10. HAFTA

Genel Sınıflandırma Problemi

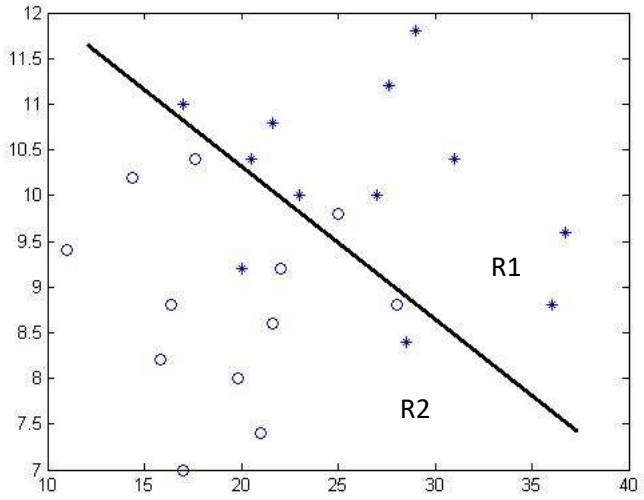
Bu bölümde iki kitle olması durumundaki sınıflandırma problemi üzerinde durulacaktır. İkiden çok kitle için de benzer sonuçlar genelleştirilebilir.

Atama veya sınıflandırma kuralları örneklemlerin değerlendirilmesinden elde edilir. İki kitlenin birinden geldiği bilinen ve rasgele seçilen bir birimin özelliklerinin ölçüm değerlerine göre değerlendirilir. Örneklem uzayının R_1 ve R_2 gibi iki ayrık bölgeye ayrıldığı kabul edilsin. Eğer yeni bir birime ait gözlem değeri; R_1 bölgesinde ise bu gözleme sahip birim Π_1 kitlesine, R_2 bölgesinde ise bu gözleme sahip birim Π_2 kitlesine atanır. Böylece birimlerin bir kümesi Π_1 de, diğer kümesi Π_2 kitlesinde yer alır. Ancak bazı birimler özelliklerinin gözlem değerlerine göre gerçek kitlesinden farklı kitleye de atanabilir. Yani bazı birimler gerçek kitlesindense diğer kitleye hatalı atanmış olur. İyi bir sınıflandırma yönteminde birkaç tane birimin hatalı atanması yani hatalı sınıflandırma olasılıklarının küçük çıkması beklenir.

Örnek 20 : Π_1 : üzerine binilerek kesme yapan araca sahip olanların grubunu ve Π_2 : böyle araca sahip olmayanların grubunu göstersin. Bir satış kampanyasında en iyi satış profilini belirlemek için bu tür aleti üreten firmalar, $x_1 : gelir$ ve $x_2 : sahip\, oldugu\, arazi\, büyüklügü$ verilerine bağlı bu tür araca sahip olan ve olmayan ailelerin sınıfıyla ilgilenmektedir. Her bir kitleden 12 birimlik rasgele örneklemler alınmış ve elde edilen değerler aşağıdadır:

Π_1 grubu		Π_2 grubu	
x_1	x_2	x_1	x_2
20,0	9,2	25,0	9,8
28,5	8,4	17,6	10,4
21,6	10,8	21,6	8,6
20,5	10,4	14,4	10,2
29,0	11,8	28,0	8,8
36,7	9,6	16,4	8,8
36,0	8,8	19,8	8,0
27,6	11,2	22,0	9,2

