Die praktische Anwendung der Diskriminanzanalyse zur Gruppierung im Data Mining

Wilfried Schollenberger

WS Unternehmensberatung und Controlling Systeme GmbH
69120 Heidelberg
Tel: 06221 / 401
E-Mail: wisch@ ws-unternehmensberatung.de
WEB: www.ws-unternehmensberatung.de

Abstract

1. Einführung

Der Anlass zu diesem Vortrag sind zwei Präsentationen auf der DISK 1999 in Ulm, wo die Mitarbeiter einer Bank vorstellten, wie sie versuchen ausfallgefährdete (Kredit-)Engagements zu erkennen. Grundlage sind Indikatorwerte, die 12 Monate vor der Wertberichtigung eines Engagements, bzw. bei nicht wertberichtigten Engagements 12 Monate vor der Analyse, erhoben wurden. Mit der Diskriminanzanalyse sollte eine optimale Formel gefunden werden, um ausfallgefährdete Engagements frühzeitig zu erkennen. Die vorgestellten "Probleme" und "Lösungen", vor allem aber das "Verfahren", die lineare Diskriminanzanalyse solange mit verschiedenen geschichteten Stichproben zu wiederholen, bis die Ergebnisse gemessen an einer Teststichprobe und der Gesamtheit eine zufriedenstellende Klassifikation erbringen (im Original: "Die Wahl einer geeigneten Stichprobe mit hohen Trefferquoten auf Lern- und Test-Stichprobe ist das A und O der Analyse"), überzeugten mich davon, dass selbst bei einem so alten Verfahren, eine strukturierte Einführung sinnvoll ist.

Neben einer grundsätzlichen Einführung in die "Paradigmen" der Datenanalyse, z.B. dass die Analyse einer Stichprobe niemals bessere Ergebnisse bringen kann als die der Gesamtheit, scheinen vor allem die Abweichungen von den Voraussetzungen, z.B. Verteilungsannahmen, Probleme zu bereiten. Es ist schwierig, aufgrund einer mathematischen Analyse vorherzusagen, wie sich konkrete Abweichungen, z.B. nicht

normalverteilte Indikatoren, auf die Güte des Ergebnisses auswirken. In der Literatur können immer nur konkrete Fälle behandelt werden, und eine vollständige Simulation ist sehr aufwendig.

In diesem Vortrag wird ein anschaulicher Ansatz versucht. Nach einem einfachen Beispiel mit drei Gruppen, wird an verschiedenen Spezialfällen gezeigt, wie sich Abweichungen von den Verteilungsannahmen auf die Güte des Ergebnisses auswirken, und wie das Ergebnis im Hinblick auf extern vorgegebene Optimalitätskriterien verbessert werden kann.

Dieser Beitrag soll auch dazu anregen, eine Methode an verschiedenen kontrollierten Beispielen zu testen, um so das Verhalten und mögliche Probleme kennen zu lernen. Eine besondere Bedeutung spielt dabei die graphische Kontrolle der Verteilungen und der Ergebnisse. Dabei bin ich mir durchaus bewusst, dass diese in gewisser Hinsicht "unwissenschaftlich" ist:

- Es wird auf ein vollständiges Verständnis der zugrundeliegenden Methoden und Algorithmen, sowie ihrer Herleitung verzichtet.
- Exemplarische Tests können vollständige Monte-Carlo-Simulationen, also die wiederholte Durchführung des Testprogramms und die simulierte Anwendung der Ergebnisse auf Bestände, die mit dem gleichen Algorithmus erzeugt wurden und sich nur zufällig unterscheiden, nicht ersetzen.

Trotzdem bin ich der Meinung, dass der hier vorgeschlagene Weg in der Praxis ein guter Mittelweg ist, um mit begrenztem Aufwand die Qualität der Anwendung von Methoden und der Durchführung von Analysen zu verbessern. Dabei konzentriere ich mich auf die Standard-Problemstellung im Data Mining, bei der eine große Datenmenge zur Verfügung steht, und die Inferenzstatistik (schließende Statistik) keine Bedeutung spielt (vgl. weiter unten die 4 Thesen). Dabei ist auch zu berücksichtigen, dass der praktische Vorteil, also die Optimierung des Verhältnisses von Kosten und Nutzen, oft bedeutsamer ist, als die exakte wissenschaftliche Erkenntnis um jeden Preis.

Eine Anmerkung zu "Lern-" und "Teststichproben": In der Literatur findet man häufig den Hinweis, dass die Daten in eine Lern- und eine Teststichprobe aufgeteilt werden können. Die Analyse wird mit der Lernstichprobe durchgeführt. Mit den Ergebnissen wird dann die Teststichprobe klassifiziert. Wenn dabei die Anteile der Fehlzuordnungen von denen der Analyse nicht abweichen ist dies ein Indikator für ein zuverlässiges Ergebnis. Grundsätzlich lohnt sich dieses Vorgehen bei der parametrischen (linearen und quadratischen) Diskriminanzanalyse nicht. Wenn die beiden Gruppen durch eine Zufallsauswahl getrennt werden, sind auch die Unterschiede in den Fehlzuordnungen nur zufällig, weil beide Stichproben der gleichen Grundgesamtheit mit denselben Parametern entstammen. Ein Maß für die Güte der Analyse ist der (gewichtete) Anteil der Fehlzuordnungen selbst. Deshalb wird in diesem Vortrag auf eine "Teststichprobe" verzichtet. Bei der nicht-parametrischen Diskriminanzanalyse (nearest neighbours) wird dagegen die graphische Analyse des Ergebnisses die Schwächen des Verfahrens aufzeigen.

