Diskriminanz-Analyse

Prof. Dr. E. Reh, Technische Hochschule Bingen, Email: e.reh@th-bingen.de

Einführung

Die Diskriminanz-Analyse ist eine Methode zur
a) Untersuchung der Gruppenstruktur a priori vorgegebener Gruppen / Cluster deklarierter Objekte auf Grund ihrer diversen Merkmale und

b) Zuordnung neuer Proben (überwachtes Lernen [1] analog kNN- oder SIMCA-Methode, vgl. [4] Kap. 7.5.1, 7.5.2).
Der Unterschied zur Clusteranalyse besteht darin, dass hier auf bereits vorgegebene Cluster mit deklarierten Objekten zurückgegriffen wird.

Die Diskriminanz-Analyse ist in diversen Anwendungsbereichen weit verbreitet, z.B.
- Marketing-Analyse (Kunde/Nichtkunde abhängig von Kunden-Eigenschaften),
- Qualitätskontrolle (valides Produkt/Ausschuss im Hinblick auf diverse Kontroll-Kennwerte).
So ist die Aufgabe in der klinisch, chemischen Diagnostik zu ermitteln, wie die Proben von Gesunden und Kranken optimal differenziert werden können und ob die Laborparameter einer Patientenprobe diesen z.B. als Tumor-Patienten klassifizieren.
Dazu liegt bereits ein Basis-Datensatz von a priori zugeordneten Objekten vor mit ihren Merkmalen / Parametern, an Hand dessen eine neue Probe sicher zum Cluster der Tumor-erkrankten oder Gesunden zugeordnet werden soll. In manchen Fällen genügt der Einsatz eines Parameters (z.B. PSA-Wert bei Prostata-Tumorverdacht), in anderen Fällen müssen mehrere Parameter herangezogen werden (z.B. CEA, CA19-9 bei Lungen-Karzinom).

Darüber hinaus gibt es diverse Varianten der Diskriminanz-Analyse, z.B. die euklidische, lineare, quadratische, regularisierte, PLS-, Maximum-Likelihood-Diskriminanz-Analyse. Nur die wichtigsten sollen an dieser Stelle behandelt werden.

Primäres Ziel ist die Differenzierung jeweils zweier benachbarter Cluster und die Zuordnung neuer Objekte. Hierzu ist oft eine nominale Zuordnungsvariable c definiert: $c \in (+1; -1)$, d.h. c = +1, Objekt ist zu Cluster 1 zugeordnet (relevanter Cluster, z.B. pathologische Objekte).

Es stehen zwei Optionen für die Klassifizierung zur Verfügung:
a) Cluster-Distanz: Als Distanzmaß eines Objekts O zum Cluster 1 (d(O, C1)) bzw. zum Cluster 2 (d(O, C2)) wird oft der Euklidische- bzw. der Mahalanobis-Abstand des Objekts zum Zentroid des entsprechenden Clusters (vgl. [4], Kap. 7.4) verwendet.
Zur visuellen Differenzierung beider Cluster repräsentieren Punkte P_i mit gleichem Abstand zum Zentroid der beiden Cluster eine Linie („Grenzlinie“) bzw. Hyperfläche.
Im einfachen Fall (z.B. Euklidische Diskriminanz-Analyse mit 2 Merkmalen) ist dies eine Gerade wie z.B. in Abbildung 1a, allgemein wird der Klassifikator durch eine gekrümmte Hyperfläche im p-dimensionalen Raum (p: Merkmalszahl) repräsentiert.
Die Zuordnung eines Objekts O erfolgt zu dem Cluster mit der kleinsten Zentroid-Distanz. Dies ist in der Chemometrie der direkte und geläufige Ansatz zur Klassifizierung.

