Einfache Diskriminanzanalyse

Einführendes Beispiel

Gegeben sind zwei Gruppen von Personen, für die jeweils mehrere Merkmale erhoben wurden, z.B. Alter, Einkommen, Zahl der Kinder und Autobesitz (1 = ja). Es gibt ein weiteres (abhängiges) Merkmal, die sich in den jeweiligen Gruppen unterscheidet, z.B. „kreditwürdig“ und „nicht kreditwürdig“.

Gruppe 1 („kreditwürdig“), fiktives Beispiel:

```
X1
    Alter Einkommen Kinder Autobesitz
x01    55      4503      0          1
x02    44      1651      1          1
x03    32      1518      2          0
x04    18      1253      1          0
x05    36      2471      0          1
x06    43      2126      0          1
x07    40      1272      3          0
x08    34      1393      2          1
x09    22      1626      1          0
x10    51      1941      0          1
```

Gruppe 2 („nicht kreditwürdig“), fiktives Beispiel:

```
X2
    Alter Einkommen Kinder Autobesitz
x11    40      2869      3          0
x12    39      1431      3          1
x13    34      1567      3          0
x14    31      1323      5          0
x15    44      3384      4          1
x16    55      1432      1          0
x17    56      2161      4          0
x18    47      2476      3          1
x19    62      1896      2          1
x20    43      2121      1          0
```

Für eine neue Person werden die vier o.g. Merkmale erhoben:

	Alter	Einkommen	Kinder	Autobesitz
x	41	2500	1	1

Es soll entschieden werden, welcher der beiden Gruppen diese Person zugeordnet werden kann.

Durchführung[1]

Gegeben sind zwei unabhängige Gruppen, für die (metrische) Daten mehrerer Merkmale vorliegen. Beide Gruppen entstammen mehrdimensionelen Normalverteilungen und haben gleiche[2] Kovarianzmatrizen $\underline{\Sigma}^{(1)} = \underline{\Sigma}^{(2)} = \underline{\Sigma}$.

Durch die Diskriminanzanalyse sollen die folgenden beiden Aufgaben gelöst werden:

1. Prüfung der Frage, ob die beiden Gruppen gleiche Mittelwertsvektoren haben können

2. Angabe einer linearen Funktion der Merkmalsdaten, die die Zuordnung weiterer Merkmalsträger zu einer der beiden Grundgesamtheiten ermöglicht (→ Diskriminanzfunktion)

Der Vektor $\underline{x}' = (x_1, x_2, \ldots, x_m)$ enthält die beobachteten Merkmale.

Die Erwartungswerte für die beiden Gruppen sind in vektorieller Schreibweise

$\underline{\mu}^{(1)'} = (\mu_1^{(1)}, \mu_2^{(1)}, \ldots, \mu_m^{(1)})$ bzw. $\underline{\mu}^{(2)'} = (\mu_1^{(2)}, \mu_2^{(2)}, \ldots, \mu_m^{(2)})$.

[1] Die Darstellung folgt einem von Prof. Kristof mündlich vorgetragenen Seminarinhalt, ca. 1994
[2] Die Prüfung der H_0: $\underline{\Sigma}^{(1)} = \underline{\Sigma}^{(2)}$ wird im Anhang dargestellt

Aus den beiden Gruppen werden nun zwei Stichproben der Größe n_1 und n_2 erhoben, es sei $n = n_1 + n_2$. Hieraus ergeben sich die arithmetischen Mittelwertsvektoren

$\underline{\bar{x}}^{(1)'} = (\bar{x}_1^{(1)}, \bar{x}_2^{(1)}, \ldots, \bar{x}_m^{(1)})$ bzw. $\underline{\bar{x}}^{(2)'} = (\bar{x}_1^{(2)}, \bar{x}_2^{(2)}, \ldots, \bar{x}_m^{(2)})$.

Der ersten Aufgabe entspricht die Prüfung der Nullhypothese

H_0: $\underline{\mu}^{(1)} = \underline{\mu}^{(2)}$ bzw. H_0: $\underline{\mu}^{(1)} - \underline{\mu}^{(2)} = \underline{0}$.

Die H_0 enthält mehrere skalare Gleichungen. Gleichbedeutend lässt sich auch eine skalare Gleichung

H_0: $\underline{a}'\underline{\mu}^{(1)} = \underline{a}'\underline{\mu}^{(2)}$ bzw. H_0: $\underline{a}'(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}) = 0$

schreiben, wobei $\underline{a}$ um ein m-dimensionaler Vektor ist. Diese Gleichung soll für beliebige $\underline{a}' = (a_1, \ldots, a_m)$ gelten.