Wenn in der Praxis des Data-Mining trotzdem Lern- und Teststichproben (oder Trainings- und Validierungs-Samples) erstellt werden, geschieht dies häufig, um die Angemessenheit eines Verfahrens ohne tiefere Einsicht zu beurteilen. Das ist durchaus sinnvoll, wenn der Versuch mehrfach mit dem identischen Verfahren durchgeführt wird. Dadurch erhält man auf empirischem Weg einen Aufschluss über die Angemessenheit des gewählten Verfahrens. Vollkommen falsch ist dagegen der Versuch, solange "Lernstichproben" zu ziehen, bis das Ergebnis, d.h. die Modellparamater oder der Entscheidungsbaum, zufällig auch bei der Teststichprobe zufriedenstellende Ergebnisse liefert.

2. Einführendes Beispiel mit drei Gruppen und erfüllten Annahmen

Für das erste Beispiel erzeugen wir eine Datei mit drei Gruppen und zwei Variablen. Die drei Gruppen unterscheiden sich nur in den Mittelwerten der beiden Variablen, x und y. Die Variablen sind in allen drei Gruppen normalverteilt und unkorreliert.

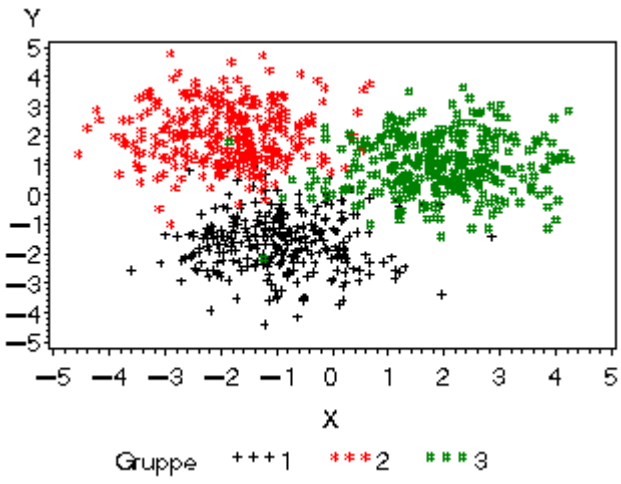
Erstellen der Datei:

```
%let grcount = 300; * Festlegung der Gruppengröße;

DATA data.bsp0;
  group = "1";
  do i = 1 to &grcount ;
     x = rannor(0) - 1;
     y = rannor(0) - 1.5;
     output;
  end;
  group = "2";
  do i = 1 to &grcount ;
     x = rannor(0) - 2;
     y = rannor(0) + 2;
     output;
  end;
  group = "3";
  do i = 1 to &grcount ;
     x = rannor(0) + 2;
     y = rannor(0) +1;
     output;
  end;
  stop;
RUN;
```

Die Verteilungsparameter der drei Gruppen, zeigen, dass eine Variable allein nicht ausreicht, um eine Beobachtung einer der drei Gruppen zuzuordnen. Der Plot zeigt aber, dass die Kombination der beiden Variablen sehr wohl geeignet ist, die Gruppenzugehörigkeit einer Beobachtung zu schätzen. Die meisten Beobachtungen jeder Gruppe befinden sich in einem abgegrenzten Teil der durch x und y aufgespannten Ebene.

Name	Gruppe	Mittel-wert	Stdev	Min	Max
X	1	-0.93	0.99	-3.61	2.86
	2	-1.88	0.97	-4.56	0.66
	3	1.98	1.07	-1.85	4.27
Y	1	-1.61	0.94	-4.40	0.94
	2	2.07	0.98	-1.00	4.83
	3	1.02	0.97	-2.14	3.66



Aufgrund der gleichen Fallzahl und der gleichen Varianz aller Variablen in allen Gruppen, wäre eine geometrische Lösung des Problems sehr einfach: Man nehme die Mittelpunkte (x,y) jeder Gruppe und bestimme die drei Mittelsenkrechten. Das Ergebnis wäre in diesem Fall identisch mit dem der folgenden Diskriminanzanalyse. Allerdings verwenden wir im folgenden die Diskriminanzfunktionen der Prozedur DISCRIM. Mit ihr lassen sich für jede Kombination der Variablen (x und y) Score-Werte für die Zugehörigkeit zu den Gruppen bestimmen. Die Beobachtung wird der Gruppe zugeordnet, für die der höchste Score-Wert ermittelt wurde. Durch gleichsetzen von jeweils zwei Diskriminanzfunktionen und umformen erhält man die Funktionen für die oben angesprochenen Mittelsenkrechten.

Das einzig Erklärungsbedürftige an dem folgenden Prozeduraufruf ist die OUTSTAT=-Option, mit der wir eine Kalibrierungsdatei erzeugen, die wir anschließend als Eingabedatei zum Gruppieren von weiteren Dateien verwenden können.

Prozeduraufruf:

```
PROC Discrim
     data = data.bsp0   /* Eingabedatei */
     outstat = work.ergBsp0 /* Kalibrierungsdaten */
  ;
  var x y;
  class group;
RUN;
```

Als Ergebnis erhalten wir (unter anderem) die Diskrimanzfunktionen und eine Zuordnungsmatrix, aus der wir ablesen können, wie gut die Gruppen durch diese Funktionen reproduziert werden können.