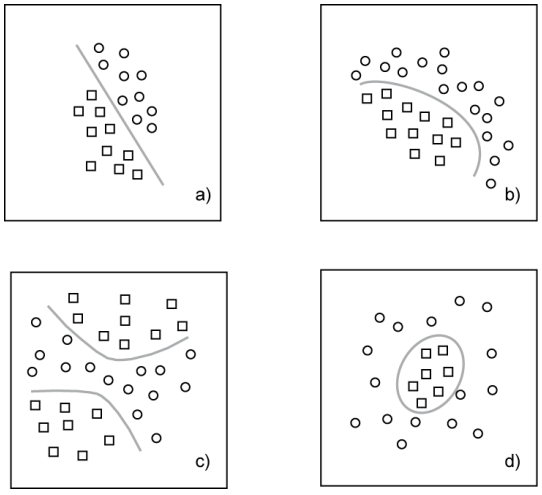


Abb. 1: Diverse 2-Cluster-Anordnungen

b) Diskriminanz-Funktion: Es wird eine Diskriminanz-Funktion verwendet, allgemein:
$y_d = b_1 x_1 + b_2 x_2 + \ldots b_p x_p$ (1).
Bei der numerischen Behandlung[1] erfolgt eine Schätzung der Diskriminanz-Gewichte b_j (Diskriminanz-Koeffizienten) ausgehend vom Merkmals-Vektor der Objekte **x** analog zur multivariaten Regression, dies ist bei vielen Merkmalen (p > 2) aufwändig.
Aus dem Mittelwert der Diskriminanz-Variablen der Objekte von Cluster 1 bzw. 2 wird ein Klassifikations-Wert y_k (kritischer Diskriminanzwert) für die Zuordnung abgeleitet zu:
$y_k = (n_1 \bar{y}_{d1} + n_2 \bar{y}_{d2}) / (n_1 + n_2)$ (2)
mit $\bar{y}_{d1}$, $\bar{y}_{d2}$: Mittelwert Diskriminanz-Variablen Objekte Cluster 1 bzw. 2; n_1, n_2: Objektzahl in Cluster 1 bzw. 2

Objekt O wird Cluster 1 zugeordnet bei $y_{dO} > y_k$ wenn gilt $\bar{y}_{d1} > y_k$, ansonsten zu Cluster 2, u.u..

Nicht immer ist es möglich, selbst durch einen komplexeren Klassifikator („Grenzlinie“) alle Objekte des einen von denen dem anderen Cluster zu differenzieren. Ziel der Diskriminanz-Analyse ist es, eine optimale Differenzierung zu erhalten, in solchen Fällen ist z.B. die prozentuale, korrekte Klassifizierung (Trefferquote, %CC) kleiner 100 %.
Hier werden basierend auf den Resultaten der Diskriminanz-Analyse alle Objekte des Basis-Datensatz klassifiziert (engl. autoprediction) und daraus die Zahl der korrekten Zuordnungen bzw. der prozentuale Anteil %CC bestimmt.

Auch für mehr als 2 Cluster findet die Diskriminanz-Analyse Anwendung. Bei m Clustern werden m Klassifikatoren (Hyperebenen) bzw. orthogonale Diskriminanz-Funktionen benötigt, wobei jede folgende einen maximalen Anteil derjenigen Streuung erklärt, die nach der Schätzung der vorigen Diskriminanz-Funktion als Rest verbleibt. Für diesen Mehr-Cluster-Fall gibt es jedoch Alternativen (z.B. kNN-, SIMCA-Methode, vgl. [4], Kap. 7.5).

Euklidische Diskriminanz-Analyse (EDA)
a) Als Maß für die Distanz eines Objekts O (mit Merkmalsvektor $\mathbf{x_o}$) z.B. zum betrachteten Cluster C1 (d.h. Abstand zum Zentroid des Clusters) dient die Euklidische Distanz (vgl. [4] Gleichung 7.20) in der Form

$$d(O, C1) = \sqrt{((\boldsymbol{x_O} - \boldsymbol{z_1})^T (\boldsymbol{x_O} - \boldsymbol{z_1}))} \quad (3)$$

mit $\mathbf{x_o}$: Merkmals-Spaltenvektor Objekt O; $\mathbf{z_1}$: Zentroid-Vektor Cluster 1 (vgl. [4] Gleichung 7.21)

Zur Zuordnung eines neuen Objekts dienen die Distanzen zu Cluster 1 bzw. 2, $\Delta d = d(O, C2) - d(O, C1)$, d.h. Objekt wird dem Cluster zugeordnet, zu dem es die kleinste Distanz aufweist.

