

Sınıflandırma Problemleri

(İktisatçılar İçin) Makine Öğrenmesi (TEK-ES-2020)

Hüseyin Taştan
Yıldız Teknik Üniversitesi

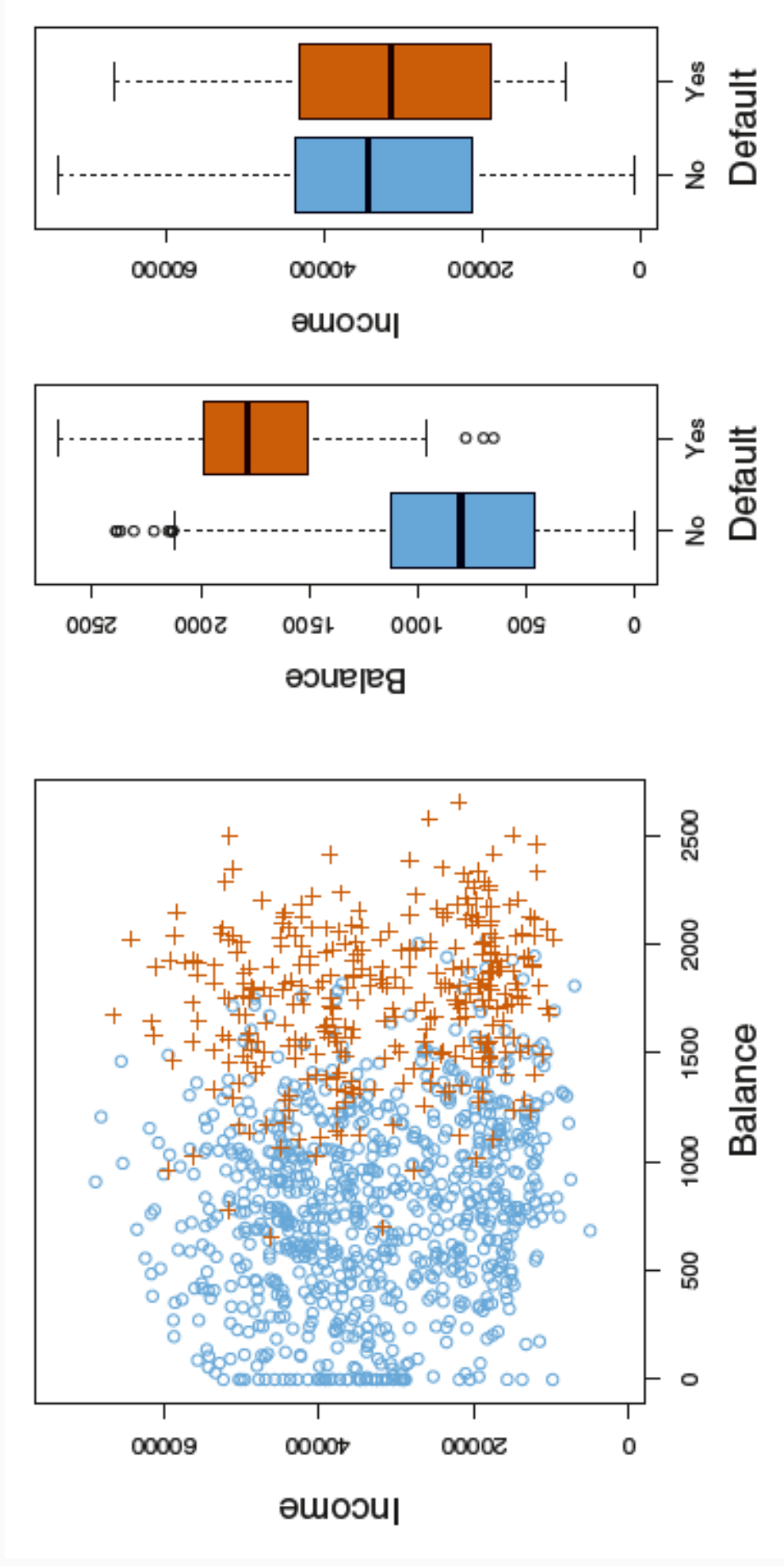
Plan

- Sınıflandırma problemleri
- Lojistik regresyon
- Doğrusal Diskriminant Analizi (LDA)
- Kareseel Diskriminant Analizi (QDA)
- Sınıflandırma performansının ölçümü
- ROC eğrisi ve AUC

Sınıflandırma Problemi

- Regresyon analizinde Y tepki değişkeni niceldir.
- Y kategorik bir değişken = sınıflandırma problemi
- Sınıflandırıcı: verilmiş bir X değişken seti için Y 'nin kategorisini kestiren yöntem.
- İkili ya da çoklu olabilir.
- Örnek: bir banka kredi başvuru sahibinin özelliklerinden hareketle kişinin geri ödeyememe olasılığını hesaplamak isteyebilir. Kişinin sınıflandırıldığı gruba göre başvurusu red ya da kabul edilir.
- Veri seti: default (Yes/No), X değişkenleri: kredi kartı bakiyesi (balance), gelir (income), öğrenci kuklası (student)

Sınıflandırma: Örnek



Mavi: Default=NO, Kavuniçi: Default=YES (ISLR Fig-4.1, p.129)

Regresyon kullanabilir miyiz?

- Standart regresyon analizi ile gözlemleri sınıflandırabilir miyiz? Burada iki sınıf olduğunu varsayıyoruz. Örneğin

$$default = \beta_0 + \beta_1 income + \beta_2 balance + \epsilon$$

Burada default ikili (0/1) değerler almaktadır.