	Linear Discriminant Function for GROUP		
Variable	1	2	3
Constant	-1.77060	-4.19353	-2.41198
X	-0.85279	-1.91946	1.90192
Y	-1.70518	2.30936	1.03173

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	3	Total
1	285	6	9	300
	95.00	2.00	3.00	100.00
2	5	290	5	300
	1.67	96.67	1.67	100.00
3	10	6	284	300
	3.33	2.00	94.67	100.00
Total	300	302	298	900
	33.33	33.56	33.11	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for GROUP

	1	2	3	Total
Rate	0.0500	0.0333	0.0533	0.0456
Priors	0.3333	0.3333	0.3333	

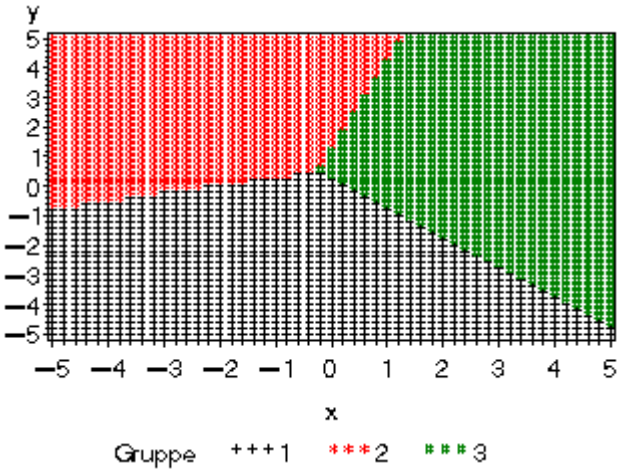
In unserem Beispiel sehen wir, dass rund 5 % der Beobachtungen falsch zugeordnet werden. Am anschaulichsten lässt sich das Ergebnis darstellen, wenn man eine Datei mit gleichverteilten Beobachtungen gruppiert und dieses Ergebnis graphisch ausgibt. Dabei wird die Kalibrierungsdatei zur Eingabedatei. Die zu gruppierende Datei wird mit der TESTDATA=-Option und die Ergebnisdatei mit der TESTOUT=-Option angegeben.

Datei erzeugen:	**DATA work.bsp0t;** **do x = -5 to 5 by 0.2;** **do y = -5 to 5 by 0.2;** **output;** **end;** **end;** **stop;** **RUN;**
Prozeduraufruf:	**PROC Discrim** **data = work.ergBsp0 /* Kalibrierungsdaten */** **testdata=work.bsp0t /* zu klassifizieren */** **testout=work.bsp0terg (rename=_into_=group);** **;** **var x y;** **class group;** **RUN;**

Die Ergebnisdatei enthält über das Gruppierungsergebnis hinaus noch eine Schätzung der Wahrscheinlichkeiten für die Zugehörigkeit der Beobachtung zu den drei Gruppen. Unter der Annahme, dass die Beobachtungen **in den Gruppen** multivariat normalverteilt sind, kann für jede Gruppe die "Dichte" für diese Merkmalskombination geschätzt werden. Nach Normierung auf die Summe 1 erhält man die relative Wahrscheinlichkeit, dass die Beobachtung zu der jeweiligen Gruppe gehört.

x:	-5
y:	-5
_1:	0.9999999648
_2:	3.5197486E-8
_3:	6.259404E-13
group:	1

Im Plot erhält man die drei Felder mit den geraden Begrenzungen.



Das Ergebnis ist insofern "optimal", als die Zahl der Fehlzuordnungen insgesamt minimiert wird. Die Fehlzuordnungen resultieren daraus, dass im gleichen Wertebereich Beobachtungen aus zwei Gruppen vorkommen. Deshalb kann das Ergebnis für eine Gruppe nur verbessert werden, wenn gleichzeitig mehr Fehlzuordnungen aus einer anderen Gruppe in Kauf genommen werden.

Eine solche "Manipulation" der Analyse kann durchaus sinnvoll sein, wenn z.B. eine Fehlzuordnung von Beobachtungen aus der Gruppe 3 zu einer anderen Gruppe mit besonders großen Nachteilen ("Kosten") verbunden wäre. In die Diskriminanzanlyse lässt sich dieser Aspekt mit den a-priori-Wahrscheinlichkeiten einführen. Die a-priori-Wahrscheinlichkeiten für alle Gruppen müssen sich auf 1 summieren. Wenn nichts anderes angegeben wird, arbeitet die Prozedur DISCRIM mit gleichen Wahrscheinlichkeiten für alle Gruppen, also 1 / (Anzahl der Gruppen). Das entspricht der Anweisung PRIORS EQUAL. Mit PRIORS PROP wird die Prozedur angewiesen, die a-priori-Wahrscheinlichkeiten proportional zur Gruppengröße zu setzen, also (Anzahl Beobachtungen in der Gruppe) / (Anzahl Beobachtungen in der Datei). Andere a-priori-Wahrscheinlichkeiten können manuell angegeben werden, dabei muß für jede Gruppe der formatierte Wert der Class-Variablen in Hochkomma angegeben werden.

Im folgenden Beispiel wählen wir für die dritte Gruppe eine a-priori-Wahrscheinlichkeit von 95 %, um möglichst alle Beobachtungen dieser Gruppe korrekt zuzuordnen.