23,0	10,0	15,8	8,2
31,0	10,4	11,0	9,4
17,0	11,0	17,0	7,0
27,0	10,0	21,0	7,4



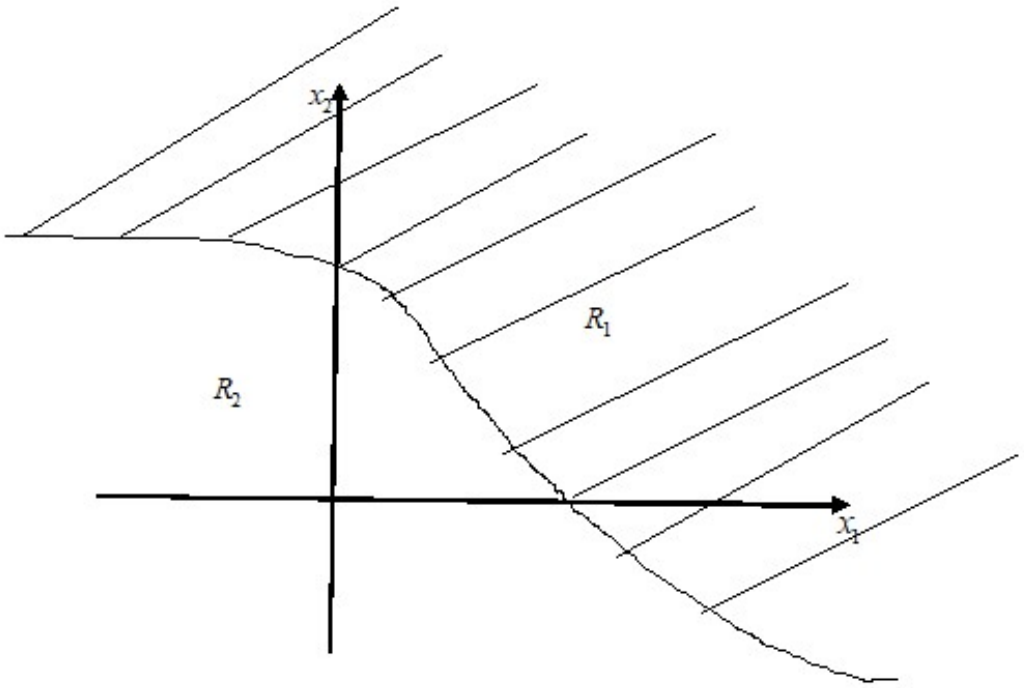
Yorum : Arazi büyüklüğü gelire göre daha iyi ayrım yapmasına rağmen, aleti olanların geliri ve arazisi olmayanlara göre daha çok olduğu görülmektedir. Diğer taraftan iki grup arasında çakışma vardır. Örneğin eğer Π_1 ve Π_2 kitlelerindeki $\left(x_1, x_2\right)$ değerlerini şekildeki R1 ve R2 bölgelerini ayıran düz doğruya göre atamasını yaparsak hata yapmış oluruz. Π_1'deki bazı birimler Π_2'ye, Π_2'deki bazı birimler Π_1'e yanlışlıkla sınıflandırılabilir. Amaç R1 ve R2 bölgelerini belirleyecek öyle bir kural oluşturulsun ki bu hatalı atamalar minimum olsun.

İki kitleden biri diğerine göre göreceli olarak daha büyük(geniş) olabileceğinden, kitlelerden biri diğerine göre daha büyük olabilirliliğe sahip olacaktır. Diğer bir ifade ile gözlemlerin ait oldukları kitlelere ilişkin ön bilgi adı verilen önsel(prior) olasılıklar farklı olabilir. Dolayısıyla bir birimin geldiği kitleye ilişkin önsel olasılık, büyük kitle için daha büyük olacaktır. Ancak tüm kitleler için önsel olasılıkların toplamı bire eşit olmalıdır. Bu önsel olasılıklar, kitle ağırlıkları olarak da ifade edilebilir.

Sınıflandırma da diğer önemli bir kavram da maliyettir. Birimlerin ait oldukları kitleden farklı bir kitleye atanması, bir çok uygulamada büyük maliyetlere neden olabilir. İyi bir sınıflandırma yöntemin de eğer mümkünse hatalı sınıflandırma maliyetleri hakkında bilgi olmalıdır.

$f_1(\underline{x})$ ve $f_2(\underline{x})$, Π_1 ve Π_2 kitleleri için $\underline{X}_{px1}$ rasgele vektörüne ilişkin olasılık yoğunluk fonksiyonları olsun. $\underline{x}$ ölçümüne sahip bir gözlem Π_1 ve Π_2 kitlelerinden birine atanmalıdır. Ω, $\underline{x}$'in olası tüm değerlerinden oluşan örneklem uzayı olsun. Ayrıca R_1, Π_1'e sınıflandırılan birimler için $\underline{x}$ değerlerinin bir kümesi ve R_2, Π_2'ye sınıflandırılan birimler için $\underline{x}$ değerlerinin bir kümesi olsun. Burada $\Omega = R_1 \cup R_2$ dir ve bir birim iki kitleden sadece birine atanacağından R_1 ve R_2 bölgeleri ayrıktır, yani $R_1 \cap R_2 = \varnothing$ dir.

$p = 2$ için sınıflandırma bölgeleri aşağıdaki şekilde gösterilmiştir.