Nun wird eine skalare Variable $Y(\underline{a})$ definiert, die von der Wahl des Vektors $\underline{a}$ abhängt:

$Y(\underline{a}) = \underline{a}'\underline{X} = a_1X_1 + \ldots a_mX_m$, $\underline{a} \neq \underline{0}$.

Für die Erwartungswerte der beiden Gruppen gilt

$\varepsilon Y^{(1)}(\underline{a}) = \underline{a}'\underline{\mu}^{(1)}$ bzw. $\varepsilon Y^{(2)}(\underline{a}) = \underline{a}'\underline{\mu}^{(2)}$

Die Varianz von $Y^{(1)}(a)$ und $Y^{(2)}(a)$ ist gleich, nämlich

$\underline{a}'\underline{\Sigma}a$.

Die Mittelwerte der Beobachtungsdaten in den beiden Gruppen sind

$\overline{y}^{(1)}(\underline{a}) = \underline{a}'\underline{x}^{(1)}$ bzw. $\overline{y}^{(2)}(\underline{a}) = \underline{a}'\underline{x}^{(2)}$.

Wenn die beiden Gruppen der Beobachtungswerte die Kovarianzmatrizen $\underline{S}^{(1)}$ und $\underline{S}^{(2)}$ liefern, dann hat die neue Variable y(a) die Varianz

$\underline{a}'\underline{S}^{(1)}\underline{a}$ bzw. $\underline{a}'\underline{S}^{(2)}\underline{a}$.

Folglich ergibt Y(a) zwei unabhängige und varianzhomogene Gruppen skalarer Messwerte, die Normalverteilungen entstammen. Für sie gilt: H_0: $\varepsilon Y^{(1)}(\underline{a}) = \varepsilon Y^{(2)}(\underline{a})$.

Die Prüfung der H_0 erfolgt durch den üblichen t-Test, wobei allerdings hier t^2 betrachtet wird, da ohnehin nur eine zweiseitige Alternativhypothese interessiert. Also ergibt sich als Prüfgröße

$$t^2(\underline{a}) = \frac{(\overline{y}^{(1)}(\underline{a}) - (\overline{y}^{(2)}(\underline{a}))^2}{\underline{a}'\underline{S}^{(1)}\underline{a}(n_1-1) + \underline{a}'\underline{S}^{(2)}\underline{a}(n_2-1)} \cdot \frac{n_1 n_2}{n_1+n_2}.$$

Setzt man

$$\underline{S} = \frac{\underline{S}^{(1)}(n_1-1) + \underline{S}^{(2)}(n_2-1)}{n_1+n_2-2}, \text{ dann wird}$$

$$t^2(\underline{a}) = \frac{(\overline{y}^{(1)}(\underline{a}) - (\overline{y}^{(2)}(\underline{a}))^2}{\underline{a}'\underline{S}\underline{a}} \cdot \frac{n_1 n_2}{n_1+n_2}$$

$$= \frac{(\underline{a}'(\overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)})^2}{\underline{a}'\underline{S}\underline{a}} \cdot \frac{n_1 n_2}{n_1+n_2}$$

$$= \frac{\underline{a}'(\overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)})(\overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)})'\underline{a}}{\underline{a}'\underline{S}\underline{a}} \cdot \frac{n_1 n_2}{n_1+n_2}.$$

Das Ziel besteht darin, $t^2(\underline{a})$ durch Wahl von $\underline{a}$ zu maximieren. Aus der oben dargestellten Form ergibt sich, dass sich das maximierende $\underline{a}$ durch jedes Vielfache seiner selbst ($\neq \underline{0}$) ersetzen lässt.

Es ist (ohne Beweis) möglich, $\underline{S} = \underline{S}^{1/2}\underline{S}^{1/2}$ zu bilden. Es gilt

$\underline{S}^{1/2} = \underline{S}^{1/2\prime}$ und $\underline{S}^{-1/2} = \underline{S}^{-1/2\prime}$.