Prozeduraufruf:	**PROC Discrim** **data = data.bsp0 /* Eingabedatei */** **outstat = work.ergBsp0B /* Kalibrierungsdaten */** **;** **var x y;** **class group;** **priors "1"=0.025 "2"=0.025 "3"=0.95; /* EQUAL, PROP oder Werte */** **RUN;**

Die Diskriminanzfunktionen unterscheiden sich von denen der ersten Analyse nur durch die deutlich kleineren Konstanten in den Diskriminanzfunktionen für die erste und die zweite Gruppe. Bis auf zwei Ausreißer werden jetzt alle Beobachtungen aus der dritten Gruppe korrekt zugeordnet. Allerdings werden jetzt auch rund 12 % aus den anderen beiden Gruppen fälschlicherweise der dritten Gruppe zugeordnet.

erste Analyse

Variable	Linear Discriminant Function for GROUP 1	2	3
Constant	-1.77060	-4.19353	-2.41198
X	-0.85279	-1.91946	1.90192
Y	-1.70518	2.30936	1.03173

mit "optimiertem" PRIORS-Statement

Variable	Linear Discriminant Function for GROUP 1	2	3
Constant	-5.45948	-7.88241	-2.46327
X	-0.85279	-1.91946	1.90192
Y	-1.70518	2.30936	1.03173

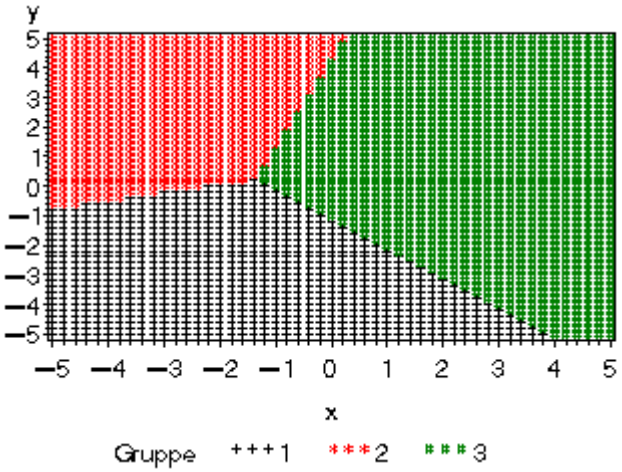
Number of Observations and Percent Classified into GROUP

From GROUP	1	2	3	Total
1	260	6	34	300
	86.67	2.00	11.33	100.00
2	5	258	37	300
	1.67	86.00	12.33	100.00
3	1	1	298	300
	0.33	0.33	99.33	100.00
Total	266	265	369	900
	29.56	29.44	41.00	100.00
Priors	0.025	0.025	0.95	

Error Count Estimates for GROUP

	1	2	3	Total
Rate	0.1333	0.1400	0.0067	0.0132
Priors	0.0250	0.0250	0.9500	

Im Plot zeigt sich, dass die Fläche der dritten Gruppe jetzt wesentlich größer ist.



3. Verletzte Annahmen

Dieses erste Beispiel ist vor allem deshalb sehr einfach, weil alle Annahmen für die einfache (d.h. lineare) Diskriminanzanlyse erfüllt sind. Die Variablen sind innerhalb der Gruppen multivariat normalverteilt und unkorreliert (obwohl nur die Homogenität der Covarianzmatrizen, also gleiche Korrelationskoeffizienten gefordert werden - aber dann hätte das geometrische Beispiel nicht so einfach funktioniert).

In der Praxis sind diese Annahmen oft nicht erfüllt:

- Die Varianzen in den Gruppen können sich unterscheiden.
- Die Fallzahlen in den Gruppen sind verschieden.
- Die Variablen in den verschiedenen Gruppen sind unterschiedlich korreliert.
- Die Verteilungen in den Gruppen entsprechen nicht der multivariaten Normalverteilung und können sich auch zwischen den Gruppen unterscheiden.

Insbesondere in der Einführungsliteratur wird nicht immer genau erklärt, welche Vorraussetzungen für welche Methoden gelten. Z.B. wird pauschal "multivariate Normalverteilung" verlangt, ohne explizit darauf hinzuweisen, dass diese Annahme für die Gruppen gilt, während die Verteilung der Gesamtheit z.B. im Zwei-Gruppen-Fall durchaus bimodal sein sollte. In der Praxis habe ich bisher nur wenige Menschen gefunden, die in der Lage waren, genau zu erklären, an welcher Stelle der Herleitung eines Verfahrens welche Annahmen eingehen, und wie dies zu begründen ist.

Die meisten der oben angeführten Probleme lassen sich durch explizit gesetzte a-priori-Wahrscheinlichkeiten und die Verwendung von quadratischen Diskriminanzfunktionen lösen, ohne dass die Voraussetzungen des Verfahrens verletzt werden. Die realen Auswirkungen von verletzten Verteilungsannahmen auf die Güte des Ergebnisses sind dagegen schwer abzuschätzen.

4. Thesen zu verletzten Verteilungsannahmen

Häufig wird in der Literatur empfohlen, die Verteilungen durch geeignete nichtlineare Transformationen an die Annahmen der Methode anzupassen. Abgesehen davon, dass dies in der Praxis nur selten vollständig möglich ist, halte ich dies im Data-Mining nur in besonders begründendeten Fällen für sinnvoll:

- Verteilungsannahmen haben eine große Bedeutung für die Inferenzstatistik (schließende Statistik). Bei den (oft) großen Fallzahlen, die im Data-Mining zur Verfügung stehen, spielt aber der inferenzstatistische Schluß auf eine

Grundgesamtheit meistens keine Rolle, weil nicht-signifikante Effekte auch deskriptiv unbedeutend sind und umgekehrt. Darüber hinaus wäre die Annahme zu prüfen, ob die vorhandenen Daten als Zufallsstichprobe einer (welcher) Grundgesamtheit aufgefasst werden können.