- Sınıflandırma kuralı: OLS tahmin değeri 0.5'den büyükse default=YES (1) grubuna değilse default=NO (0) grubuna sınıflandırma yapabiliriz.
- Ancak kestirim değerlerinin 0 ile 1 arasında olmasının garantisi yoktur. Herhangi bir değeri alabilirler hatta negatif olabilirler.
- Ayrıca doğrusal olasılık modelinde hata terimi sabit varyanslı değildir.
- Tepki değişkeni ikiden fazla gruba sähipse uygulanamaz.

Lojistik Regresyon

- Doğrudan tepki değişkenini modellemek yerine sınıflandırma olasılığını modelleyebiliriz.

$$p(X) = \beta_0 + \beta_1 X$$

Burada $p(X) = Pr(Y = 1 | X)$ tepki değişkeninin grup=1 olarak sınıflandırılma (koşullu) olasılığıdır.

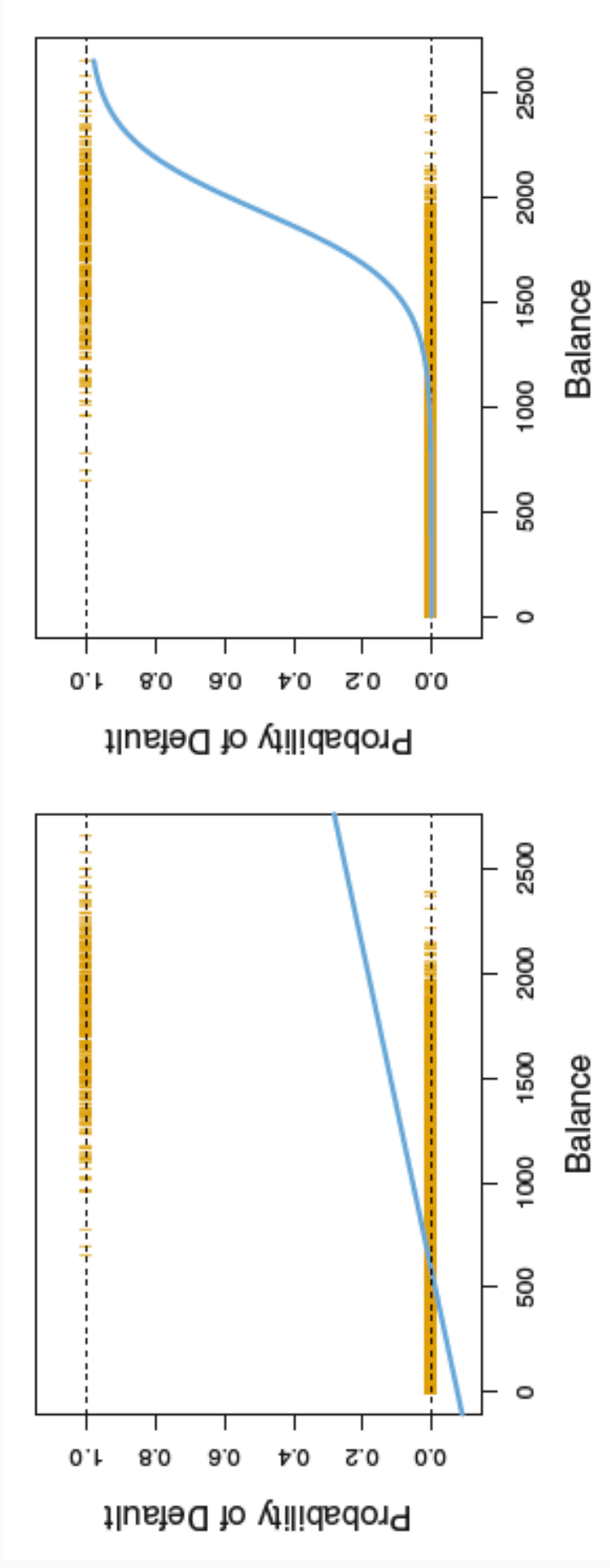
- Tanım gereği $0 \leq p(X) \leq 1$. Örneğin, lojistik fonksiyon

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

- Logit modeli

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

Lojistik vs. Doğrusal Olasılık Modeli



Sol: doğrusal regresyon ile tahmin edilen koşullu olasılıklar, **Sağ:** Lojistik regresyon ile tahmin edilen koşullu olasılıklar (ISLR Fig-4.2, p.131)

Lojistik Regresyonun Tahmini

- Modelin tahmininde OLS kullanılamaz.
- En Yüksek Olabilirlik (Maximum Likelihood) yöntemi tutarlı ve etkin tahminciler verir.
- Katsayılar hesaplandıktan sonra koşullu olasılıklar hesaplanabilir.
- Sınıflandırma işlemi koşullu olasılık tahminlerine göre yapılabilir.

Çoklu Lojistik Regresyon

- Çok sayıda nicel ya da nitel kestirim değişkeni için model kolayca geliştirilebilir:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

- Modelin doğrusal olmayan yapısından dolayı katsayılar koşullu olasılıklardaki değişim olarak yorumlanamaz. Ancak işaretleri yorumlanabilir.
- $\hat{\beta}_j$ 'lar bulunduğundan sonra X_j değerleri birlikte yukarıdaki denklemden yerlerine yazılarak koşullu olasılıklar tahmin edilir.