Der o.g. Ausdruck lässt sich damit schreiben als

$$t^2(\underline{a}) = \frac{\underline{a}'\underline{S}^{1/2}\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{1/2}\underline{S}^{-1/2}\underline{a}}{\underline{a}'\underline{S}^{1/2}\underline{S}^{1/2\prime}\underline{a}} \cdot \frac{n_1 n_2}{n_1 + n_2}.$$

Wir nun gesetzt

$\underline{h} = \underline{S}^{1/2\prime}\underline{a} = \underline{S}^{1/2}\underline{a}$,

dann lässt sich t^2 auch in Abhängigkeit von $\underline{h}$ betrachten und schreiben als

$$t^2(\underline{h}) = \frac{\underline{h}'\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1/2}\underline{h}}{\underline{h}'\underline{h}} \cdot \frac{n_1 n_2}{n_1 + n_2},$$

wobei $h \neq \underline{0}$, wie auch von $\underline{a}$ verlangt.

Es geht also jetzt darum, $t^2(\underline{h})$ durch Wahl des geeigneten $\underline{h}$ zu maximieren. Offensichtlich hängt $t^2(\underline{h})$ nicht von der Länge von $\underline{h}$ ab, so dass $\underline{h}$ auch als Einheitsvektor genommen werden darf.

Bei $\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1/2}$ handelt es sich um eine symmetrische Matrix mit Rang 1, denn $\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ ist ein Spaltenvektor.

Also ist $\underline{h}'\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ ein Skalar, dessen Quadrat in $t^2(\underline{h})$ entscheidend eingeht. Soll diese Zahl und damit ihr Quadrat maximal werden, dann muss $\underline{h}$ als der parallel zu $\underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ verlaufende Einheitsvektor gewählt werden, also wird

$$\underline{h} = \underline{S}^{-1/2}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}).$$

Da $\underline{a} = \underline{S}^{-1/2}\underline{h}$, wird entsprechend das maximierende

$$\underline{a} = \underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}).$$

Die Länge von $\underline{a}$ ist unwesentlich, aber $\underline{a} \neq \underline{0}$.

Damit erhält man die Diskriminanzfunktion $Y(\underline{a}) = \underline{a}'\underline{X}$, die die Gruppen am besten trennt

$$Y = (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}\underline{X}.$$

Damit wird die zweite Aufgabe, die Zuweisung zukünftiger Fälle zu einer der beiden Gruppen, gelöst: Ist $\underline{x}$ ein Vektor, der neue Messwerte enthält, dann wird $\underline{x}$ der Gruppe i = 1 oder i = 2 zugeordnet, zu der der Abstand von

$$(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}\underline{x} \qquad \text{zu} \qquad (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}\underline{x}^{(i)}$$

der kleinere ist.

Zu lösen ist noch die erste Aufgabe, nämlich die Frage, ob die beiden Gruppen gleiche Mittelwertsvektoren haben. Hierfür wird das maximale $t^2(\underline{a}) = t^2_{max}$ berechnet. Hierfür muss $\underline{a} = \underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ in die Formel zur Bestimmung des $t^2(\underline{a})$ eingesetzt werden. Dies ergibt

$$t^2_{max} = \frac{(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})}{(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underbrace{\underline{S}^{-1}\underline{S}}_{=\underline{I}}\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})} \cdot \frac{n_1 n_2}{n_1 + n_2}.$$

Durch Kürzen durch die Zahl $(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ ergibt sich

$$t^2_{max} = (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}) \cdot \frac{n_1 n_2}{n_1 + n_2}.$$

Bei $(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})'\underline{S}^{-1}(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ handelt es sich um die sog. quadrierte „Mahalanobis-Distanz“ zwischen den Gruppen D^2. D lässt eine Verallgemeinerung auf $m \geq 2$ Gruppen zu. Damit wird

$$t^2_{max} = \frac{n_1 n_2}{n_1 + n_2} D^2.$$

Die Prüfung der H_0 kann nun (ohne weiteren Beweis) erfolgen durch die Prüfgröße

$$F = \frac{n_1 + n_2 - m - 1}{m(n_1 + n_2 - 2)} \cdot t^2_{max}, \; df_1 = m, \; df_2 = n_1+n_2-m-1.$$

Damit sind beide Aufgaben gelöst.

Anhang

Prüfung der Gleichheit der Kovarianzmatrizen

Für die Prüfung der Gleichheit von Kovarianzmatrizen sind mehrere Verfahren bekannt. Ein Verfahren zur Prüfung der H_0: $\underline{\Sigma}^{(1)} = \underline{\Sigma}^{(2)}$ findet sich z.B. bei Hartung/Elpelt, Multivariate Statistik, 7. Auflage, Oldenbourg 2007:

Die Prüfung der H_0 erfolgt durch einen χ^2-Test mit der Prüfgröße

$$\chi^2 = c[(\sum_j n_j - k) \cdot \ln|\underline{S}| - \sum_j (n_j - 1) \cdot \ln|\underline{S}^{(j)}|],\ df = m(m+1)(k-1)/2, \text{ wobei}$$

$$c = 1 - \frac{2m^2 + 3m - 1}{6(m+1)(k-1)} (\sum_j \frac{1}{n_j - 1} - 1/(\sum_j (n_j - 1)))$$

k bezeichnet die Zahl der Gruppen, hier also k = 2.

