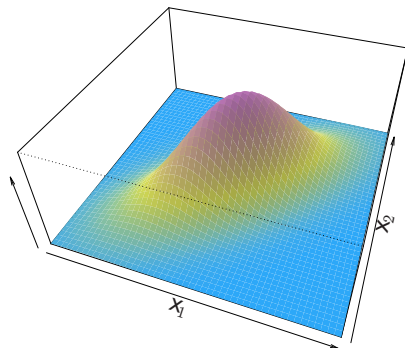
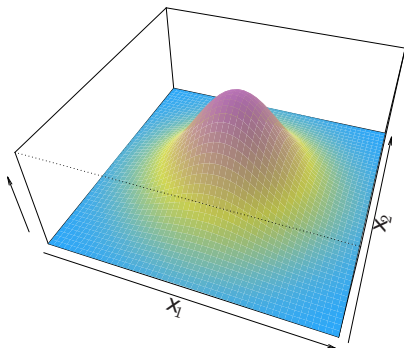
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

- Çokdeğişkenli normal dağılımın çizim üzerinde gösterimi Şekil 7'deki gibidir.



Çokdeğişkenli Normal Dağılımlar

- Sol panelde görülen çokdeğişkenli normal dağılımda X_1 ve X_2 değişkenleri bağımsızdır. Sağ paneldeki dağılımda ise iki değişken arasında 0,7 oranında korelasyon bulunmaktadır.



Şekil 7: Çokdeğişkenli normal dağılım örneği



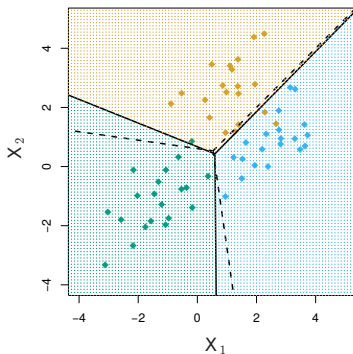
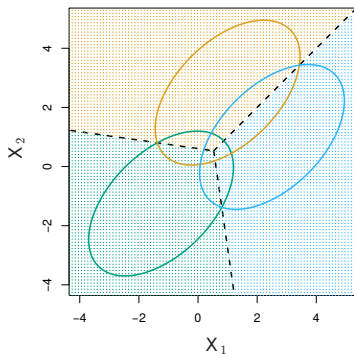
Çoklu Doğrusal Diskriminant Çözümlemesi Tahmini (1)

- Çoklu diskriminant çözümlemesi modelinin tahmin edilmesi de yine, ikili modeldeki gibidir.
- Eldeki eğitim veri seti kullanılarak $\mu_1, \dots, \mu_K$, $\pi_1, \dots, \pi_K$ ve Σ şeklindeki bilinmeyen parametreler tahmin edilir.
- Bunların formülleri önceki formüller ile benzerdir.
- Daha sonra, tahmin edilen diskriminant fonksiyonu kullanılarak sınıflandırma yapılır: Belli bir $X = x$ değeri için her bir sınıfa ait diskriminant skoru bulunur ve en yüksek skora sahip sınıf seçilir.
- Ayrıca karar sınırları da diskriminant fonksiyonları kullanılarak kolayca tahmin edilebilir.
- İki adet X değişkeni ve toplam üç adet sınıftan oluşan bir çoklu doğrusal DÇ tahmini Şekil 8'de gösterilmiştir.



Çoklu Doğrusal Diskriminant Çözümlemesi Tahmini (2)

- Sol paneldeki $p = 2$ olan üç sınıflı örnek için doğrusal DÇ tahminleri sağ panelde verilmiştir. Düz çizgiyle gösterilen tahminler, kesikli çizgiyle gösterilen Bayes karar sınırlarına oldukça yakındır.