Doğrusal Diskriminant Analizi

- Doğrusal diskriminant analizinde (LDA - Linear Discriminant Analysis) her grup için ayrı ayrı olmak üzere X değişkenlerinin dağılımı modellenir.
- Daha sonra Bayes Teoremi'nden hareketle asıl ilgilendiğimiz $Pr(Y = k | X = x)$ olasılıkları tahmin edilir.
- Sınıfların birbirinden çok ayrık olduğu durumlarda lojistik regresyon istikrarsız olabilir. LDA'nde böyle bir problem yoktur.
- n küçük olsa da değişkenler yaklaşık olarak normal dağılıyorsa LDA daha istikrarlı.
- LDA ikiden daha fazla grup olduğunda da uygulanabilir.

Sınıflandırmada Bayes Teoreminin Kullanımı

- Gözlemleri $K \geq 2$ sınıfa ayırmak istediğimizi düşünelim. Çıktı değişkeni $Y_i = k$, $k = 1, 2, \dots, K$ değerlerini almaktadır (sıralama önemsiz)
- Önsel olasılık (prior): π_k , rassal çekilmiş bir gözlemin k sınıfına ait olma olasılığı
- k sınıfı için X değişkeninin yoğunluk fonksiyonu: $f_k(x) \equiv \Pr(X = x | Y = k)$ (basitlik için X 'in kesikli bir değişken olduğunu varsaydık)
- $f_k(x)$ 'in yorumu: X 'in k sınıfından çekilme olasılığı yükseldikçe daha büyük değerler alır.
- Şimdi Bayes teoremini uygulayabiliriz:

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

- $p_k(X)$ 'i doğrudan tahmin etmek yerine bileşenleri tahmin edebiliriz.

Sınıflandırmada Bayes Teoreminin Kullanımı

Ardıl (posterior) olasılık dağılımı:

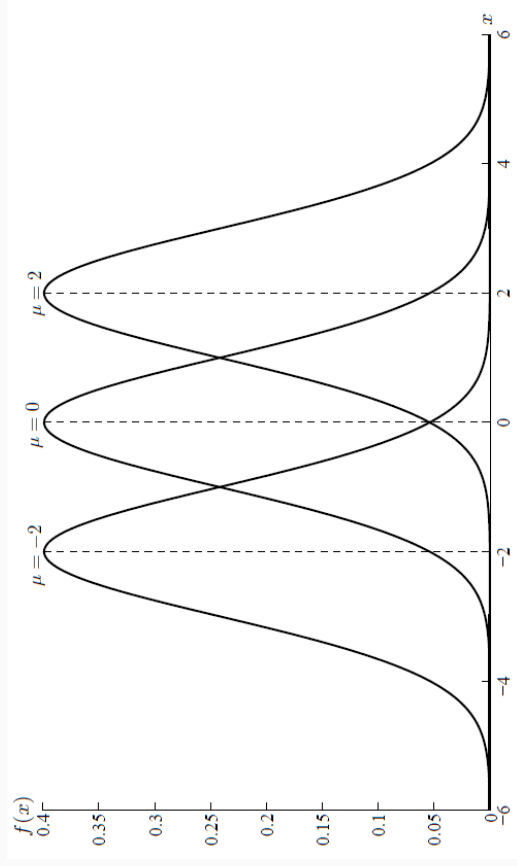
$$p_k(X) \equiv \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

- π_k verilerden kolayca tahmin edilebilir. k grubunun örneklemdaki oranını hesaplamak yeterli.
- Ancak $f_k(x)$ 'in tahmini daha zor. Dağılımsal varsayımlar yapmamız gerekir.
- LDA: normal dağılım. Örneğin, $p = 1$ için

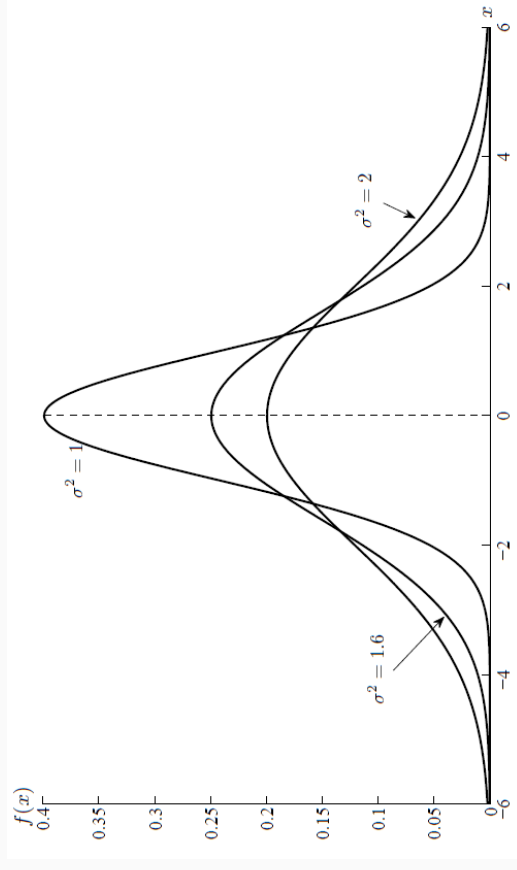
$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

Normal Dağılım

Farklı ortalamalar, varyans aynı



Farklı varyanslar, ortalama aynı



LDA, p=1 için koşullu olasılıklar

- Tek kestirim değişkeni $p = 1$ için normal dağılım ve her grup için aynı varyans (σ^2) varsayımları altında koşullu olasılıklar

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- Bayes sınıflandırıcısı verilmiş bir $X = x$ gözlemini $p_k(x)$ 'in en yüksek olduğu gruba atar.
- $p_k(x)$ 'in doğal logaritmasını alırsak **diskriminant** fonksiyonunu elde ederiz:

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

kategori kestirimlerini en yüksek $\delta_k(x)$ değerine göre yapabiliriz.

