

## 9. HAFTA

### DİSKRİMİNANT ANALİZİ

Diskriminant analizi önceden bilinen farklı kitlelerden (gruplardan, sınıflardan, kümelerden) birine, üzerinde ölçüm yapılan yeni bir birimin (bireyin) atanması (sınıflandırılması) biçiminde tanımlanan istatistiksel bir tekniktir.

#### İki Kitle İçin Fisher Sınıflandırma Yöntemi

Genel düşünce gözlemleri, bilinen iki kitleden veya sınıftan birine dağıtmak ya da yeni bir gözlemi sınıflardan birine atamaktır.  $\Pi_1$  ve  $\Pi_2$  sınıfları veya kitleleri gösterebilir.

$\underline{X}' = (X_1, X_2, \dots, X_p)$  birey (birim) üzerinden alınacak ölçümlere karşılık gelen  $p$  boyutlu rasgele vektör olsun.  $\underline{X}$  'in gözlem değeri kitleden kitleye (gruptan gruba) değişecektir.  $\underline{X}$  rasgele vektörüne ilişkin gözlem değeri,  $\underline{x}$ ,  $\Pi_1$  'de ise  $\underline{X}$  'in olasılık yoğunluk fonksiyonu  $f_1(\underline{x})$  ve  $\Pi_2$  'de ise  $\underline{X}$  'in olasılık yoğunluk fonksiyonu  $f_2(\underline{x})$  dir.

$$\underline{\mu}_1 = E(\underline{X} / \Pi_1) : \Pi_1 \text{ 'e ait } \underline{X} \text{ rasgele vektörünün beklenen değeri}$$

$$\underline{\mu}_2 = E(\underline{X} / \Pi_2) : \Pi_2 \text{ 'ye ait } \underline{X} \text{ rasgele vektörünün beklenen değeri}$$

ve iki kitle için varyans-kovaryans matrislerinin eşit olduğu kabul edilirse

$$\begin{aligned} Cov(\underline{X} / \Pi_1) &= Cov(\underline{X} / \Pi_2) = Cov(\underline{X}) \\ \underline{\Sigma} &= E(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)', \quad i = 1, 2 \end{aligned}$$

olmak üzere,

$$Y = \underline{l}'_{1 \times p} \underline{X}_{p \times 1}$$

doğrusal bileşimi göz önüne alınsın. Buradan,

$$\begin{aligned} \mu_{1Y} &= E(Y / \Pi_1) \\ &= E(\underline{l}' \underline{X} / \Pi_1) \\ &= \underline{l}' \underline{\mu}_1 \end{aligned}$$

$$\begin{aligned}
\mu_{2Y} &= E(Y / \Pi_2) \\
&= E(\underline{l}'\underline{X} / \Pi_2) \\
&= \underline{l}'\underline{\mu}_2
\end{aligned}$$

ve

$$\begin{aligned}
\sigma_Y^2 &= \text{Var}(Y / \Pi_1, \Pi_2) \\
&= \text{Var}(\underline{l}'\underline{X}) \\
&= \underline{l}'\text{Cov}(\underline{X})\underline{l} \\
&= \underline{l}'\underline{\Sigma}\underline{l}
\end{aligned}$$

dir. En iyi doğrusal birleşim iki kitle için  $Y$  'nin ortalamaları arası kare uzaklığının,  $Y$  'nin varyansına oranlanarak bulunan ifadeyi maksimum yapacak şekilde elde edilmiştir. Yani en iyi doğrusal birleşim

$$\begin{aligned}
\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} &= \frac{(\underline{l}'\underline{\mu}_1 - \underline{l}'\underline{\mu}_2)^2}{\underline{l}'\underline{\Sigma}\underline{l}} \\
&= \frac{\underline{l}'(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\underline{l}}{\underline{l}'\underline{\Sigma}\underline{l}} \\
&= \frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\underline{\Sigma}\underline{l}}
\end{aligned}$$

oranından elde edilir, burada  $\underline{\delta} = (\underline{\mu}_1 - \underline{\mu}_2)$  iki kitle ortalama vektörlerinin farkıdır.  $p \times p$  tipindeki  $\underline{\delta}\underline{\delta}'$  matrisi,  $\Pi_1$  ve  $\Pi_2$  kitlelerinin ortalamaları arası fark bileşenlerinin kareler ve çapraz çarpımlar toplamı matrisidir.