Zuordnungs-Variable (Klassifikator) $c = \begin{cases} +1 & für\ \Delta d > 0 \\ -1 & für\ \Delta d < 0 \end{cases}$

Die Klassifikations-Gerade (Hyperebene) wird definiert durch Punkte gleichen Abstands zu beiden Zentroiden.

---

1 Die Diskriminanz-Funktion folgt allgemein aus der Linearkombination der Merkmale **x**:
$y_d = b_1 x_1 + b_2 x_2 + \ldots b_p x_p$ für i-tes Objekt (a)
mit y_d: Diskriminanz-Variable (metrisch); b_j: Diskriminanz-Gewicht; x_j: Merkmal j des i-ten Objekts

Diskriminanz-Gewicht b_j gibt die Bedeutung des Merkmals x_j für die Cluster-Differenzierung wieder (teilweise wird b_0 hinzu genommen zur Normierung für Gesamtmittel $\bar{y}_d = 0$).
Die Schätzung der Diskriminanz-Gewichte b_j erfolgt derart, dass Streuung der Diskriminanz-Variablen y_d innerhalb Cluster gering, zwischen den Cluster maximal ist, d.h.

Zielfunktion $z = \frac{Abweichung\ \bar{y}_d\ zwischen Clustern}{Abweichung\ y_{di}\ innerhalb Cluster} = \frac{\sum_{c=1}^{m} n_c (\bar{y}_{dc} - \bar{y}_d)^2}{\sum_{c=1}^{m} \sum_{i=1}^{n_c} (y_{di} - \bar{y}_{dc})^2}$ (b)

mit $\bar{y}_d$: Gesamtmittel Diskriminanz-Variable über alle Objekte; $\bar{y}_{dc}$: Mittelwert Diskriminanz-Variable in Cluster c; y_{di}: Diskriminanz-Variable i-tes Objekt; m: Clusterzahl; n_c: Objektzahl in Cluster c

Schätzung Diskriminanz-Gewichte b_j analog multivariater Regression mit Ziel maximaler Zielfunktion z:
- Einsetzen Gleichung a für y_{di} in Zielfunktion z
- Bestimmung Maxima durch Null setzen der partiellen Ableitung von z $\frac{\delta z}{\delta b_j} = 0$ (c) $\Rightarrow b_j$

Es wird vorausgesetzt eine symmetrische Streuung der Objekte um den Zentroid, eine vergleichbare Streuung bei beiden Clustern und eine multivariate Normalverteilung der Variablen x_j. Eine Autoskalierung bzw. Standardisierung der Rohdaten (Zentrierung und Skalierung auf einen vergleichbaren Wertebereich, vgl. [4] Gleichung 7.4) ist daher zwingend erforderlich.
Die Euklische Diskriminanz-Analyse sollte daher nur zum Einsatz kommen, wenn bei der Linearen Diskriminanz-Analyse oder Quadratischen Diskriminanz-Analyse die Berechnung der inversen Varianz-Kovarianz-Matrix nicht möglich ist.

b) Alternativ kann die Diskriminanz-Funktion verwendet werden.
Aus der Maximierung des Quotienten der maximalen Streuung der Diskriminanz-Variablen y_d zwischen den Clustern und der minimalen Streuung innerhalb der Cluster resultieren die beiden Diskriminanz-Gewichte bei 2 Merkmalen:

$$b_1 = \frac{(z_{12} - z_{22})\,SQ_{12} - (z_{11} - z_{21})\,SQ_{22}}{SQ_{12}^2 - SQ_{11}\,SQ_{22}} \quad (4) \qquad b_2 = \frac{(z_{11} - z_{21})\,SQ_{12} - (z_{12} - z_{22})\,SQ_{11}}{SQ_{12}^2 - SQ_{11}\,SQ_{22}} \quad (5)$$