- Wenn es sich nicht um Skalenfehler, z.B. db(A) statt Sone für Lärm oder Vermögen in EURO als Indikator für Wohlstand, handelt, ist es meistens unmöglich eine Transformation zu finden, die für alle Gruppen der Diskriminanzanalyse eine ausreichende Verbesserung bringt.
- Die Methode sollte zu den Daten passen. Eine Anpassung der Daten an die Anforderungen der Methode ist immer der schlechtere Weg.
- Viele Verfahren sind zwar unter bestimmten Annahmen entwickelt worden, aber in der Praxis robust gegen bestimmte Abweichungen von diesen Annahmen. Die Auswirkungen dieser Abweichungen können an künstlich und kontrolliert erzeugten Daten geprüft werden.

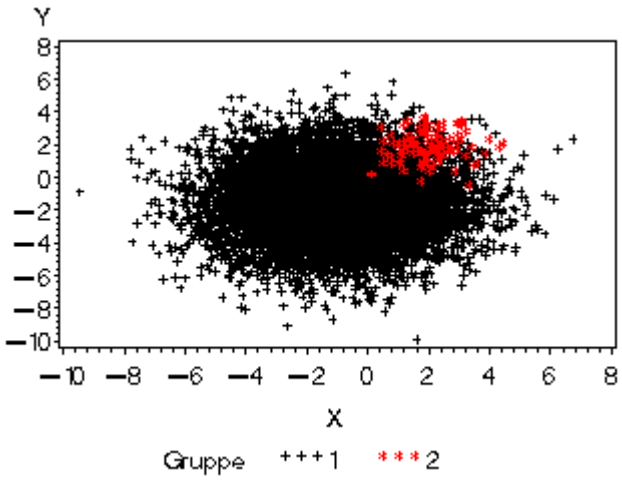
In den folgenden Beispielen wird zunächst gezeigt, wie sich ungleiche Gruppengrößen, verschiedene Varianzen und korrelierte Indikatoren in der Analyse berücksichtigen und behandeln lassen. Im Anschluss wird am Beispiel von in den jeweiligen Intervallen gleich-verteilten Gruppen gezeigt, wie sich diese Verletzung der Verteilungsannahme auswirkt, und diskutiert, wann eine Transformation oder eine nichtparametrische Methode (nearest neighbour) sinnvoll ist.

5. Beispiel mit ungleichen Gruppengrößen und verschiedenen Varianzen

Im ersten Beispiel sind gleich mehrere Probleme kombiniert. Ziel ist die Identifikation einer kleinen Gruppe in einer großen Gesamtheit. Die kleine Gruppe hat auch eine kleinere Varianz in den Variablen und befindet sich im Streuungsbereich der großen Gruppe.

Erstellen der Datei:

```
DATA data.bsp1;
  group = "1";
  do i = 1 to 10000;
     x = rannor(0) * 2 - 1;
     y = rannor(0) * 2 - 1.5;
     output;
  end;
  group = "2";
  do i = 1 to 100 ;
     x = rannor(0) + 2;
     y = rannor(0) + 2;
     output;
  end;
  stop;
RUN;
```



Verteilungsparameter

Name	Gruppe	Mittelwert	Stdev	Min	Max
X	1	-1.02	1.99	-9.45	6.73
	2	2.06	0.88	0.15	4.45
Y	1	-1.48	1.99	-9.86	6.32
	2	1.95	0.89	-0.45	3.71

In dieser Situation ist es offensichtlich, dass die kleine (interessante) Gruppe nicht eindeutig identifiziert werden kann. Mit der Diskriminanzanalyse kann aber eine Vorauswahl getroffen werden, in der sich (fast) alle Beobachtungen der kleinen Gruppe befinden und die dann weiter untersucht werden muss.

Im ersten Versuch führen wir eine lineare Diskriminanzanalyse durch. Die zusätzliche Option POOL=YES entspricht der Voreinstellung und soll hier nur auf den Fehler hinweisen, denn tatsächlich sind die Kovarianzmatrizen der beiden Gruppen nicht gleich. PRIORS EQUAL ist ebenfalls Voreinstellung und führt dazu, dass der Anteil (!) der Fehlzuordnungen aus der ersten Gruppe genauso gewichtet wird, wie derjenige aus der zweiten Gruppe. In unserem Fall (10000 zu 100 Beobachtungen) wird also eine Fehlzuordnung aus der zweiten Gruppe wie 100 Fehlzuordnungen aus der ersten Gruppe gewertet.

Prozeduraufruf:

```
PROC Discrim
      data = data.bsp1          /* Eingabedatei */
      outstat = wok.ergBsp1 /* Kalibrierungsdaten */
     pool = YES          /* = Voreinstellung */
  ;
  var x y;
  class group;
  priors EQUAL; /* = Voreinstellung */
RUN;
```

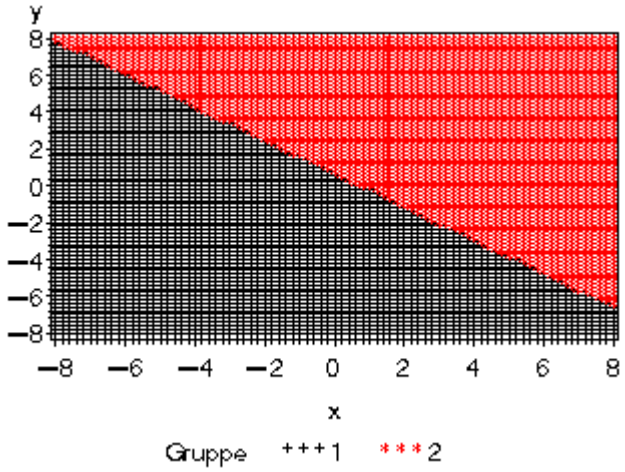
Im Ergebnis erhalten wir trotzdem mit über 12 % einen relativ hohen Anteil an Fehlzuordnungen aus der ersten Gruppe. Würden wir dagegen die a-priori-Wahrscheinlichkeiten proportional zu den Häufigkeiten in den Gruppen setzen, würden wir gerade mal 5 Beobachten aus der zweiten Gruppe richtig zuordnen.