$\frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\underline{\Sigma}\underline{l}}$  ifadesi her  $c \neq 0$  için  $\underline{l} = c\underline{\Sigma}^{-1}\underline{\delta} = c\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$  'nin seçilmesiyle maksimum olur.  $c = 1$

alınmasıyla elde edilen

$$\begin{aligned}
Y &= \underline{l}'_{1 \times p} \underline{X}_{p \times 1} \\
&= (\underline{\mu}_1 - \underline{\mu}_2)'\underline{\Sigma}^{-1}\underline{X}
\end{aligned}$$

Lineer birleşimine Fisher'in Lineer Diskriminant Fonksiyonu (LDF) denir. Oranın maksimumu

$$\max_{\underline{l}} \frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\underline{\Sigma}\underline{l}} = \underline{\delta}' \underline{\Sigma}^{-1} \underline{\delta}$$

ile verilir.

Linear diskriminant fonksiyonu, çok deęişkenli  $\Pi_1$  ve  $\Pi_2$  kitlelerini öyle tek deęişkenli kitlelere dönüştürür ki, bu tek deęişkenli kitlelerin ortalamaları arası fark, kitle varyansına göre mümkün olduğunca büyük olsun.

Yeni bir  $\underline{x}_0$  gözlemi için diskriminant fonksiyonun deęeri  $y_0 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{x}_0$  biçiminde tanımlansın ve bu linear birleşime göre oluşan iki tek deęişkenli kitlenin ortalamalarının orta noktası

$$\begin{aligned} m &= \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) \\ &= \frac{1}{2}(\underline{l}' \underline{\mu}_1 + \underline{l}' \underline{\mu}_2) \\ &= \frac{1}{2}(\underline{l}'(\underline{\mu}_1 + \underline{\mu}_2)) \\ &= \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \end{aligned}$$

dır. Buradan,

$$E(Y_0 / \Pi_1) - m \geq 0$$

ve

$$E(Y_0 / \Pi_2) - m < 0$$

dır. Yani eęer yeni birim  $\underline{X}_0$ ,  $\Pi_1$ 'den ise  $Y_0$ 'ın beklenen deęeri orta noktadan büyük,  $\underline{X}_0$ ,  $\Pi_2$ 'den ise  $Y_0$ 'ın beklenen deęeri orta noktadan küçük olacaktır. Böylece sınıflandırma kuralı:

$$y_0 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{x}_0 \geq m \text{ ise } \underline{x}_0, \Pi_1' \text{ e atanır}$$

$$y_0 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{x}_0 < m \text{ ise } \underline{x}_0, \Pi_2' \text{ ye atanır}$$

biçimindedir. Ayrıca aynı sınıflandırma kuralı

$$y_0 - m \geq 0 \text{ ise } \underline{x}_0, \Pi_1' \text{ e atanır}$$

$$y_0 - m < 0 \text{ ise } \underline{x}_0, \Pi_2' \text{ ye atanır}$$

biçiminde de ifade edilebilir.

Burada verilen diskriminant fonksiyonunda kitle parametreleri  $\underline{\mu}_1$   $\underline{\mu}_2$  ve  $\Sigma$  kitle parametreleri bilinmemektedir. Kitle parametreleri bilinmiyorsa, ilgili kitlelerden alınan örneklem parametreler tahmin edilir. Her bir kitleden alınan  $n_1$  ve  $n_2$  birimlik örneklemlerden sırasıyla kitle parametrelerinin tahmin edicileri  $\hat{\underline{\mu}}_1 = \bar{\underline{X}}_1$ ,  $\hat{\underline{\mu}}_2 = \bar{\underline{X}}_2$  ve  $\hat{\Sigma} = S_{pooled}$  olmak üzere Fisher'in Örneklem Linear Diskriminant Fonksiyonu

$$Y = \hat{\underline{l}}' \underline{X}$$

$$= (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S_{pooled}^{-1} \underline{X}$$

olarak elde edilir. Burada  $S_{pooled}$  iki kitle için birleştirilmiş örneklem varyans kovaryans matrisidir ve  $S_{pooled} = S$  alınacaktır.