mit z_{11}-z_{21}: Differenz 1. Merkmal der Zentroide Cluster 1 bzw. 2; z_{12}-z_{22}: Differenz 2. Merkmal der Zentroide Cluster 1 bzw. 2; SQ_{11}: Summe gruppeninterner Abweichungsquadrate Merkmal x_1; SQ_{22}: Summe gruppeninterner Abweichungsquadrate Merkmal x_2; SQ_{12}: Summe gruppeninterner Abweichungsprodukte Merkmal x_1, x_2

Die Summenquadrate SQ sind definiert zu

$$SQ_{11} = \sum_{c=1}^{2} \sum_{i=1, i \in n_c}^{n_c} (x_{i1} - z_{c1})^2 \qquad SQ_{22} = \sum_{c=1}^{2} \sum_{i=1, i \in n_c}^{n_c} (x_{i2} - z_{c2})^2 \qquad SQ_{12} = \sum_{c=1}^{2} \sum_{i=1, i \in n_c}^{n_c} (x_{i1} - z_{c1})(x_{i2} - z_{c2})$$

mit c: Index Cluster 1, 2; n_c: Objektzahl in Cluster c;
x_{i1}-z_{c1}: Differenz zwischen 1. Merkmal Objekt i (im Cluster c) und 1. Merkmal des Zentroid von Cluster c;
x_{i2}-z_{c2}: Differenz zwischen 2. Merkmal Objekt i (im Cluster c) und 2. Merkmal des Zentroid von Cluster c

Aus Gleichung 4 und 5 resultiert die Diskriminanz-Funktion $y_d = b_1 x_1 + b_2 x_2$ für alle Objekte und aus den Mittelwerten der Diskriminanz-Variablen y_d der Objekte von Cluster 1 bzw. 2 nach Gleichung 2 der Klassifikations-Wert y_k.
Die Auswahl-Regeln für die Klassifizierung eines Objekts O sind
Für $y_{dO} > y_k$ erfolgt Zuordnung zu Cluster 1 wenn gilt $\bar{y}_{d1} > y_k$, ansonsten zu Cluster 2, bzw.
Für $y_{dO} < y_k$ erfolgt Zuordnung zu Cluster 1 wenn gilt $\bar{y}_{d1} < y_k$, ansonsten zu Cluster 2.
Für die Zordnungs-Variable gilt: $c = \begin{cases} +1 & für (y_k - y_{dO})(y_k - \bar{y}_{d1}) > 0 \\ -1 & für (y_k - y_{dO})(y_k - \bar{y}_{d1}) < 0 \end{cases}$
Die Klassifikations-Gerade für 2 Merkmale resultiert aus Diskriminanz-Funktion und
Klassifikations-Wert $y_d = y_k = b_1 x_1 + b_2 x_2$ $\Rightarrow$ $x_2 = \frac{y_k}{b_2} - \frac{b_1}{b_2} x_1$ (6).

Lineare Diskriminanz-Analyse (LDA)
Die Lineare Diskriminanz-Analyse basiert auf den Arbeiten von R.A. Fisher [2], der als Distanzmaß eine modifizierte Mahalanobis-Distanz des Objekts O (Merkmalsvektor $\mathbf{x_O}$) zum Cluster 1 einführte, die sich von der Mahalanobis-Distanz (vgl. [4] Gleichung 7.21) ableitet:

$$d(O, C1) = \sqrt{((\boldsymbol{x_O} - \boldsymbol{z_1})^T \boldsymbol{C_{12}^{-1}} (\boldsymbol{x_O} - \boldsymbol{z_1}))} \quad (7)$$

mit $\boldsymbol{C_{12}}$: vereinte Varianz-Kovarianz-Matrix berechnet aus Varianz-Kovarianz-Matrices von Cluster 1 und 2

$$\boldsymbol{C_{12}} = \frac{(n_{C1} - 1)\boldsymbol{C_1} + (n_{C2} - 1)\boldsymbol{C_2}}{(n_{C1} + n_{C2} - 2)} \quad (8)$$

mit n_{C1}, n_{C2}: Objektzahl von Cluster 1 bzw. Cluster 2; $\boldsymbol{C_1}$, $\boldsymbol{C_2}$: Varianz-Kovarianz-Matrix Cluster 1 bzw. Cluster 2