LDA katsayılarının tahmini

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

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

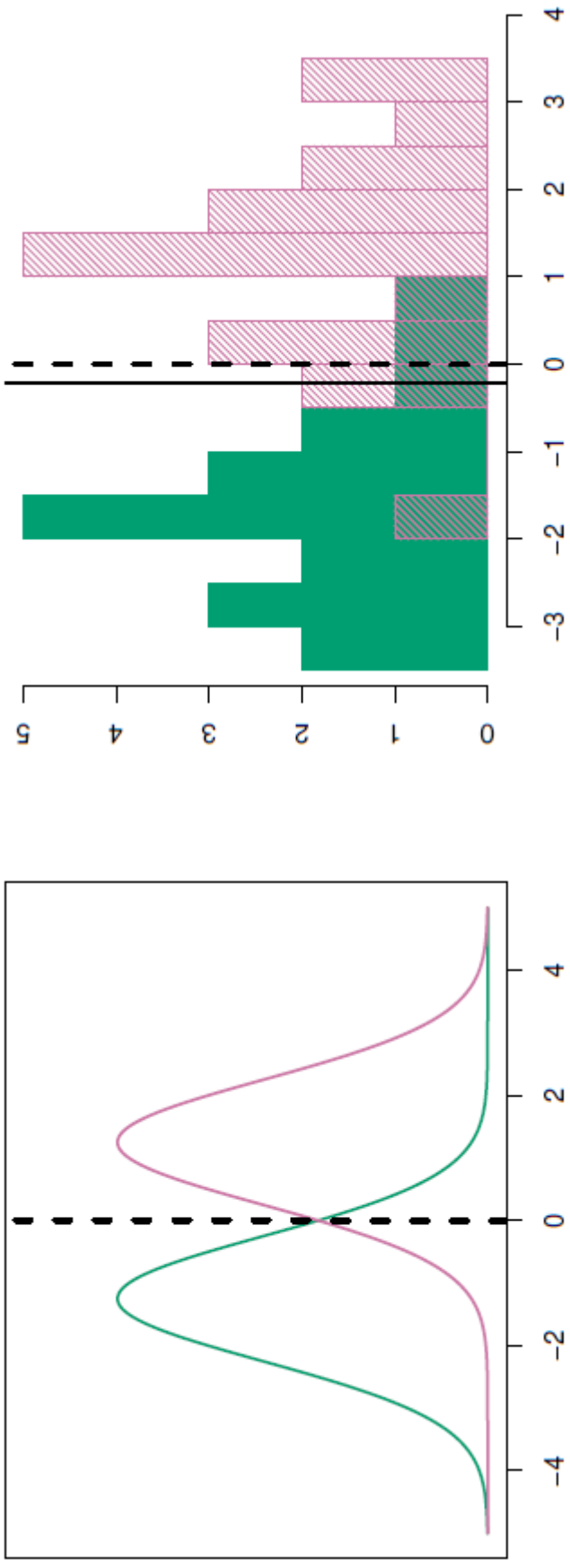
$$\hat{\pi}_k = n_k/n$$

Bu tahminleri diskriminant fonksiyonu içine yazarsak:

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

LDA: Örnek

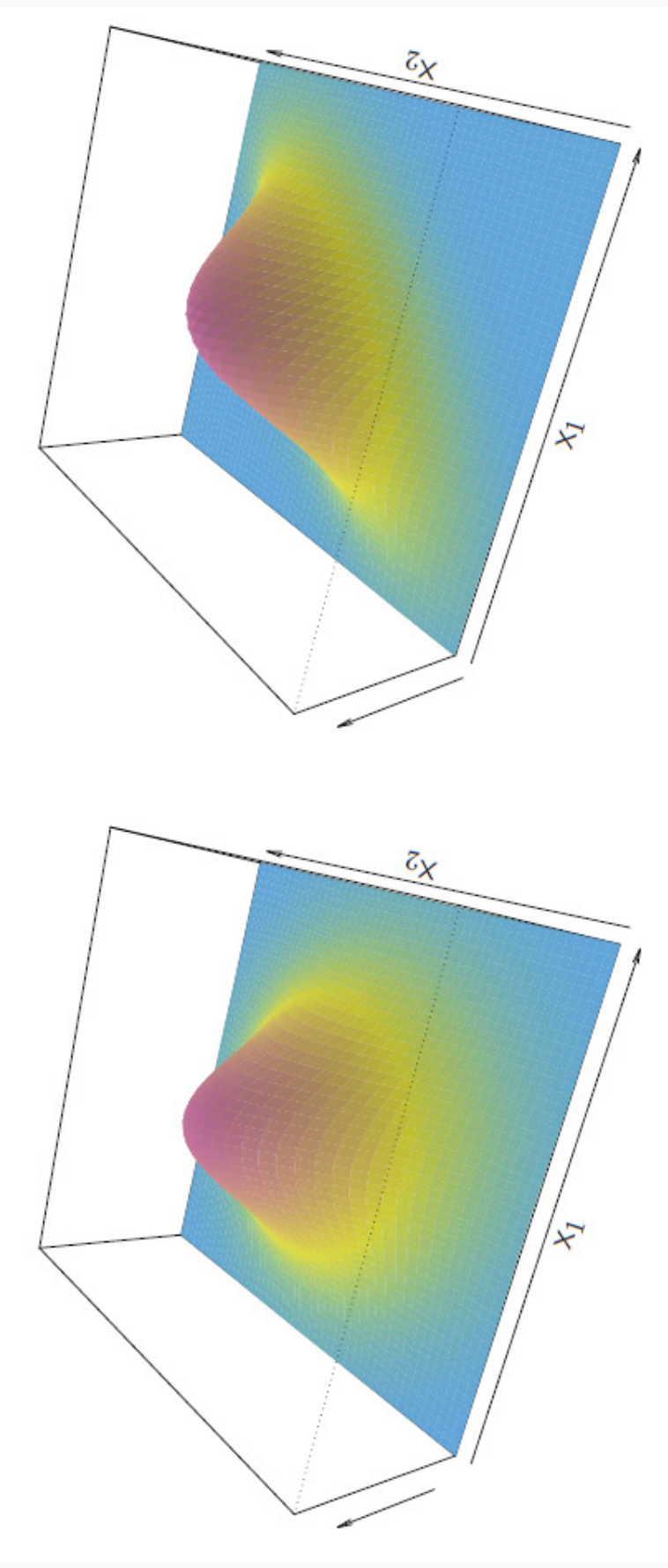
İki grup için farklı normal dağılım yoğunluk fonksiyonları (sol) ve verilerden hareketle oluşturulan histogramlar:



Not: (Sol) dikey kesikli çizgi Bayes karar sınırındır (simülasyonla oluşturulduğu için biliniyor), (sağ) Dikey çizgi eğitim verisiyle tahmin edilen LDA karar sınırındır. (ISLR, Fig. 4.4, p.140)

Çok değişkenli LDA

- Kestirim değişkenleri: $\mathbf{X} = (X_1, X_2, \dots, X_p)$
- Dağılımsal varsayım (Çok değişkenli Normal - Gaussian - dağılım): $\mathbf{X} \sim N(\mu, \Sigma)$



Sol: X_1 ve X_2 ilişkisiz, Sağ: ilişkili, kovaryans = 0.7

Çok değişkenli LDA

Çoklu Gaussian yoğunluk fonksiyonu.

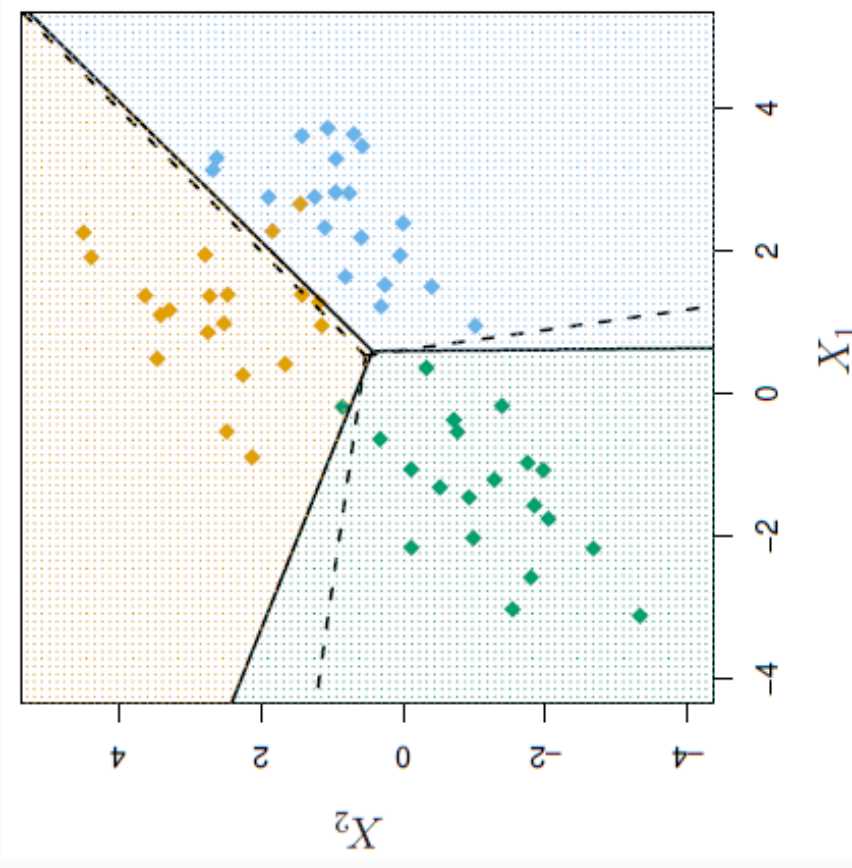
$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

$p = 1$ durumundakine benzer adımları takip ederek sınıflandırmada kullanacağımız diskriminant fonksiyonunu elde ederiz:

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

Verilmiş bir x gözlemi $\delta_k(x)$ 'in en büyük olduğu gruba atanır.

Örnek



- 2 kestirim değişkeni X_1, X_2
- Grup sayısı $K = 3$
- Her gruptaki gözlem sayısı 20
- Kesikli çizgiler LDA sınıflandırma sınırları
- Düz çizgiler Bayes sınırları
- LDA hata oranı = 0.0770
- Bayes hata oranı = 0.0746

Örnek: Borç ödememe (default) verileri

| | default | student | balance | income |
|----|---------|---------|------------|-----------|
| 1 | No | No | 729.52650 | 44361.625 |
| 2 | No | Yes | 817.18041 | 12106.135 |
| 3 | No | No | 1073.54916 | 31767.139 |
| 4 | No | No | 529.25060 | 35704.494 |
| 5 | No | No | 785.65588 | 38463.496 |
| 6 | No | Yes | 919.58853 | 7491.559 |
| 7 | No | No | 825.51333 | 24905.227 |
| 8 | No | Yes | 808.66750 | 17600.451 |
| 9 | No | No | 1161.05785 | 37468.529 |
| 10 | No | No | 0.00000 | 29275.268 |
| 11 | No | Yes | 0.00000 | 21871.073 |