Şekil 8: İki X ve üç sınıf için örnek doğrusal diskriminant çözümlemesi



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 **Diskriminant Çözümlemesi**
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - **Almaşık DÇ modelleri**
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



İkinci Dereceden Diskriminant Çözümlemesi (1)

- Yukarıda tartıştığımız doğrusal DÇ modellerinde sınıflar arasındaki tek fark ortalamalardan kaynaklanıyordu.
- Varyansın her sınıf için ortak olduğunu ya da çoklu durumda X 'lere ait tek bir kovaryans matrisi olduğunu varsaymıştık.
- **İkinci dereceden** (quadratic) DÇ modelinde sabit varyans varsayımını kaldırırız. Bu durumda diskriminant fonksiyonu şöyle olur:

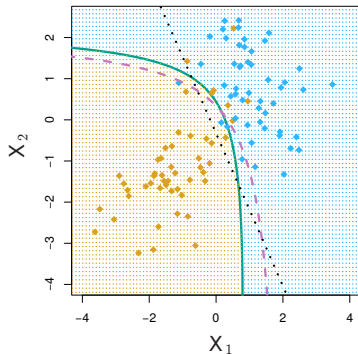
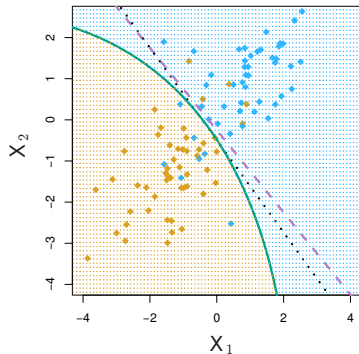
$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log \pi_k$$

- Yukarıdaki denklem x 'e göre ikinci dereceden bir fonksiyondur.
- Görüldüğü gibi, doğrusal ve ikinci dereceden DÇ yaklaşımları temelde aynıdır. Aradaki fark hesaplama yoğunluğundadır.
- İkinci dereceden modelde her sınıf için ayrı bir kovaryans matrisi tahmin edildiği için toplam $Kp(p+1)/2$ parametre hesaplanır. Doğrusal modelde ise Kp parametre bulunmaktadır.
- Dolayısıyla belli bir durumda hangi yöntemin seçileceği sınıflar arası ilişkiye ve varyans-yanlılık ödünleşmesine bağlı olur.



İkinci Dereceden Diskriminant Çözümlemesi (2)

- Sol panelde sınıfların X 'ler ile korelasyonu aynı, sağda farklıdır. Sol panelde siyah noktalı çizgiyle gösterilen doğrusal DÇ karar sınırı mor renkli ideal Bayes sınırına daha yakındır. Sağda ise yeşil renkli ikinci dereceden DÇ karar sınırı daha başarılıdır.



Şekil 9: Doğrusal ve ikinci dereceden DÇ tahminlerinin karşılaştırılması



Sade Bayes Diskriminant Çözümlemesi

- X değişkeni sayısı arttıkça ikinci dereceden DÇ tahmini zorlaşır.
- Örnek olarak, 50 adet X olduğu zaman her bir sınıf için $50 \times 51/2 = 1275$ parametre tahmin etmek gerekir.
- Bu zorluktan kaçınmak için her bir sınıftaki X 'lerin birbirinden bağımsız olduğu yönünde basitleştirici bir varsayım yapılabilir.
- Bu durumda, hesaplanacak covaryans matrisi köşegen matrise indirgendiği için yalnızca $2Kp$ parametre tahmin edilir.
- **Sade Bayes** (naive Bayes) denilen bu DÇ modelinde diskriminant fonksiyonu aşağıdaki gibi olur:

$$\delta_k(x) = -\frac{1}{2} \sum_{j=1}^p \left[\frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} \right] + \log \pi_k$$

- Sade Bayes uygulamada genellikle başarılı sonuçlar vermektedir.
- Sonuç olarak, diskriminant çözümlemesi oldukça esnek bir sınıflandırma yaklaşımı sunmaktadır. Farklı istatistiksel dağılımlar ve varsayımlar kullanarak daha başka modeller de oluşturulabilir.



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Sınıflandırma Kesinliğinin Ölçülmesi

- Diğer yöntemlerde olduğu gibi sınıflandırma çözümlemesinde de tahminlerin kesinliği önemlidir.
- Bu konuya yönelik olarak kredi kartı örneğimize geri dönelim.
- 10.000 gözlemlik bu veri setinde temerrüte düşme durumunu çoklu doğrusal DÇ modeli ile tahmin ettiğimiz zaman %2,75 gibi düşük bir eğitim hata oranı elde edilmektedir.
- Bu sonuca bakarak uyguladığımız sınıflandırmanın son derece başarılı olduğunu düşünebiliriz.
- Ancak dikkat edilmesi gereken iki nokta vardır:
 - 1 Tahmin için eldeki veriler kullanılarak *eğitim* hata oranı minimize edilmiştir. Dolayısıyla yeni veriler kullanıldığı zaman bulunacak *test* hata oranının daha yüksek çıkması beklenir.
 - 2 Bu veri setindeki *öncül* temerrüt oranı %3,33'tür. Demek ki herkes için “temerrüt yoktur” dersek, bu **boş** (null) sınıflandırıcının hata oranı %2,75'ten yalnızca biraz daha yüksektir.