Ein Objekt O wird entsprechend den Distanzen dem Cluster 1 (z.B. Tumorpatient) zugeordnet wenn gilt: d(O, C2) - d(O, C1) > 0, ansonsten Cluster 2 (gesunder Proband).
Zuordnungs-Variable $c = \begin{cases} +1 & für \Delta d > 0 \\ -1 & für \Delta d < 0 \end{cases}$
Die Klassifikations-Gerade (Hyperebene) wird ebenfalls definiert durch Punkte gleichen Abstands zu beiden Zentroiden.

Anders als bei Verwendung der euklidischen Distanz behandelt die Mahalanobis-Distanz auch Cluster mit asymmetrischer Verteilung der Objekte (resultierend aus Korrelationen zwischen Variablen). Auch gravierend unterschiedliche Rohdaten-Merkmale werden ähnlich der Autoskalierung korrigiert.
Voraussetzungen sind multivariate Normalverteilung der Variablen x_j und vergleichbare Varianzen der Variablen in den diversen Clustern. Zur Berechnung der Inversen von $\boldsymbol{C_{12}}$ muss die Zahl der Merkmale kleiner als die Zahl der Proben sein (p < n-1).

b) Alternativ kann nach Fisher folgende Diskriminanzfunktion verwendet werden:

$$y_{dO} = \boldsymbol{x_O}^T \boldsymbol{C_{12}}^{-1} (\boldsymbol{z_1} - \boldsymbol{z_2}) \quad (9)$$

mit $\boldsymbol{x_O}$: Merkmalsvektor Objekt O, $\boldsymbol{z_1}$: Zentroidvektor Cluster C1, $\boldsymbol{z_2}$: Zentroidvektor Cluster C2

Der Klassifikations-Wert y_k (Gleichung 2) resultiert aus den mittleren Diskriminanz-Variablen der Objekte in Cluster 1 bzw. 2 nach Gleichung 2. Zur Klassifikation von Objekt O gilt:
Für $y_{dO} > y_k$ erfolgt Zuordnung zu Cluster 1 wenn gilt $\bar{y}_{d1} > y_k$, ansonsten zu Cluster 2, u.u..

Für die Zordnungs-Variable gilt: $$c = \begin{cases} +1 & für (y_k - y_{dO})(y_k - \bar{y}_{d1}) > 0 \\ -1 & für (y_k - y_{dO})(y_k - \bar{y}_{d1}) < 0 \end{cases}$$

Eine Anwendung wird exemplarisch in Beispiel 1 exerziert.

Quadratische Diskriminanz-Analyse (QDA)

Es ist offensichtlich, das durch eine Gerade (Hyperebene) als Klassifikator in vielen Fällen keine optimale Differenzierung der Cluster erreicht werden kann (Abbildung 1b, c).
Bei der quadratischen Diskriminanz-Analyse wird die Zuordnung eines neuen Objekts zu den beiden Clustern 1 bzw. 2 getrennt betrachtet, Gleichung 7 ändert sich dadurch zu

$$d(O, C1) = \sqrt{((\boldsymbol{x_O} - \boldsymbol{z_1})^T \boldsymbol{C_1}^{-1} (\boldsymbol{x_O} - \boldsymbol{z_1}))} \quad (10)^2$$ analog für Cluster 2

Die Klassifikation erfolgt nach den Distanzen zu Cluster 1 bzw. 2, d.h. Objekt O wird dem Cluster zugeordnet, zu dem es die kleinste quadratische Mahalanobis-Distanz aufweist, d.h. $\Delta d^2 = d(O, C2)^2 - d(O, C1)^2$.