- n=10000
- Balance = kredi kartı bakiyesi
- income = gelir düzeyi
- Student = öğrenci kuklası
- default = borç ödeyememe ikili değişken (Yes=borçlu, No= borçlu değil)

Örnek

- Kredi kartı bakiyesi ve öğrenci kuklası değişkenleri ile LDA tahmin ettiğimizde eğitim seti hata oranı %2.75. Bu oldukça düşük bir hata oranı gibi görünüyor.
- Verilerde borcunu ödeyemeyenlerin oranı (default=YES) %3.33. Bir gözlemi kredi kartı bakiyesinden ve öğrenci olup olmadığından bağımsız olarak "borçlu değil" (default=NO) diye sınıflandırsak (null classifier) yapacağımız hata oranı %3.33 olur.
- Bu açıdan baktığımızda LDA hata oranı aslında çok da başarılı değil.

Sınıflandırma Performansının Ölçümü

Hata Matrisi (Confusion Matrix)

| | GERÇEK GERÇEK | | |
|----------|---------------|-------|--|
| | - | + | • Tablo hücreleri gözlem sayılarıdır |
| TAHMİN - | A | B | Toplam A = Doğru tahmin edilen gerçek negatif sayısı |
| TAHMİN + | C | D | • D = Doğru tahmin edilen gerçek pozitif sayısı |
| Toplam | A + C | B + D | A + B. C = Yanlış pozitif sayısı |
| | | D | + C + B = Yanlış negatif sayısı |

- Yanlış pozitif oranı (false positive rate): $FP = C / (A + C)$
- Doğru pozitif oranı. $TP = D / (B + D)$ (duyarlılık - sensitivity)
- Doğru negatif oranı. $A / (A + C)$ (özgüllük - specificity)
- Pozitif kestirimsel değer: $PP = D / (C + D)$, Negatif kestirimsel değer: $NP = A / (A + B)$

Sınıflandırma Performansının Ölçümü

Hata Matrisi (Confusion Matrix)

| | | <i>True default status</i> | | Total |
|---------------------------------|-----|----------------------------|-----|--------|
| | | No | Yes | |
| <i>Predicted default status</i> | No | 9,644 | 252 | 9,896 |
| | Yes | 23 | 81 | 104 |
| Total | | 9,667 | 333 | 10,000 |

- Default = YES (Pozitif grup - borcunu ödemeyenler)
- Default = NO (Negatif grup - borcunu zamanında ödeyenler)

- Bu örnekte Doğru tahmin edilen gözlem sayısı $9644+81=9725$ 'dir. Toplamdaki payı ise %97.25'dir.
- Yanlış sınıflanan gözlem sayısı ise $252+23=275$. Yanlış sınıflama oranı ya da hata oranı % 2.75'dir.
- Gerçekte borçlu olmadığı halde yanlışlıkla borçlu olarak sınıflandırılanların oranı $23/9667 = 0.00238$, yani %0.238 gibi çok küçük bir orandır.
- Ancak Gerçekte borçlu olanlar içinde %75.7'si yanlışlıkla borçlu değil grubundadır (252/333)

Sınıflandırma Performansının Ölçümü

Hata Matrisi (Confusion Matrix)

| | | <i>True default status</i> | | |
|-------------------------------------|-------|----------------------------|-----|--------|
| | | No | Yes | Total |
| <i>Predicted default status</i> | No | 9,644 | 252 | 9,896 |
| | Yes | 23 | 81 | 104 |
| | Total | 9,667 | 333 | 10,000 |

- Default = YES (Pozitif grup - borcunu ödemeyenler)
- Default = NO (Negatif grup - borcunu zamanında ödeyenler)

- Özgüllük (specificity): $(1 - 23/9667) \times 100 = \% 99.8$
- Genel hata oranı düşük olsa da Default (YES) grubu içindeki hata oranı çok yüksek.
- Doğru pozitif oranı ya da duyarlılık (sensitivity) $\% 24.3 (= 100 \times 81/333)$
- Yanlış negatif oranı $= \%75.7 = 100 \times 252/333 = 1 - \text{duyarlılık}$. Yüksek risk grubundaki bireyleri tahmin etmek isteyen bir banka için bu oldukça yüksek.
- Yukarıdaki sınıflamada 0.5 eşik değerini kullandık. Yani koşullu olasılık 0.5'den büyükse default = YES grubuna atandı.

Farklı Eşik Değeri

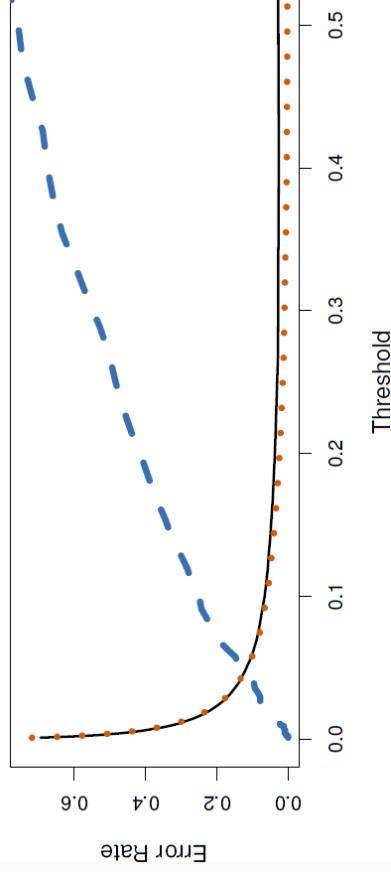
$$Pr(default = YES|X) > 0.2$$

| <i>Predicted default status</i> | <i>True default status</i> | | |
|-------------------------------------|----------------------------|-----|--------|
| | No | Yes | Total |
| | No | 138 | 9,570 |
| Yes | 235 | 195 | 430 |
| Total | 9,667 | 333 | 10,000 |