Karışıklık Matrisi

- Genel olarak, kredi kartı örneğindeki gibi bir sınıflandırma probleminde iki farklı şekilde hata yapabiliriz:
 - 1 Temerrütteki birisini *temerrüte düşmeyenler* sınıfına atamak
 - 2 Temerrüte düşmeyen birisini *temerrüttekiler* sınıfına atamak
- Bu iki tür hatayı incelemenin uygun bir yolu **karışıklık matrisi** (confusion matrix) denilen aşağıdaki çizelgeyi incelemektir:

		Kestirilen Durum		
		Evet	Hayır	Toplam
Gerçek Durum	Evet	81	252	333
	Hayır	23	9644	9667
Toplam		104	9896	10000

- Çizelgeye göre temerrüte düşmeyen toplam 9667 kişi içinde 9644'ü doğru tahmin edilmiştir. Burada hata oranı %0,2'dir.
- Ancak temerrüte düşen 333 kişinin yalnızca 81'i doğru sınıflandırılmış, 252 kişide hata yapılmıştır. Buradaki hata oranı %75,7'dir.
- Bir banka için asıl önemli olan ikinci hata oranı çok yüksektir.



Özgüllük ve Duyarlılık (1)

- Birçok uygulamada sınıflardan biriyle daha fazla ilgileniriz. Bu, genellikle düşük olasılıklı ancak yüksek riskli bir durumu gösterir.
- Kredi kartı örneğinde ilgilendiğimiz sınıf temerrüte düşecek kişilerdir. Başka bir örnekte bir şirketin ele geçirilmeye hedef olması ya da bir ürünün iade edilmesi daha önemli olabilir.
- Bu durumda bir sınıflandırıcının başarı performansını ölçmek için iki farklı ölçütü birlikte incelemek gerekir:
 - 1 **Duyarlılık** (sensitivity)
 - 2 **Özgüllük** (specificity)
- Duyarlılık, asıl ilgilendiğimiz sınıftaki başarı oranıdır. Buna **pozitif sınıfı** da denir. Özgüllük ise **negatif sınıfı** denilen diğer sınıftaki başarı oranını ölçer.
- Kredi örneğinde özgüllük %99,8 olsa da duyarlılık %24,3'tür.
- Peki, bu model duyarlılık konusunda nasıl başarısız olmuştur?
- Bunun nedeni uygulanan yöntemin sınıflar arasında ayırım yapmadan hataları *toplam olarak* minimize etmeye çalışmasıdır.



Özgüllük ve Duyarlılık (2)

- Bayes sınıflandırıcı, sınıfları eşit gördüğü için öntanımlı olarak daha yüksek olasılığa sahip olan sınıfı seçer:

$$\Pr(\text{Temerrüt} = \text{Evet} | X = x) > 0,5$$

- Öyleyse yukarıdaki eşik değerini değiştirerek bir bankanın ihtiyacına daha uygun bir sınıflandırma geliştirebiliriz.
- Örnek olarak, riskli kişileri belirlemek için %50 yerine %20 eşiği kullanılırsa aşağıdaki karışıklık matrisi elde edilir:

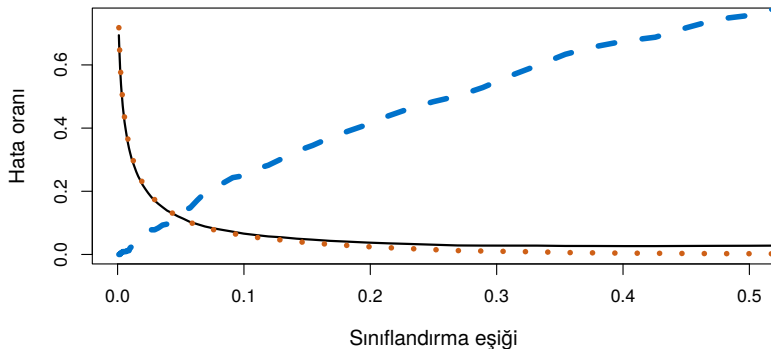
		<i>Kestirilen Durum</i>		
		Evet	Hayır	Toplam
<i>Gerçek Durum</i>	Evet	195	138	333
	Hayır	235	9432	9667
Toplam		430	9570	10000