Zuordnungs-Variable: $$c = \begin{cases} +1 & für\ \Delta d^2 > 0 \\ -1 & für\ \Delta d^2 < 0 \end{cases}$$

Zur grafischen Differenzierung dient eine gekrümmte Kurve bzw. Hyperfläche im p-dimensionalen Raum, definiert durch Punkte, deren quadratische Mahalanobis-Distanz zu Cluster 1 und Cluster 2 gleich ist. Alle Diskriminanz-Analyse-Varianten erlauben keine gekrümmten Klassifikatoren und sind daher für anspruchsvollere Fälle nicht zu empfehlen.

Voraussetzung für die Quadratische Diskriminanz-Analyse ist eine multivariate Normalverteilung der Variablen x_j, zumeist wirkt sich dies jedoch erst bei starker Abweichung von der Normalverteilung auf die Klassifizierung aus.

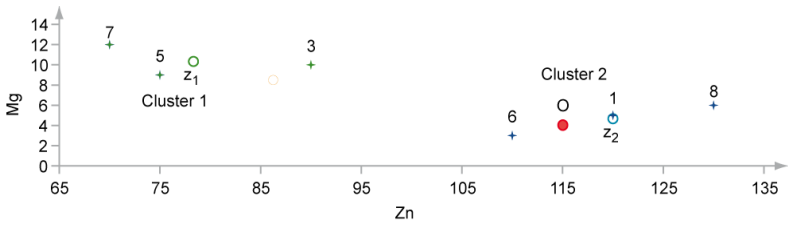
Die Quadratische Diskriminanz-Analyse behandelt erfolgreich auch Problemstellungen wie in Abbildung 1 b-d und ist für die meisten Aufgabenstellungen in der Chemie ausreichend, sie sollte daher vorrangig eingesetzt werden.

---

2 Gleichung 3, 7 bzw. 10 impliziert eine Gewichtung der Cluster nach der Häufigkeit / Wahrscheinlichkeit entsprechend der Zahl der zugehörigen Objekte. Dies ist in manchen Fällen nicht gerechtfertigt, z.B. einer sehr großen Anzahl Proben gesunder Probanden steht nur eine geringe Zahl pathologischer Proben gegenüber. Entsprechend dem Bayes Theorem der Wahrscheinlichkeitslehre kann dies bei der Definition der Mahalanobis-Distanzen durch die Integration einer a priori-Wahrscheinlichkeit p_C (für Cluster c) berücksichtigt werden.
Bei der Linearen Diskriminanz-Analyse gilt: $d(O, C1) = \sqrt{((\boldsymbol{x_O} - \boldsymbol{z_1})^T \boldsymbol{C_{12}}^{-1} (\boldsymbol{x_O} - \boldsymbol{z_1}) - 2\ln(p_1) + \ln(|\boldsymbol{C_{12}}|))}$.
$|\boldsymbol{C_{12}}|$ steht für die Determinante der vereinten Varianz-Kovarianz-Matrix $\boldsymbol{C_{12}}$, damit kann d(O, C1) nicht mehr als Distanz-Maß zu Cluster 1 interpretiert werden.
Für die Quadratische Diskriminanz-Analyse gilt: $d(O, C1) = \sqrt{((\boldsymbol{x_O} - \boldsymbol{z_1})^T \boldsymbol{C_1}^{-1} (\boldsymbol{x_O} - \boldsymbol{z_1}) - 2\ln(p_1) + \ln(|\boldsymbol{C_1}|))}$.
Im vorliegenden Fall würde die Zuweisung einer a priori Wahrscheinlichkeit $p_C = 0{,}5$ für beide Cluster eine gleiche Gewichtung beider Cluster bewirken.
Eine analoge Gewichtung ist bei der euklidischen Diskriminanz-Analyse nicht möglich.