- Duyarlılık = $100 \times 195 / 333 = \%58.6$
- Yanlış pozitif oranı ise $\%41.44 = 100 \times 138 / 333 (=1\text{-duyarlılık})$
- Duyarlılık beklendiği gibi yükseldi ancak Borçlu olmadıkları halde yanlışlıkla borçlu olarak sınıflandırılanların oranı da yükseldi, $\%2.43 = 100 \times 235 / 9667$.
- Genel hata oranı: $\%3.73$

- Borcunu ödemeyecek müşterileri öngörmek isteyen bir banka 0.5 yerine yukarıdaki gibi daha düşük bir eşik değeri kullanabilir.
- Eşik değerini artırınca duyarlılık yükseldi (yanlış negatif oranı düştü) ancak genel hata oranı arttı. Bu ikisi arasındaki ödünüm izleyen grafikte verilmiştir.

Eşik değeri ve hata oranı

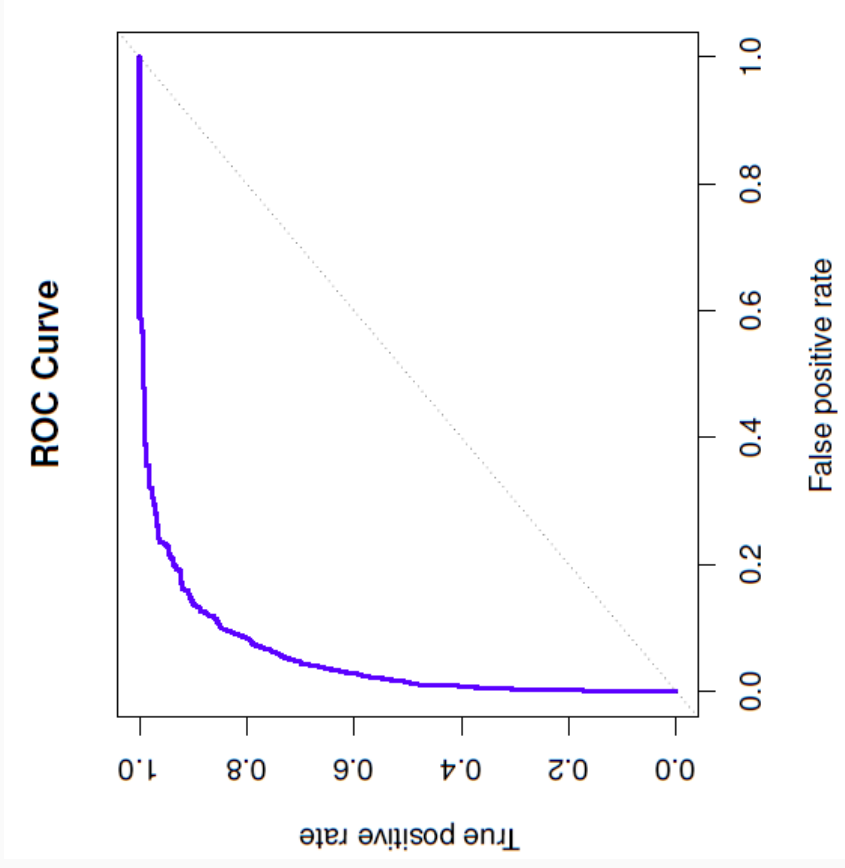


- Yandaki grafikte yatay ekseninde sınıflamada kullanılan eşik değeri
- Dikey eksen, siyah: genel hata oranı, kavuniçi nokta: yanlış pozitif oranı
- Mavi kesikli çizgi: Yanlış negatif oranı (bocunu ödemedikleri halde yanlışlıkla "borçlu değil" olarak sınıflandırma oranı)

- Eşik değeri 0.5 alındığında genel hata oranı en düşüktür. Ancak yanlış negatif oranı da en yüksektir.
- Eşik değeri azaldıkça, borcunu ödemeyenler içinde yanlış sınıflananların oranı (yanlış negatif oranı) azalmaktadır.
- Diğer taraftan, borcunu ödeyenler içinde yanlış sınıflananların oranı da artmaktadır.
- Hangi eşik değerini kullanacağımıza nasıl karar verebiliriz?

ROC Eğrisi

Bu amaçla literatürde ROC eğrisi olarak bilinen "Karar Değerlendirme Eğrisi" kullanılabilir.



- Yatay eksen: yanlış pozitif oranı
- Dikey eksen: doğru pozitif oranı
- Tüm eşik değerleri için bu oranlar hesaplanıp grafik çizilir.
- Genel performans ölçütü = AUC
- $AUC = \text{Area Under the Curve (Eğrinin altındaki alan)}$
- Maks. $AUC = 1$, yüksek AUC tercih edilir.
- Default verileri için $AUC = 0.95$

Hata Matrisi: Özet

| | | TAHMİN EDİLEN | | Toplam |
|--------------|---|------------------------------------|-----------------------------------|-----------|
| | | – | + | |
| GERÇEK DURUM | – | Doğru Negatif (DN) | Yanlış Pozitif (YP) Tip I Hata | N |
| | + | Yanlış Negatif (YN) Tip II Hata | Doğru Pozitif (DP) | P |
| Toplam | | N* | P* | N+P=N*+P* |