- Yeni durumda pozitif sınıfı (temerrüt=evet) için hata oranı %41,4'e düşmüştür ($138/333 \times 100$). Ancak negatif sınıfı (temerrüt=hayır) için hata oranı da %2,43'e yükselmiştir ($235/9667 \times 100$).



Özgüllük ve Duyarlılık (3)

- İki hata oranı arasındaki ödünleşme Şekil 10'da görülmektedir.
- Mavi çizgi pozitif sınıfa ait hata oranını gösteren **yanlış pozitif** (false positive) oranıdır. Kırmızı çizgi de diğer sınıfa ait **yanlış negatif** (false negative) oranıdır. Siyah çizgi ise genel hata oranıdır.



Şekil 10: Yanlış pozitif ve yanlış negatif oranları arasındaki ödünleşme



Özgüllük ve Duyarlılık (4)

- Sınıflandırma çözümlemesi yaparken ortaya çıkabilecek sonuçlar ve bunların nasıl adlandırıldığı Çizelge 4'te gösterilmiştir.

		Kestirilen Durum		
Gerçek Durum		Sınıf 1 (+)	Sınıf 2 (-)	Toplam
	Sınıf 1 (+)	Doğru Pozitif (DP)	Yanlış Negatif (YN)	TP
	Sınıf 2 (-)	Yanlış Pozitif (YP)	Doğru Negatif (DN)	TN

Çizelge 4: Karışıklık matrisinin yorumlanması

- Doğru pozitif oranı, DP/TP olarak bulunur. Yanlış pozitif oranı ise YP/TN olur. Negatif sınıfa ait oranlar da benzer şekildedir.
- Klinik tıp literatüründe yaygın kullanılan bu tanımları daha iyi anlamak için pozitif sınıfı “hastalık var” olarak düşünebilirsiniz.
- Farklı kaynaklarda “gerçek durum” ve “kestirilen durum” eksenleri ya da sınıfların sırası yer değiştirebilir.



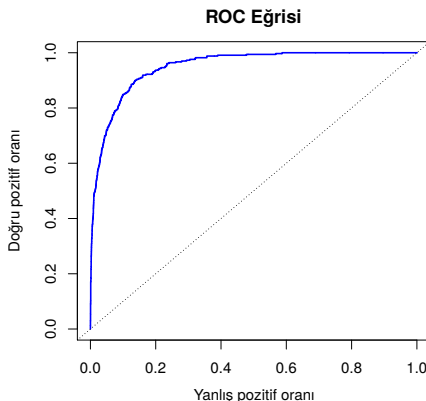
ROC Eğrisi

- Yukarıda görüldüğü gibi sınıflar eğer dengesiz ise eşik değerini değiştirerek daha iyi bir sınıflandırma elde edilebilir.
- En iyi eşik noktasını bulmak için ise **ROC eğrisi** (receiver operating characteristics curve ya da kısaca ROC curve) denilen şekilsel araçtan yararlanabiliriz.
- ROC eğrisi, pozitif sınıfa ait doğru sınıflandırma oranı ile yanlış sınıflandırma oranı arasındaki ilişkiyi farklı eşik düzeyleri için verir.
- İyi bir sınıflandırıcının doğru pozitif oranı yüksek, yanlış pozitif oranı düşüktür.
- Hiçbir X değişkenini dikkate almayan **bilgisiz sınıflandırıcı** (no information classifier) için ise bu iki oran birbirine eşit olur.
- ROC eğrisi altında kalan alana **AUC skoru** (area under the curve score) adı verilir. Bu değer 1'e ne kadar yakınsa sınıflandırıcı da o kadar iyidir.
- Kredi kartı örneğimize ait ROC eğrisi Şekil 11'de gösterilmiştir.