Bsp. 1: Euklidische, Lineare Diskriminanz-Analyse zweier Saftsorten an Hand der Zn- und Mg-Konzentration

Objekt	Zn [µg/g]	Mg [mg/g]	Cluster
1	120	5	2
3	90	10	1
5	75	9	1
6	110	3	2
7	70	12	1
8	130	6	2



Tab. 1: Basisdaten (Zn/Mg-Merkmale)

Zentroid Cluster 1, $\boldsymbol{z_1}$

z_{11} (Zn)	78,33
z_{12} (Mg)	10,33

Zentroid Cluster 2, $\boldsymbol{z_2}$

z_{21} (Zn)	120,00
z_{22} (Mg)	4,67

$\boldsymbol{z_1 - z_2}$

$z_{11} - z_{21}$ (Zn)	-41,67
$z_{12} - z_{22}$ (Mg)	5,67

Cluster	x_1	x_2	$(x_{i1} - z_{c1})^2$	$(x_{i2} - z_{c2})^2$	$(x_{i1} - z_{c1})*(x_{i2} - z_{c2})$	y_d	
1	90	10	136,11	0,11	-3,89	-3,58	
1	75	9	11,11	1,78	4,44	-2,42	
1	70	12	69,44	2,78	-13,89	0,78	
						$\bar{y}_{d1}$	-1,74
2	120	5	0,00	0,11	0,00	-11,83	
2	110	3	100,00	2,78	16,67	-12,18	
2	130	6	100,00	1,78	13,33	-12,32	
						$\bar{y}_{d2}$	-12,11
			416,67	9,33	16,67		
			SQ_{11}	SQ_{22}	SQ_{12}		

1) Euklidische Diskriminanz-Analyse (mittels Diskriminanz-Funktion)

Berechnung Diskriminanz-Funktion (Gl. 4, 5) -Variablen y_d (Gl. 1) und Klassifikations-Wert y_k (Gl. 2)

b_1	b_2
-0,134	0,846

y_k
-6,93

Klassifikation Objekt O

x_{O1} (Zn)	115
x_{O2} (Mg)	4

Diskriminanz-Variable (Gl. 1)

Objekt O	y_{dO}
	-12,01

Diskriminanz-Variable $y_{dO} < y_k$ und $y_{d2} < y_k$, d.h. Objekt wird Cluster 2 zugewiesen.

2) Lineare Diskriminanz-Analyse (mittels Diskriminanz-Funktion)

Varianz-Covarianz-Matrix $\boldsymbol{C_1}$

	Zn	Mg
Zn	72,22	-4,44
Mg	-4,44	1,56

Varianz-Covarianz-Matrix $\boldsymbol{C_2}$

	Zn	Mg
Zn	66,67	10,00
Mg	10,00	1,56

vereinte Varianz-Covarianz-Matrix $\boldsymbol{C_{12}}$ (Gl. 6)

	Zn	Mg
Zn	69,44	2,78
Mg	2,78	1,56

inverse Varianz-Covarianz-Matrix $\boldsymbol{C_{12}^{-1}}$

	Zn	Mg
Zn	0,016	-0,028
Mg	-0,028	0,692

Berechnung Diskriminanz-Variablen y_d (Gl. 9), und Klassifikations-Wert y_k (Gl. 2),

Cluster	y_d	
1	-21,51	
1	-14,54	
1	4,71	
	$\bar{y}_{d1}$	-10,45
2	-70,98	
2	-73,11	
2	-73,94	
	$\bar{y}_{d2}$	-72,68

Objekt O	y_{dO}
	-72,05

y_k
-41,56

Diskriminanz-Variable $y_{dO} < y_k$ und $y_{d2} < y_k$, d.h. Objekt wird Cluster 2 zugewiesen.

Kritisch ist bei der Quadratische Diskriminanz-Analyse die numerische Inversion der Varianz-Kovarianz-Matrix. Ist diese nicht möglich, liegt dies nahezu immer an hohen Korrelationen zwischen den Merkmalen (Kollinearität).
Hier sollte vorrangig eine effektive Auswahl der Merkmale erfolgen (Monte-Carlo-Methoden, Bootstrapping), dies erlaubt meist auch eine wesentlich effizientere Klassifizierung.
Alternativ kann auf die Euklidische Diskriminanz-Analyse ausgewichen werden.