Rechenbeispiel

Ausgangsdaten

Gruppe 1 („kreditwürdig“), fiktives Beispiel:

```
X1
    Alter Einkommen Kinder Autobesitz
x01    55      4503      0          1
x02    44      1651      1          1
x03    32      1518      2          0
x04    18      1253      1          0
x05    36      2471      0          1
x06    43      2126      0          1
x07    40      1272      3          0
x08    34      1393      2          1
x09    22      1626      1          0
x10    51      1941      0          1
```

Gruppe 2 („nicht kreditwürdig“), fiktives Beispiel:

X2

	Alter	Einkommen	Kinder	Autobesitz
x11	40	2869	3	0
x12	39	1431	3	1
x13	34	1567	3	0
x14	31	1323	5	0
x15	44	3384	4	1
x16	55	1432	1	0
x17	56	2161	4	0
x18	47	2476	3	1
x19	62	1896	2	1
x20	43	2121	1	0

Testvektor

	Alter	Einkommen	Kinder	Autobesitz
x	41	2500	1	1

Kovarianzmatrizen

S1

	Alter	Einkommen	Kinder	Autobesitz
Alter	136.944444	7115.8889	-4.3333333	4.2222222
Einkommen	7115.888889	938955.3778	-620.6666667	248.0666667
Kinder	-4.333333	-620.6667	1.1111111	-0.3333333
Autobesitz	4.222222	248.0667	-0.3333333	0.2666667

S2

	Alter	Einkommen	Kinder	Autobesitz
Alter	99.655556	765.8889	-5.65555556	1.28888889
Einkommen	765.888889	466357.1111	139.44444444	102.55555556
Kinder	-5.655556	139.4444	1.65555556	0.04444444
Autobesitz	1.288889	102.5556	0.04444444	0.26666667

Gesamtkovarianzmatrix

```
                Alter   Einkommen       Kinder  Autobesitz
Alter       118.300000   3940.8889   -4.9944444   2.7555556
Einkommen  3940.888889 702656.2444 -240.6111111 175.3111111
Kinder       -4.994444   -240.6111    1.3833333  -0.1444444
Autobesitz    2.755556    175.3111   -0.1444444   0.2666667
```

Prüfung auf Gleichheit der Kovarianzen

```
"Chi²-Test auf Varianzhomogenität"
"Chi²="
12.92608
"Freiheitsgrade df="
10
"Irrtumswahrscheinlichkeit"
0.2278344
```

→ *Die H_0 ist nicht widerlegt – gleiche Kovarianzen.*

Prüfung auf Gleichheit der Mittelwertsvektoren

```
"F-Test auf Mittelwertunterschiede"
"F =
12.36453
"Freiheitsgrade df1="
2
"Freiheitsgrade df2="
17
"Irrtumswahrscheinlichkeit"
0.0004842702
```

→ *Die H_0 ist widerlegt – es gibt unterschiedliche Mittelwertsvektoren.*

Ergebnis unter Anwendung der Diskriminanzfunktion

```
"Diskriminanzfunktion Y = (x1quer-x2quer)'S^(-1)X"
"(x1quer-x2quer)'S^(-1) ="
      Alter    Einkommen    Kinder Autobesitz
-0.1755371 -0.0002046068 -1.866666   1.687284
"Abstand Testvektor zu Gruppe 1: "
0.04680291
"Abstand Testvektor zu Gruppe 2: "
5.189939
```

→ *Damit ist der Testvektor der Gruppe 1 zuzuordnen (die Person ist „kreditwürdig“).*