Böylece iki tek değişkenli kitlenin örneklem ortalama değerleri  $\bar{y}_1 = \hat{\underline{l}}' \underline{x}_1$  ve  $\bar{y}_2 = \hat{\underline{l}}' \underline{x}_2$  arasındaki orta nokta

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

$$= \frac{1}{2}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S^{-1} (\bar{\underline{X}}_1 + \bar{\underline{X}}_2)$$

dir ve örneklem bağı sınıflandırma kuralı

$$y_0 = (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S^{-1} \underline{x}_0 \geq \hat{m} \text{ ise } \underline{x}_0, \Pi_1' \text{ e atanır}$$

$$y_0 = (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S^{-1} \underline{x}_0 < \hat{m} \text{ ise } \underline{x}_0, \Pi_2' \text{ e atanır}$$

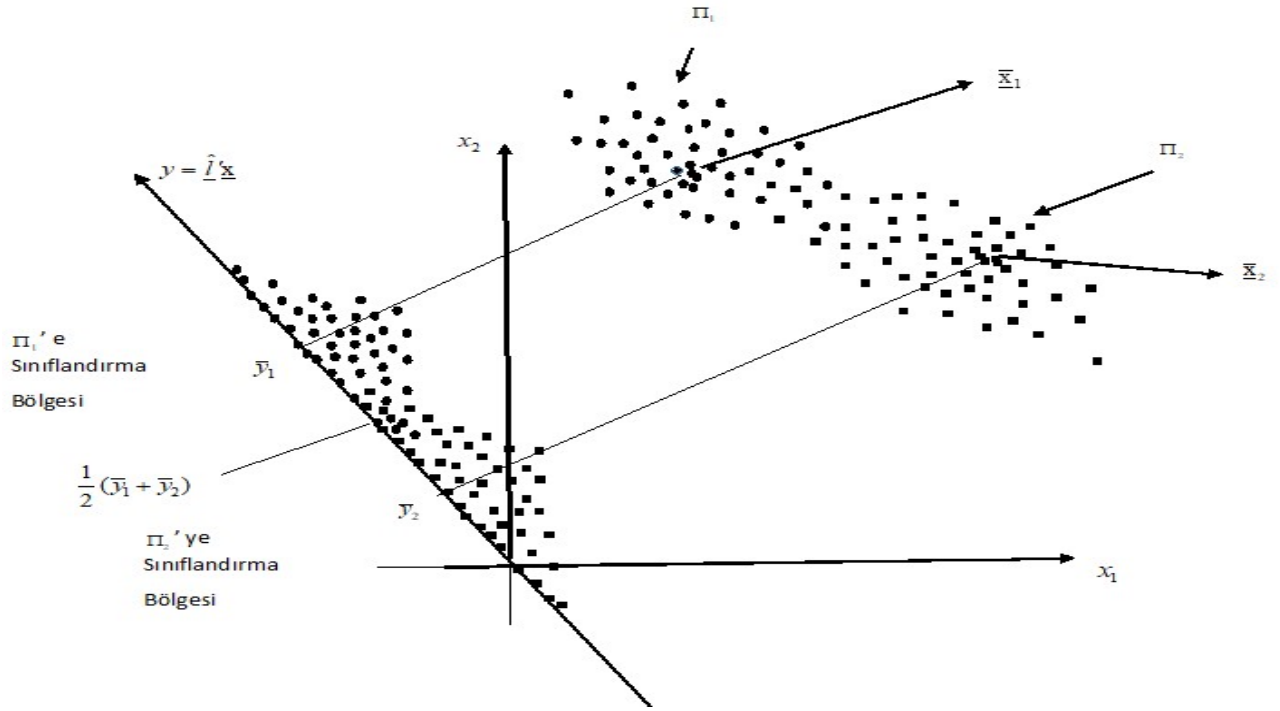
biçiminde ya da

$$y_0 - \hat{m} \geq 0 \text{ ise } \underline{x}_0, \Pi_1' \text{ e atanır}$$

$$y_0 - \hat{m} < 0 \text{ ise } \underline{x}_0, \Pi_2' \text{ ye atanır}$$

biçiminde de ifade edilebilir.