Örnek ROC Eğrisi

- İyi bir sınıflandırma için ROC eğrisinin sol üst köşeye yakın olması istenir. Eğri altındaki alan ise sınıflandırıcının iyiliğini ölçer. 45 derece çizgisi bilgisiz sınıflandırıcıyı temsil etmektedir.



Şekil 11: Temerrüt olasılığı modeline ait ROC eğrisi



Ders Planı

- 1 Doğrusal Olasılık Modeli
 - DOM tahminine yönelik eleştiriler
- 2 Lojistik Regresyon
 - İkili lojistik model
 - Lojistik tahmin ve kestirim
 - Çoklu lojistik model
- 3 Diskriminant Çözümlemesi
 - İkili doğrusal DÇ modeli
 - Çoklu doğrusal DÇ modeli
 - Almaşık DÇ modelleri
- 4 Diğer Konular
 - Sınıflandırma kesinliğinin ölçülmesi
 - Sınıflandırıcıların karşılaştırılması



Sınıflandırıcıların Karşılaştırılması (1)

- Bu bölümde aşağıdaki sınıflandırma yöntemlerini ele aldık:
 - 1 Doğrusal olasılık modeli
 - 2 Lojistik regresyon
 - 3 Doğrusal diskriminant çözümlemesi
 - 4 2. dereceden diskriminant çözümlemesi
 - 5 Sade Bayes sınıflandırıcı
- Ayrıca 2. Bölümde K-enyakın komşu yöntemini de görmüştük.
- Şimdi, son olarak, bu altı yaklaşımı birbiriyle karşılaştıralım ve farklı durumlarda hangisinin daha iyi sonuçlar verebileceğini kısaca tartışalım.



Sınıflandırıcıların Karşılaştırılması (2)

- Doğrusal olasılık modeli, lojistik regresyon ve doğrusal diskriminant çözümlemesi yöntemlerinin üçü de doğrusal karar sınırları üretirler. Aralarındaki temel bir fark kullanılan tahmin yöntemidir.
- DOM, enküçük karelere dayanmaktadır. Lojistik regresyon $\Pr(Y|X)$ şeklindeki koşullu ençok olabilirlik yöntemini, doğrusal DÇ ise tam bilgi ençok olabilirlik tahmincisini kullanır.
- Sınıf sayısı iki olduğu zaman bu üç yöntemin benzer sonuçlar vermesi beklenir. Bu noktada lojistik regresyon kuramsal olarak DOM'a tercih edilebilir. DOM'un ise hesaplama maliyeti düşüktür.
- Sınıf sayısı ikiden fazla ise DOM kullanılamaz. Bu durumda lojistik regresyon yerine DÇ yöntemi daha kullanışlıdır.
- Ancak doğrusal DÇ yöntemi sınıfların normal dağılımlı ve ortak varyanslı olduğunu varsayar. Dolayısıyla bu yöntemin başarısı normallik varsayımının ne kadar doğru olduğuna da bağlıdır.



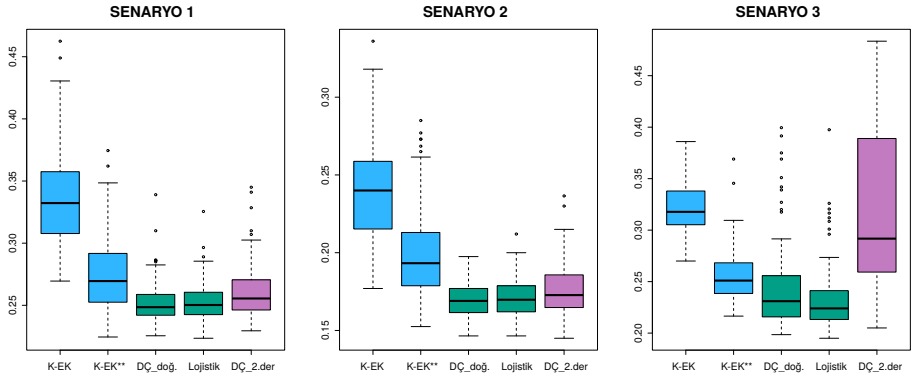
Sınıflandırıcıların Karşılaştırılması (3)