Π_1 kitlesine ait bir birimin Π_2 kitlesine atanmasının(sınıflandırılmasının) koşullu olasılığı

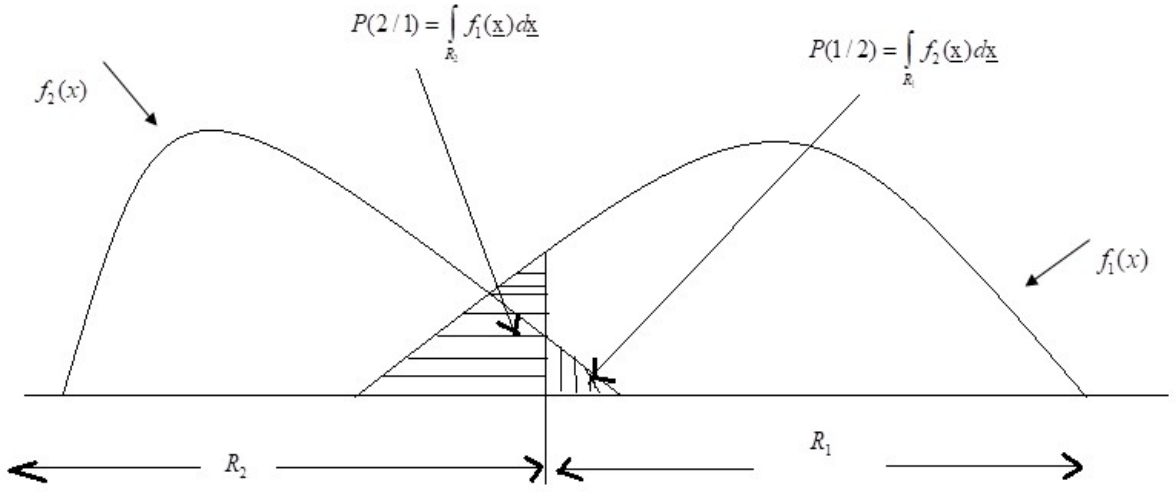
$$P(2/1) = P(\underline{X} \in R_2 / \Pi_1)$$
$$= \int_{R_2} f_1(\underline{x}) d\underline{x}$$

dir. Benzer biçimde Π_2 kitlesine ait bir birimin Π_1 kitlesine atanmasının(sınıflandırılmasının) koşullu olasılığı

$$P(1/2) = P(\underline{X} \in R_1 / \Pi_2)$$
$$= \int_{R_1} f_2(\underline{x}) d\underline{x}$$

dir.

$p = 1$ olduğunda yani birim üzerinde sadece bir özelliğe ilişkin ölçüm yapıldığında, ölçüme karşılık gelen rasgele değişkenin her iki kitle üzerinde olasılık yoğunluk fonksiyonlarına bağlı olarak sınıflandırma bölgeleri için koşullu olasılıklar aşağıdaki şekilde gösterilmiştir.



Π_2 'ye sınıflandırma bölgesi Π_1 'ye sınıflandırma bölgesi

$p_1 = P(\underline{X} \in \Pi_1)$, Π_1 'in ve $p_2 = P(\underline{X} \in \Pi_2)$, Π_2 'nin önsel olasılığı olsun. Birimlerin hatalı veya doğru sınıflandırılmasına ilişkin olasılıklar; önsel olasılıklar ile koşullu olasılıkların çarpılmasıyla elde edilebilir.

$$P(\Pi_1\text{'e doğru sınıflandırma}) = P(\text{Gözlemin } \Pi_1\text{'den gelmesi ve } \Pi_1\text{'e doğru sınıflandırılması})$$
$$= P(\underline{X} \in R_1 / \Pi_1) P(\underline{X} \in \Pi_1)$$
$$= P(1/1) p_1$$

$$P(\Pi_1\text{'e hatalı sınıflandırma}) = P(\text{Gözlemin } \Pi_2\text{'den gelmesi ve } \Pi_1\text{'e hatalı sınıflandırılması})$$
$$= P(\underline{X} \in R_1 / \Pi_2) P(\underline{X} \in \Pi_2)$$
$$= P(1/2)\, p_2$$

$$P(\Pi_2\text{'ye doğru sınıflandırma}) = P(\text{Gözlemin } \Pi_2\text{'den gelmesi ve } \Pi_2\text{'ye doğru sınıflandırılması})$$
$$= P(\underline{X} \in R_2 / \Pi_2) P(\underline{X} \in \Pi_2)$$
$$= P(2/2)\, p_2$$

$$P(\Pi_2\text{'ye hatalı sınıflandırma}) = P(\text{Gözlemin } \Pi_1\text{'den gelmesi ve } \Pi_2\text{'ye hatalı sınıflandırılması})$$
$$= P(\underline{X} \in R_2 / \Pi_1) P(\underline{X} \in \Pi_1)$$
$$= P(2/1)\, p_1$$