R-Datenskript (Beispiel)

```
# Demo Matrix "Datensatz Diskriminanzanalsye"

# Eingabe der Datenfelder zunächst als Vektor,
# byrow = T: Zeilenweises einlesen!

# Alle Variablen löschen

rm(list=ls())


rnamen1 <- c("x01","x02","x03","x04","x05","x06","x07","x08","x09","x10")
rnamen2 <- c("x11","x12","x13","x14","x15","x16","x17","x18","x19","x20")
cnamen <- c("Alter","Einkommen","Kinder","Autobesitz")


X1 <- matrix(c(   55, 4503, 0, 1,
                  44, 1651, 1, 1,
                  32, 1518, 2, 0,
                  18, 1253, 1, 0,
                  36, 2471, 0, 1,
                  43, 2126, 0, 1,
                  40, 1272, 3, 0,
                  34, 1393, 2, 1,
                  22, 1626, 1, 0,
                  51, 1941, 0, 1
                  ),
            nrow=10,
            byrow=T,
            dimnames = list(rnamen1, cnamen)
            )

X2 <- matrix(c(   40, 2869, 3, 0,
                  39, 1431, 3, 1,
                  34, 1567, 3, 0,
                  31, 1323, 5, 0,
                  44, 3384, 4, 1,
                  55, 1432, 1, 0,
                  56, 2161, 4, 0,
                  47, 2476, 3, 1,
```

```
                  62, 1896, 2, 1,
                  43, 2121, 1, 0
                  ),
            nrow=10,
            byrow=T,
            dimnames = list(rnamen2, cnamen)
            )

# Zahl der Gruppen k

k <- 2

# Prüfvektor

x <- c(41, 2500, 1, 1)
```

R-Programmskript

```
# Programm zur Durchführung einer einfachen Diskriminanzanalyse

# V 0.2 vom 24./25.01.2010

# (c) Dr. Alexander Preuß

# Eine vollständige Datenmatrix X muss vorhanden sein!



# Bestimmung von n1, n2 und p (=Zahl der Merkmale)

n1 <- nrow(X1)

n2 <- nrow(X2)

m <- ncol(X1)

# Bestimmung der Mittelwertvektoren x1quer und x2quer

einsi1 <- c(rep(1,times=n1))
x1quer <- t(X1)%*%einsi1/n1

einsi2 <- c(rep(1,times=n2))
x2quer <- t(X2)%*%einsi2/n2

# Bestimmung der Kovarianzmatrizen S1 und S2

S1 <- cov(X1)
S2 <- cov(X2)

# Bildung der Gesamtkovarianzmatrix S

S <- (S1*(n1-1)+S2*(n2-1))/(n1+n2-2)

# Prüfung auf Varianzhomogenität der Kovarianzmatrizen

c <- 1-((2*m*m+3*m-1)/(6*(m+1)*(k-1)))*((((1/(n1-1))-1)/(n1-1+n2-
1))+(((1/(n2-1))-1)/(n1-1+n2-1)))
```

```
chi2 <- c*((n1+n2-2)*log(det(S))-(n1-1)*log(det(S1))-(n2-1)*log(det(S2)))

df_chi2 = m*(m+1)*(k-1)/2

print ("Chi²-Test auf Varianzhomogenität")
print ("Chi²=")
print (chi2)
print ("Freiheitsgrade df=")
print (df_chi2)
print ("Irrtumswahrscheinlichkeit")
print (1-pchisq(chi2,df_chi2))
print ("-----")


# Bestimmung der quadrierten Mahalanobisdistanz

D <- t((x1quer-x2quer))%*%solve(S)%*%(x1quer-x2quer)

# Bestimmung von t2max

t2max <- D*(n1*n2/(n1+n2))

# Bestimmung der Prüfgröße F

F <- (n1+n2-k-1)/(k*(n1+n2-2))*t2max
df_f1 <- k
df_f2 <- n1+n2-k-1

print ("F-Test auf Mittelwertunterschiede")
print ("F =")
print (F)
print ("Freiheitsgrade df1=")
print (df_f1)
print ("Freiheitsgrade df2=")
print (df_f2)
print ("Irrtumswahrscheinlichkeit")
print (1-pf(F,df_f1,df_f2))
print ("-----")

# Bestimmung der Entfernung des Prüfvektors zu den beiden Gruppen
```

```
d1 <- t((x1quer-x2quer))%*%solve(S)%*%x1quer
d2 <- t((x1quer-x2quer))%*%solve(S)%*%x2quer
d <- t((x1quer-x2quer))%*%solve(S)%*%x

a1 <- abs(d1-d)
a2 <- abs(d2-d)

print ("Diskriminanzfunktion Y = (x1quer-x2quer)'S^(-1)X")
print ("(x1quer-x2quer)'S^(-1) =")
print (t((x1quer-x2quer))%*%solve(S))
print ("-----")

print ("Abstand Testvektor zu Gruppe 1: ")
print (a1)
print ("Abstand Testvektor zu Gruppe 2: ")
print (a2)
```