- İkinci derece diskriminant çözümlemesi doğrusal-dışı karar sınırları üreterek daha esnek bir yaklaşım sunar.
- Ancak bu yöntem diğerlerine göre hesaplama konusunda daha maliyetlidir ve çok daha büyük veri setlerine ihtiyaç duyar.
- Özellikle X değişkeni sayısı çok ise ikinci derece DÇ'yi hesaplamak olanaksızlaşabilir. Bu durumda uygulamada genellikle başarırlı sonuçlar verdiği görülen sade Bayes yöntemi tercih edilebilir.
- Son olarak, K-enyakın komşu yöntemi diğerlerinden tümüyle farklı, parametrik-dışı bir sınıflandırma yaklaşımı sunmaktadır.
- Karar sınırının yüksek derecede doğrusal-dışılık gösterdiği durumlarda K-EK'in diğer yöntemlere üstün gelmesi beklenir. Ancak burada da esneklik seçimi önemlidir.
- Gerçek sınıflandırmanın doğrusal ya da doğrusal-dışı olduğu altı farklı senaryo Şekil 12 ve Şekil 13'te incelenmiştir.



Doğrusal Durumda Sınıflandırıcılar

- Şekilde gerçek sınıf ayrımının doğrusal olduğu üç farklı senaryo verilmiştir. Genel olarak, doğrusal DÇ modeli ile lojistik regresyon modelinin hata oranlarının daha düşük olduğu görülmektedir.

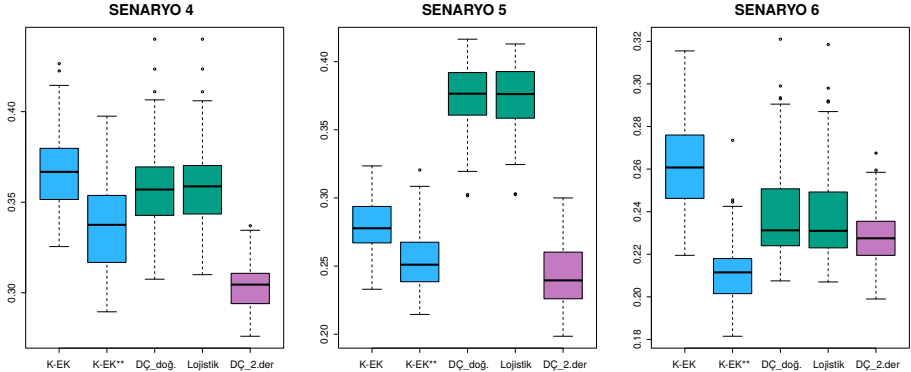


Şekil 12: Gerçek ayrım *doğrusal* iken sınıflandırıcıların hata oranı



Doğrusal-Dışı Durumda Sınıflandırıcılar

- Şekilde gerçek sınıf ayrımının doğrusal-dışı olduğu üç farklı durum verilmiştir. Burada 2. dereceden DÇ modeli ile parametrik dışı K-EK yönteminin daha düşük hata oranları verdiği görülmektedir.



Şekil 13: Gerçek ayrım *doğrusal-dışı* iken sınıflandırıcıların hata oranı



Sonuç

- Yukarıdaki şekillerde sırasıyla $K=1$ olan K-EK, K değerinin çapraz geçерleme yöntemiyle optimize edildiği K-EK, doğrusal DÇ, lojistik regresyon ve ikinci derece DÇ tahminleri karşılaştırılmıştır.
- Açıkça görüldüğü gibi her durumda diğerlerine üstünlük sağlayan tek bir yöntem bulunmamaktadır.
- Ayrıca yukarıdakilerden başka sınıflandırma yöntemleri de vardır. Ancak bunlar da yine duruma göre diğerlerinden iyi ya da daha kötü sonuçlar vereceklerdir.
- Sonuç olarak, diğerlerinde olduğu gibi sınıflandırma çözümlemesinde de başarılı kestirimler için yöntem seçimi önemlidir.
- Farklı durumlarda en iyi yöntemi ya da en uygun esneklik düzeyini seçmek için kullanabileceğimiz teknikleri bir sonraki bölümde ele alacağız.



Önümüzdeki Dersin Konusu ve Ödev

Ödev

Kitaptan **Bölüm 4** “Sınıflandırma” okunacak.

Önümüzdeki Ders

Yeniden Örnekleme