Birimlerin sınıflandırılmasında maliyetler de önemli rol oynar. Hatalı sınıflandırma maliyetleri maliyet matrisinde aşağıdaki gibi tanımlanabilir.

		Sınıflandırılan Kitle	
		Π_1	Π_2
Doğru Kitle	Π_1	0	$C(2/1)$
	Π_2	$C(1/2)$	0

Burada;

$C(1/2)$: Gözlemin Π_2 den olduğu bilindiğinde, Π_1 'e hatalı atanmasının maliyetidir.

$C(2/1)$: Gözlemin Π_1 den olduğu bilindiğinde, Π_2 'e hatalı atanmasının maliyetidir.

Birimler ait olduğu gerçek kitlesine doğru sınıflandırıldığında herhangi bir maliyet oluşmayacağından, maliyetler 0 olacaktır.

Herhangi bir kural için, ortalamaya veya hatalı sınıflandırmanın beklenen maliyeti (*ECM*)

$$ECM = C(2/1)P(2/1)p_1 + C(1/2)P(1/2)p_2$$

dır. ***İyi bir sınıflandırma kuralında ECM değeri küçük veya mümkün olduğunca küçük olmalıdır.***

İki Kitle için Sınıflandırma Kuralları

İyi bir sınıflandırma kuralı *ECM*'nin minimizasyonu ile elde edilebilir. Diğer bir ifade ile öyle R_1 ve R_2 bölgeleri seçilsin ki *ECM* mümkün olduğunca küçük olsun.

Sonuç: *ECM* değerini minimize eden R_1 ve R_2 bölgeleri aşağıdaki eşitsizlikler geçerli olacak biçimde $\underline{x}$ değeriyle tanımlanır:

$$R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

$$R_2 : \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

Ispat: $P(2/1)$ ve $P(1/2)$ için integral gösterimlerinden,

$$ECM = C(2/1)\,p_1 \int_{R_2} f_1(\underline{x})d\underline{x} + C(1/2)\,p_2 \int_{R_1} f_2(\underline{x})d\underline{x}$$

olarak yazılabilir. $\Omega = R_1 \cup R_2$ olduğundan, toplam olasılık

$$\begin{aligned} 1 &= \int_{\Omega} f_1(\underline{x})d\underline{x} \\ &= \int_{R_1} f_1(\underline{x})d\underline{x} + \int_{R_2} f_1(\underline{x})d\underline{x} \end{aligned}$$

dir. Böylece,

$$\begin{aligned} ECM &= C(2/1)\,p_1\left[1 - \int_{R_1} f_1(\underline{x})d\underline{x}\right] + C(1/2)\,p_2 \int_{R_1} f_2(\underline{x})d\underline{x} \\ &= \int_{R_1}\left[C(1/2)\,p_2 f_2(\underline{x}) - C(2/1)\,p_1 f_1(\underline{x})\right]d\underline{x} + C(2/1)\,p_1 \end{aligned}$$

olur. p_1, p_2, $C(1/2)$ ve $C(2/1)$ pozitiftir. Bununla birlikte $f_1(\underline{x})$ ve $f_2(\underline{x})$ fonksiyonları bütün $\underline{x}$ 'ler için pozitiftir ve *ECM* ifadesi de $\underline{x}$' bağlıdır. Böylece, eğer R_1 bölgesi

$$\left[C(1/2)\,p_2 f_2(\underline{x}) - C(2/1)\,p_1 f_1(\underline{x})\right] \leq 0$$

olan $\underline{x}$ değerlerini içeriyorsa *ECM* minimum olur ve $\underline{x}$'in bu değerleri hariç ifade pozitiftir. Yani R_1 bölgesi

$$C(1/2)p_2 f_2(\underline{x}) \le C(2/1)p_1 f_1(\underline{x})$$

veya

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \ge \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

sonucunu sağlayan $\underline{x}$ değerlerinin bir kümesidir. R_2 bölgesi, Ω da R_1 bölgesinin tümleyeni olduğundan, R_2 bölgesi;

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

sonucunu sağlayan $\underline{x}$ değerlerinin bir kümesidir. Bu ifadelerde yer alan maliyetlerin belirlenmesi zordur.