Ergebnis der linearen Diskriminanzanalyse mit PRIORS EQUAL

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	Total
1	8768	1232	10000
	87.68	12.32	100.00
2	1	99	100
	1.00	99.00	100.00
Total	8769	1331	10100
	86.82	13.18	100.00
Priors	0.5	0.5	

Der praktische Vorteil mag jetzt schon sehr groß sein, wenn statt 10.000 Fällen nur rund 1/8 genauer untersucht werden muss. Der Plot des Ergebnisses zeigt aber die Ursache für den relativ großen Fehler: Während sich in den Ausgangsdaten die Beobachtungen der zweiten Gruppe in einem relativ kleinen kreisförmigen Segment befinden, wird bei der linearen Diskriminanzanalyse ein gerader Abschnitt der zweiten Gruppe zugeordnet.



Ursache ist die Annahme der gemeinsamen Kovarianzmatrix, d.h. hier der gleichen Varianzen. Sie führt dazu, dass die beiden Gruppen nur durch eine Gerade getrennt werden können. Wenn wir diese Annahme mit der Option POOL=NO aufgeben, erhalten wir eine "quadratische" Diskriminanzfunktion, in die auch die Quadrate der Variablen und die Produkte eingehen.

Prozeduraufruf:	**PROC Discrim** **data = data.bsp1 /* Eingabedatei */** **outstat = work.ergBsp1 /* Kalibrierungsdaten */** **method = normal /* = Voreinstellung */** **pool = NO /* = quadratische(?) Analyse */** **;** **var x y;** **class group;** **priors EQUAL; /* = Voreinstellung */** **RUN;**

Die Parameter der Diskriminanzfunktonen werden dann nur noch in der Kalibrierungsdatei ausgegeben. Im Ergebnis werden dann mit 562 gegenüber 1232 weniger als die Hälfte der Beobachtungen falsch zugeordnet. (Dass viele Beobachtungen falsch zugeordnet würden, war ja aufgrund der Ausgangslage klar.)

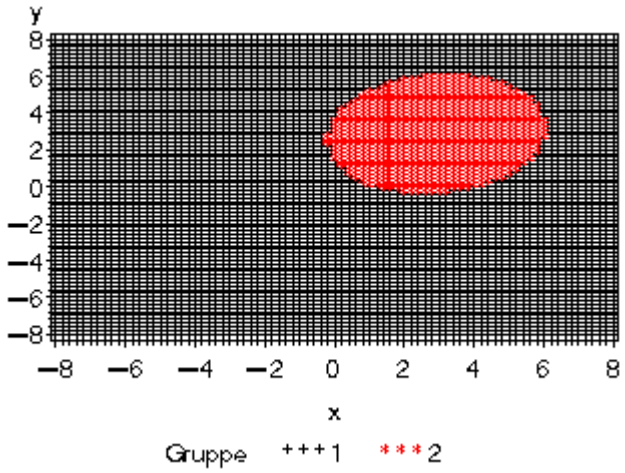
Kalibrierungsdatei

Gruppe	_TYPE_	_NAME_	X	Y
1	QUAD	X	-0.12676	-0.00004
1	QUAD	Y	-0.00004	-0.12603
1	QUAD	_LINEAR_	-0.25894	-0.37384
1	QUAD	_CONST_	-1.78456	-1.78456
2	QUAD	X	-0.64761	0.05327
2	QUAD	Y	0.05327	-0.63017
2	QUAD	_LINEAR_	2.46417	2.24229
2	QUAD	_CONST_	-4.49066	-4.49066

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	Total
1	9438	562	10000
	94.38	5.62	100.00
2	3	97	100
	3.00	97.00	100.00
Total	9441	659	10100
	93.48	6.52	100.00
Priors	0.5	0.5	

Der Plot zeigt, dass jetzt ein kreis- (bzw. ellipsen-) förmiges Segment des Wertebereichs der zweiten Gruppe zugeordnet wird, was aus der Ausgangslage resultiert, dass die zweite Gruppe bei beiden Variablen eine kleinere Streuung aufweist. Deshalb ist bei sehr großen Werten für x und y wieder eine Herkunft aus der ersten Gruppe zu erwarten.



6. Beispiel mit unterschiedlich korrelierten Indikatoren in den Gruppen

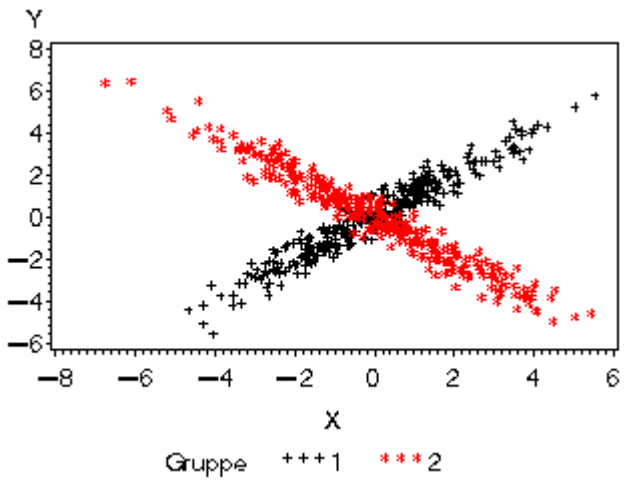
Das folgende Beispiel zeigt, dass korrelierte Indikatoren in der Diskriminanzanalyse keineswegs nachteilig sind, wie manchmal angenommen wird. Die Annahmen der linearen Diskriminanzanlyse verlangen eine gemeinsame Kovarianzmatrix, d.h. die Korrelation zwischen den Indikatoren soll in allen Gruppen gleich sein. Unterschiede in der Korrelationsmatrix können dagegen in der quadratischen Diskriminanzanalyse genutzt werden.