Biologische oder medizinische Fragestellungen sind oft komplexer, hier kann der Einsatz von alternativen Diskriminanz-Analysen wie learning vector quantization oder support vector machines[3] sinnvoll sein, sie tragen jedoch die Gefahr einer Über-Interpretation der Daten (engl. overfitting) in sich [3].

Generell behandelt die Diskriminanz-Analyse keine Ausreißer, ein Objekt wird immer dem einen oder anderen Cluster zugeordnet, eine Überlappung von Clustern ist nicht darstellbar (keine unscharfe Clusterung).
Zumeist sollte eine Validierung des Klassifikations-Modells erfolgen und auch die Bedeutung der diversen Merkmale bzw. von Hauptkomponenten für die Klassifizierung sollte untersucht werden [5].

Literatur

[1] Huberty, C.J. Applied Diskriminant Analysis, Wiley, Chinchester, 1994
[2] Fisher, R.A. Ann. Eugen., 7, 179-188, 1936
[3] Dixon, S.J., Brereton, R.G., Comparison of performance of five commen classifiers represented as boudary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines as dependant on data structure, Chemometrics Intell. Lab. Systems, 95, 1-17, 2009
[4] Reh, E., Chemometrie, Grundlagen der Statistik, Numerischen Mathematik und Software-Anwendung in der Chemie, de Gruyter, Berlin, 2017
[5] Brereton, R.G., Chemometrics for Pattern Recognition, Wiley-VCH, Weinheim, 2009

3 *Learning Vector Quantization*: Cluster sind durch eine Kombination von Geradenstücken (bzw. Hyperflächen) getrennt. Hierzu werden repräsentative Vektoren beschrieben („codebook vector“), mit dem Resultat, dass ein Cluster vom zweiten durch eine Serie linearer Klassifikatoren abgegrenzt wird (je mehr codebooks pro Cluster, desto komplexer die Trennlinie). Die Berechnung der Vektoren erfolgt iterativ unter Einsatz neuronaler Netze. Phänomenologisch kann dies mit der Berechnung eines äquidistanten Geradenstücks beschrieben werden für jeweils 2 benachbarte Objekte. Trennlinien zwischen zwei Objekten des gleichen Cluster entfallen, so dass final benachbarte Objekte unterschiedlicher Cluster an der Cluster-Grenze durch mehrere Geradenstücke getrennt sind. Entsprechend der primären Vorgabe resultieren, wie bei Einsatz neuronaler Netze häufig, jeweils unterschiedliche Ergebnisse.
Support Vector Machines: Es werden zur Differenzierung komplex gekrümmte Kurven berechnet. Die Berechnung basiert auf wenigen Objekten, die sich an der Grenze der Cluster befinden („support vector“), Objekte innerhalb der Cluster werden nicht berücksichtigt. Resultat ist eine beliebig gekrümmte Kurve (2 Cluster), wobei die Objekte des einen Cluster auf der einen, des zweiten Cluster auf der anderen Seite der Grenzlinie liegen. Die Optimierung des Klassifikators basiert bei vorgegebenen Randbedingungen (maximaler Abstand der support vectoren zur Grenzlinie) auf Verwendung der Lagrange Multiplikatoren. Dazu erfolgt eine Projektion der Objekte in einen höher dimensionalen Raum („feature space“; support vectoren sind die Objekte, die auf der hyperdimensionalen Ebene liegen). Aus der Rückdimensionierung resultiert ein komplexer, gekrümmter Klassifikator. Auf diese Weise können auch Kernel-Probleme (vgl. Abbildung 7.11d) adäquat gelöst werden. Kritisch ist die hohe Klassifikator-Komplexität auf Grund des hoch-dimensionalen Raums der Hyperebene und muss durch geeignete Maßnahmen kontrolliert werden.