Beklenen Maliyet Bölgelerinin Minimizasyonunun Özel Durumları:

a) Önsel (prior) olasılıkların eşit olması. Yani $\frac{p_2}{p_1} = 1$

$$R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \ge \left[\frac{C(1/2)}{C(2/1)}\right] \; ; \; R_2 : \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{C(1/2)}{C(2/1)}\right]$$

b) Hatalı sınıflandırma maliyetlerinin eşit olması. Yani $\frac{C(1/2)}{C(2/1)} = 1$

$$R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \ge \left[\frac{p_2}{p_1}\right] \; ; \; R_2 : \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{p_2}{p_1}\right]$$

c) Hem önsel olasılıkların hem de maliyetlerin eşit olması. Yani $\frac{p_2}{p_1} = 1$ ve $\frac{C(1/2)}{C(2/1)} = 1$

$$R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \ge 1 \; ; \; R_2 : \frac{f_1(\underline{x})}{f_2(\underline{x})} < 1$$

Önsel olasılıklar bilinmediğinde genelde eşit alınırlar ve en küçük *ECM* kuralı uygun hatalı sınıflandırma maliyetleri ile kitle olasılık yoğunluk fonksiyonlarının karşılaştırılmasını içerir. Eğer hatalı sınıflandırma maliyetleri belirlenemez ise eşit alınırlar ve kitle olasılık yoğunluk

fonksiyonlarının oranı, önsel olasılıkların oranı ile karşılaştırılır. Son olarak, hem önsel olasılıklar, hem de hatalı sınıflandırma maliyetleri bilinmediğinde önsel olasılıkların oranı ve hatalı sınıflandırma maliyetlerinin oranı bire eşit alınarak en iyi sınıflandırma bölgeleri yoğunluk fonksiyonlarının değerlerinin karşılaştırılmasıyla belirlenir. Bu durumda yeni bir gözlem $\underline{x}_0$ için $\frac{f_1(\underline{x}_0)}{f_2(\underline{x}_0)} \geq 1$ (veya $f_1(\underline{x}_0) \geq f_2(\underline{x}_0)$) ise $\underline{x}_0$ gözlem değerine sahip birim Π_1 kitlesine atanır. Diğer taraftan $\frac{f_1(\underline{x}_0)}{f_2(\underline{x}_0)} < 1$ ise $\underline{x}_0$ gözlem değerine sahip birim Π_2 kitlesine atanır.

Hatalı sınıflandırmanın beklenen maliyetinin minimizasyonundan başka bir sınıflandırma kriteri en iyi (optimal) sınıflandırma yöntemidir. Hatalı sınıflandırma maliyetleri ihmal edildiğinde, R_1 ve R_2 bölgelerinin seçimi, Toplam Hatalı Sınıflandırma Olasılığının (*TPM*) minimizasyonu ile belirlenir.

$TPM = P($ Bir birimin Π_1'e veya Π_2'ye hatalı sınıflandırılması)

$= P$ (Birim Π_1'den gelsin ve hatalı sınıflandırılsın)

$+ P$ (Birim Π_2'den gelsin ve hatalı sınıflandırılsın)

biçimindedir. Buradan,

$$TPM = p_1 \int_{R_2} f_1(\underline{x}) d\underline{x} + p_2 \int_{R_1} f_2(\underline{x}) d\underline{x}$$

dir. Matematiksel olarak bu ifade, hatalı sınıflandırma maliyetleri eşit olduğu durumdaki hatalı sınıflandırmanın beklenen maliyetinin minimizasyonu ile eşdeğerdir.

Ayrıca $\underline{x}_0$ gözlem değerine sahip bir birim, $P(\Pi_i / \underline{x}_0)$, $i = 1, 2$ sonsal (posterior) olasılığı büyük olan kitleye atanır. Burada,

$$P(\Pi_1 / \underline{x}_0) = \frac{p_1 f_1(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)}$$

ve

$$P(\Pi_2 / \underline{x}_0) = 1 - P(\Pi_1 / \underline{x}_0)$$
$$= \frac{p_2 f_2(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)}$$

dır. Böylece, $P(\Pi_1 / \underline{x}_0) > P(\Pi_2 / \underline{x}_0)$ olduğunda, $\underline{x}_0$ gözlem değerine sahip birim Π_1 kitlesine atanır.