In diesem Beispiel stammen die beiden Gruppen aus Grundgesamtheiten mit gleichen Mittelwerten für beide Indikatoren. Allerdings unterscheidet sich der Korrelationskoeffizient der beiden Indikatoren im Vorzeichen. D.h trotz der gleichen Mittelwerte, liegt ein großer Teil der Beobachtungen in verschiedenen Wertebereichen. Klar ist allerdings, dass im "Kreuzungsbereich" die Beobachtungen nicht "richtig" zugeordnet werden können.

Erstellen der Datei:

```
%let grcount = 300;

DATA data.bsp2;
 group = "1";
 do i = 1 to &grcount;
   x = rannor(0) * 2 ;
   y = rannor(0) / 2 + x ;
   output;
 end;
 group = "2";
 do i = 1 to &grcount ;
   x = rannor(0) * 2;
   y = rannor(0)  / 2 - x ;
   output;
 end;
 stop;
RUN;
```



Im Ergebnis sehen wir, dass sich die Diskriminanzfunktion (erwartungsgemäß) praktisch nur im Parameter für das Produkt aus den Indikatoren unterscheidet. Im "Kreuzungsbereich" der Verteilungen werden insgesamt 45 Beobachtungen, also rund 8%, fehlklassifiziert. Natürlich kann man auch hier mit a-priori-Wahrscheinlichkeiten die bevorzugte Zuordnung zu einer Gruppe erzwingen. Z.B. erhält man mit PRIORS "1" = 0.8 "2" = 0.2 insgesamt 62 Fehlzuordnungen, von denen aber nur noch 3, d.h. 1%, aus der ersten Gruppe stammen.

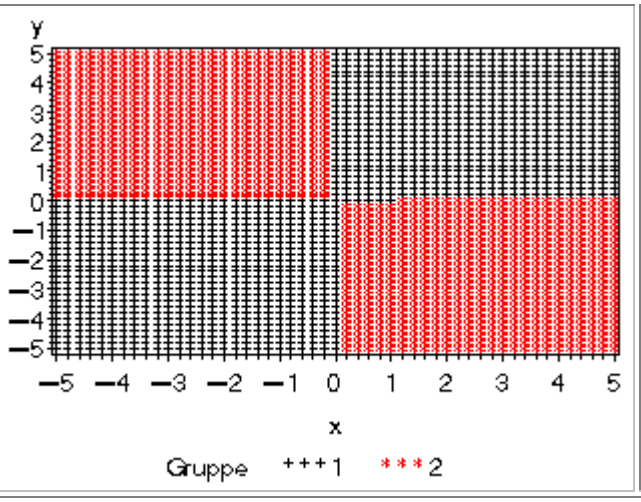
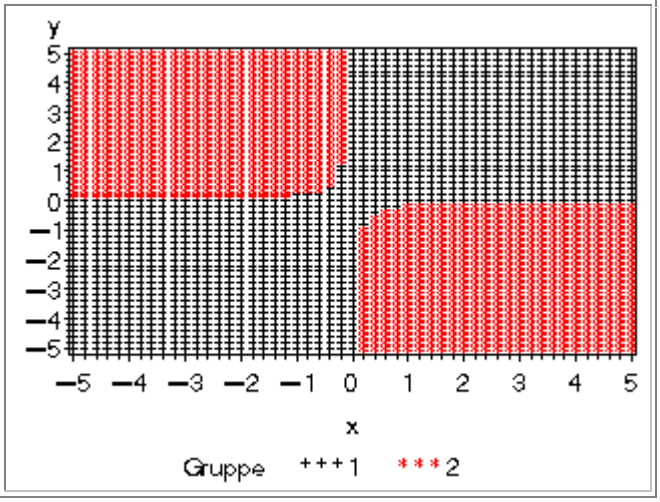
Kalibrierungsdatei (Auszug)

Gruppe	_TYPE_	_NAME_	X	Y
1	QUAD	X	-1.97830	1.84272
1	QUAD	Y	1.84272	-1.84988
1	QUAD	_LINEAR_	-0.11951	0.12543
1	QUAD	_CONST_	0.02507	0.02507
2	QUAD	X	-1.98979	-1.92558
2	QUAD	Y	-1.92558	-1.96948
2	QUAD	_LINEAR_	-0.00749	-0.01754
2	QUAD	_CONST_	-0.08507	-0.08507

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	Total
1	280	20	300
	93.33	6.67	100.00
2	25	275	300
	8.33	91.67	100.00
Total	305	295	600
	50.83	49.17	100.00

Im Ergebnisplot sehen wir, dass der Bereich in vier Rechtecke eingeteilt wurde. Durch die Priorisierung der ersten Gruppe entsteht eine breitere Verbindung zwischen den beiden Feldern dieser Gruppe.

Plot mit PRIORS EQUAL	Plot mit PRIORS "1" = 0.8 "2" = 0.2
	

6.1 Andere Verteilungen

Bisher waren die Indikatoren in allen Gruppen bis auf zufällige Abweichungen multivariat normalverteilt. Angesichts der nahezu unendlichen Zahl von vorstellbaren Abweichungen, möchte ich mich hier auf eine extreme Situation konzentrieren, und an ihr zeigen, wie die Auswirkungen von verletzten Verteilungsannahmen und der zu erwartende Nutzen einer nicht-linearen Transformation der Daten abgeschätzt werden kann.