H_0 : – (hastalık yok)

H_a : + (hastalık var)

Yanlış Pozitif Oramı = YP/N = Tip I hata Oramı = $1 - \text{Özgüllük (specificity)}$

Doğru Pozitif Oramı = DP/P = Duyarlılık (sensitivity) = $1 - \text{Tip II Hata Oramı}$

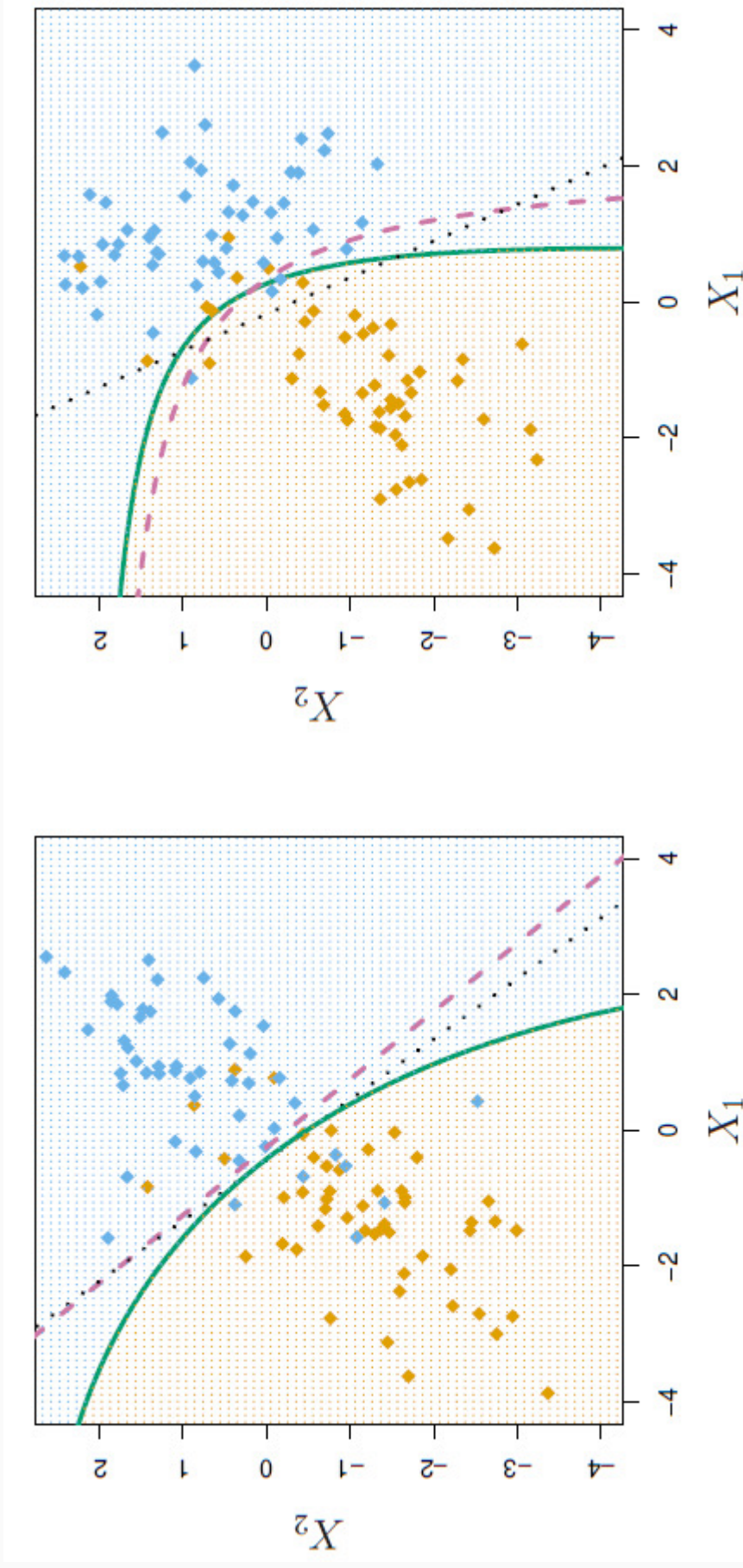
Tip I Hata: H_0 Doğru iken RED (gerçekte hasta olmadıkları halde hasta olarak sınıflandırma)

Tip II Hata: H_0 Yanlış iken KABUL (gerçekte hasta oldukları halde “hasta değil” diye sınıflandırma)

Karesel Diskriminant Analizi (QDA)

- LDA her grupta varyansın aynı olduğu varsayımını yapıyordu.
- Ancak bu varsayım sağlanmıyorsa LDA performansı kötüleşebilir.
- QDA (quadratic discriminant analysis) yöntemi LDA yöntemine benzer. Ortak varyans varsayımı yerine grup varyanslarının farklı olduğu varsayımını yapar.
- Farklı varyans varsayımı altında ortaya çıkan diskriminant fonksiyonu doğrusal değil kareseldir.
- Varyansların çok farklı olduğu verilerde QDA daha iyi bir başarıma sahip olabilir. Ancak her grupta yeterli gözlemlerin olması gerekir.
- Grup varyansları aynı ise LDA tercih edilebilir.

LDA vs. QDA



Sol: gerçek modelde varyans-kovaryans matrisi her iki grup için aynı. LDA (siyah kesikli) QDA (yeşil) yöntemine göre daha başarılı (mor kesikli: Bayes sınırı); **Sağ:** Grup varyansları farklı, QDA daha başarılı (kaynak: James et al., ISLR, Fig-4.9, p.150)