Varyans-Kovaryans Matrisleri Eşit Olan İki Çok Değişkenli Normal Kitle için Sınıflandırma

Π_1 ve Π_2 , yoğunluk fonksiyonları $f_1(\underline{x})$ ve $f_2(\underline{x})$, kitle ortalama vektörleri $\underline{\mu}_1$, $\underline{\mu}_2$ ve varyans–kovaryans matrisleri Σ_1, Σ_2 olan çok değişkenli (p-boyutlu) normal dağılıma sahip kitleler olduğunu kabul edelim.

İlk olarak her iki kitle için varyans-kovaryas matrislerinin eşit olduğu durumu göz önüne alalım. Bu durum Fisherin iki kitle için elde edilen lineer diskriminant fonksiyonu kullanılabilir.

$\Sigma_1 = \Sigma_2 = \Sigma$ **olsun**

Π_1 ve Π_2 kitleleri için $\underline{X}' = (X_1, X_2, ..., X_p)$ rasgele vektörünün ortak yoğunluk fonksiyonu

$$f_i(\underline{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma^{-1}(\underline{x}-\underline{\mu}_i)} \quad , \quad i = 1,2$$

dir, burada $\underline{\mu}_i \in \mathbb{R}^p$ ve Σ pozitif tanımlı kare matristir. Ayrıca $\underline{\mu}_1$, $\underline{\mu}_2$ ve Σ 'nın bilindiği kabul edilsin. Böylece,

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} = \frac{e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)}}{e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)}}$$
$$= e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)+\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)}$$

dir. Buradan minimum *ECM* bölgeleri

$$R_1 : (e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)+\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)}) \geq \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

$$R_2 : (e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)+\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)}) < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$$

olarak elde edilir. Bu şekilde R_1 ve R_2 bölgeleri verildiğinde, aşağıdaki sınıflandırma kuralı elde edilir.

Sonuç: Π_1 ve Π_2 çok değişkenli normal yoğunluk fonksiyonuna sahip kitleler olsunlar. *ECM*'yi minimize eden atama kuralı eğer

$$(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x}_0 - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2) \geq \ln\left(\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \right)$$

ise $\underline{x}_0$ gözlem değerine sahip birim Π_1'e aksi halde Π_2'ye atanır.

Ispat: Yukarıda verilen üstel ifade bütün $\underline{x}$'ler için negatif olmadığından, bu ifadenin doğal logaritması alınarak eşitlik düzenlendiğinde,

$$-\frac{1}{2}(\underline{x} - \underline{\mu}_1)'\Sigma^{-1}(\underline{x} - \underline{\mu}_1) + \frac{1}{2}(\underline{x} - \underline{\mu}_2)'\Sigma^{-1}(\underline{x} - \underline{\mu}_2) = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2)$$

dir ve sonuç olarak,

$$R_1 : (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2) \geq \ln\left(\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \right)$$

ve

$$R_2 : (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2) < \ln\left(\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \right)$$

olarak bulunur.

Minimum ECM kuralı, Fisher yöntemiyle karşılaştırıldığında $\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] = 1$ olduğunda $\ln(1) = 0$ dır. Bu durumda bu iki kural eşdeğerdir.

Bir çok durumda $\underline{\mu}_1$, $\underline{\mu}_2$ ve Σ kitle parametreleri bilinmediğinde, yukarıda verilen atama kuralı değiştirilmelidir. Yani bilinmeyen kitle parametreleri yerine tahmin edicileri kullanılarak örneklem sınıflandırma kuralı elde edilir. Çok değişkenli normal kitlelerin her birinden alınan n_1 ve n_2 birimlik örneklemlerden elde edilen $\overline{\underline{x}}_1$, $\overline{\underline{x}}_2$ ve $S_{pool} = S$ örneklem değerlerine bağlı olarak örneklem minimum *ECM* kuralı,

$$(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1} \underline{x}_0 - \frac{1}{2}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1}(\underline{\bar{x}}_1 + \underline{\bar{x}}_2) \geq \ln\left(\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \right)$$

ise $\underline{x}_0$ gözlem değerine sahip birim Π_1'e aksi halde Π_2'ye atanır biçiminde verilir. Bu ifadedeki ilk terim $y = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1} \underline{x}_0$, Fisher tarafından elde edilen lineer fonksiyondur. Bu fonksiyon, örneklemler arası değişkenliği, örneklem içi değişkenliğe göre maksimize eder.