Ayrıştırma ve sınıflandırma problemi için Fisher çözümü  $p = 2$  için aşağıdaki şekilde gösterilmiştir.



**Örnek 19 :** A tipi Hemophilia hastalığını taşıyan potansiyelin ortaya çıkarılması için kan örnekleri iki kadın grubu için analiz edilmiş ve iki değişken üzerinden

$$X_1 : \log_{10}(AHF \text{ aktivitesi})$$
$$X_2 : \log_{10}(AHF \text{ aynı antijen})$$

ölçüm değerleri elde edilmiştir.  $n_1 = 30$  kadından oluşan ilk grup hastalığı taşımayan grup kitlesinden rasgele seçilmiştir. Bu gruba normal grup adı verilmiştir.  $n_2 = 22$  kadından oluşan ikinci grup hemophilia hastası olduğu bilinen kadınların kitlesinden rasgele seçilmiştir.

A tipi taşıyıcılar :hemophilik hastaların kızları, birden çok hemophilik erkek çocuğu olan anneler, bir hemophilik erkek çocuğu ve diğer hemophilik bağıntılı olan anneler. Bu özellikleri taşıyanlara zorunlu taşıyanlar grubu denir.

Ölçümlerden elde edilen örneklem değerleri;

$$\bar{\underline{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \quad \bar{\underline{x}}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \text{ ve } S^{-1}_{pooled} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Buna göre Fisher'in örneklem lineer diskriminant fonksiyonunu, iki tek değişkenli örneklem ortalamalarını ve bunlar arasındaki orta noktayı bulunuz.  $x_1 = -0.210$ ,  $x_2 = -0.044$  ölçümlerine sahip bir kadın normal gruba mı, zorunlu taşıyıcılar grubuna mı sınıflandırılır? Her iki normalleştirme metodunu kullanarak orta noktaları elde edip sonucu tekrar yorumlayınız.

### Çözüm19:

$\Pi_1$  : Hastalığı taşımayan grup (bu gruptan alınan örneklem  $n_1 = 30$ )

$\Pi_2$  : Hastalığı taşıyan grup (bu gruptan alınan örneklem  $n_2 = 22$ )

Fisher'in örneklem lineer diskriminant fonksiyonu :

$$\begin{aligned} y &= \hat{l}'\underline{x} = [\bar{\underline{x}}_1 - \bar{\underline{x}}_2]' S^{-1}_{pooled} \underline{x} \\ &= [0.2418 \quad -0.0652] \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 37.61x_1 - 28.92x_2 \end{aligned}$$

olarak elde edilir. İki tek değişkenli örneklem ortalamaları

$$\bar{y}_1 = \hat{l}'\bar{\underline{x}}_1 = [37.61 \quad -28.92] \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \hat{l}'\bar{\underline{x}}_2 = [37.61 \quad -28.92] \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} = -10.10$$

Bu ortalamalar arasındaki orta nokta:

$$\hat{m} = \frac{(\bar{y}_1 + \bar{y}_2)}{2} = -4.61.$$

Fisher'in örneklem lineer diskriminant fonksiyonuna bağlı atama kuralına göre:

Eğer  $y_0 = \hat{l}'\underline{x}_0 \geq \hat{m} = -4.61$  ise  $\underline{x}_0, \Pi_1$  'e atanır.

Eğer  $y_0 = \hat{l}'\underline{x}_0 < \hat{m} = -4.61$  ise  $\underline{x}_0, \Pi_2$  'ye atanır.

Soruya dönecek olursak

$$y_0 = \hat{l}'x_0 = [37.61 \quad -28.92] \begin{bmatrix} -0.210 \\ -0.044 \end{bmatrix} = -6.62 < -4.61$$

olduğundan bu kadın zorunlu taşıyıcılar kitlesi olan  $\Pi_2$  'ye atanır.