Ayrıca,

$$\begin{aligned} W &= (\underline{\bar{X}}_1 - \underline{\bar{X}}_2)' S^{-1} \underline{X} - \frac{1}{2}(\underline{\bar{X}}_1 - \underline{\bar{X}}_2)' S^{-1}(\underline{\bar{X}}_1 + \underline{\bar{X}}_2) \\ &= (\underline{\bar{X}}_1 - \underline{\bar{X}}_2)' S^{-1}\left[\underline{X} - \frac{1}{2}(\underline{\bar{X}}_1 + \underline{\bar{X}}_2)\right] \end{aligned}$$

ifadesine Anderson sınıflandırma fonksiyonu(W istatistiği) adı da verilmektedir.

İki normal kitle aynı varyans-kovaryans matrisine sahip ise eşit önsel olasılık ve eşit hatalı sınıflandırma maliyetleri durumunda, Fisher'in sınıflandırma kuralı, minimumu *ECM* kuralıyla eşdeğerdir.

Bilinmeyen parametreler yerine örneklemlerden elde edilen tahminleri alındığında, elde edilen kural uygulamada hatalı sınıflandırmanın beklenen maliyetini minimize etmeyebilir. Optimal kuralın $f_1(\underline{x})$ ve $f_2(\underline{x})$ çok değişkenli normal yoğunluk fonksiyonlarının tamamen bilindiği durumda elde edildiğinden minimizasyon gerçekleşmeyebilir. Örneklem sınıflandırma kuralı, optimal kuralın bir tahminidir. Eğer örneklemler yeterince büyük ise, kuralın iyi çalışması beklenir.

Örnek 21: A tipi Hemophilia hastalığını taşıyan ve taşımayanlara ilişkin verilerin dağılımı çok değişkenli normal olsun. Grup elemanlarının prior olasılıkları bilindiğinde bir kadının Π_1: normal kitle veya Π_2: Zorunlu taşıyıcılar kitlesine atanmasına ilişkin daha önce elde edilen sonuçlar :

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

Şeklinde daha önce verilmişti. Hatalı sınıflandırma maliyetleri $C(1/2)=C(2/1)$ olsun. Teyze çocuğu hemophilik olan bir kişinin A tipi hemophilia hastalığına genetik olarak yakalanma

olasılığı 0.25'tir. Verilenlere göre, tahmini minimum ECM kuralını kullanarak $x_1 = -0.210,\ x_2 = -0.044$ ölçümlerine sahip bir teyze çocuğunun hangi kitleye sınıflandırılacağını araştırınız.

Çözüm 21 :

Π_1 : hastalığı taşımayan (normal) kitle $\longrightarrow$ $p_1 = 1 - 0.25 = 0.75$

Π_2 : Zorunlu taşıyıcılar kitlesine $\longrightarrow$ $p_2 = 0.25$

W sınıflandırma istatistiğinin değeri $w = \left(\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right)' S_{pooled}^{-1} \underline{x}_0 - \frac{1}{2}\left(\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right)' S_{pooled}^{-1} \left(\underline{\bar{x}}_1 + \underline{\bar{x}}_2\right)$ veya $w = \underline{\hat{l}}' \underline{x}_0 - \hat{m}$ ile elde edilir. Burada $y_0 = \underline{\hat{l}}' \underline{x}_0$ Fisher'in lineer diskriminant fonksiyonu ve $\hat{m}$ y 'nin örneklem ortalamaları arasındaki orta noktadır. Daha önceki derste çözdüğümüz örnekte, $\hat{m} = -4.61$ ve $y_0 = \underline{\hat{l}}' \underline{x}_0 = -6.62$ bulmuştuk. Buna göre,

$$w = -6.62 - (-4.61) = -2.01$$

dir. Böylece iki normal kitle için tahmini minimum ECM kuralından,

$$w = -2.01 < \ln\left(\frac{p_2}{p_1}\right) = \ln\left(\frac{0.25}{0.75}\right) = -1.10$$

olduğundan bu ölçümlere sahip kadın zorunlu taşıyıcılar (Π_2) kitlesine sınıflandırılır.