In den folgenden Beispielen konzentriere ich mich deshalb auf die Frage:

Lässt sich eine gegebene Diskriminanzfunktion bei geichverteilten Variablen reproduzieren ?

Dazu wird eine Datei mit gleichverteilten Variablen erzeugt und mit den oben gewonnenen Diskriminanzfunktionen gruppiert. D.h. wir erhalten drei bzw. zwei klar voneinander getrennte

Gruppen, für die wir die Diskriminanzfunktion kennen. Danach werden die Daten in einem Fall transformiert. Der Plot entspricht den oben gezeigten Ergebnisplots. Nach der Diskriminanzanalyse wird die Ergebnisdatei der zweiten Gruppierung so aufbereitet, dass die Fehlzuordnungen sichtbar werden. Die Fehlzuordnungen werden in der Form "Evn" angezeigt, wobei v für die Gruppe der Eingabedatei und n für die Gruppe der irrtümlichen Zuordnung steht. "E12" steht als für die Beobachtungen aus der Gruppe 1, die irrtümlich der Gruppe 2 zugeordnet wurden.

Mit einer nichtparametrischen Analyse nach der nearest-neighbour-Methode mit jeweils 6 neighbours wird ggf. kontrolliert, ob diese Methode bessere Ergebnisse bringen würde. (Die nearest-neighbour-Methode ist die einzige nicht-parametrische Alternative der Prozedur Discrim, die sich auch mit großen Fallzahlen performant durchführen lässt.)

Grundsätzlich muss berücksichtigt werden, dass die Ergebnisse der parametrischen Analyse von den Mittelwerten und Streuungen der Indikatoren in den einzelnen Gruppen abhängen. Deshalb führen andere Wertebereiche für die Indikatoren in der Ausgangsdatei auch zu anderen Ergebnissen im Bezug auf die Güte der Analyse.

6.2 3 Gruppen mit linearen Grenzen

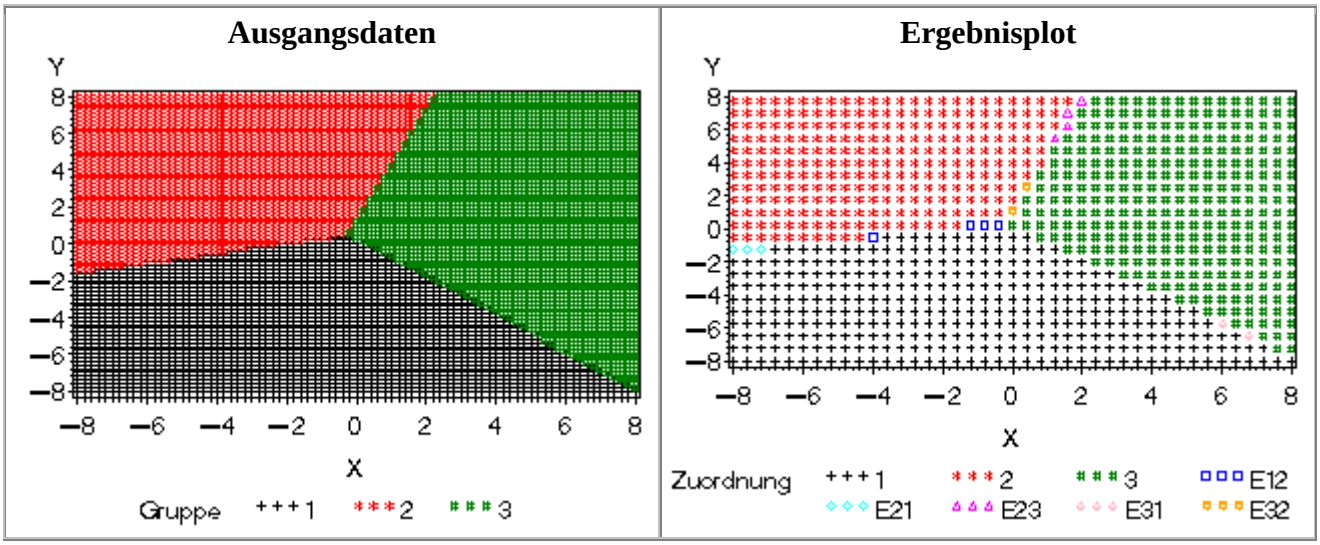
In diesem Beispiel hält sich der Anteil der Fehlzuordnungen nach einer linearen Diskriminanzanalyse mit 2 % in (hoffentlich) akzeptablen Grenzen. Allerdings zeigt der Plot dass die Trennfunktionen, insbesondere zwischen den Gruppen 2 (rot) und 3 (grün), in der Steigung deutlich abweichen. (Ob, wie gut und unter welchen Bedingungen sich das Ergebnis mit einer quadratischen Analyse verbessert werden kann, sei Ihren Versuchen überlassen. Genauso spannend ist die asymmetrische Änderung der Wertbereiche der Indikatoren.)

Number of Observations and Percent Classified into group

From group	1	2	3	Total
1	2364	28	18	2410
	98.09	1.16	0.75	100.00
2	13	1884	30	1927
	0.67	97.77	1.56	100.00
3	14	13	2197	2224
	0.63	0.58	98.79	100.00
Total	2391	1925	2245	6561
	36.44	29.34	34.22	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for group

	1	2	3	Total
Rate	0.0191	0.0223	0.0121	0.0178
Priors	0.3333	0.3333	0.3333	

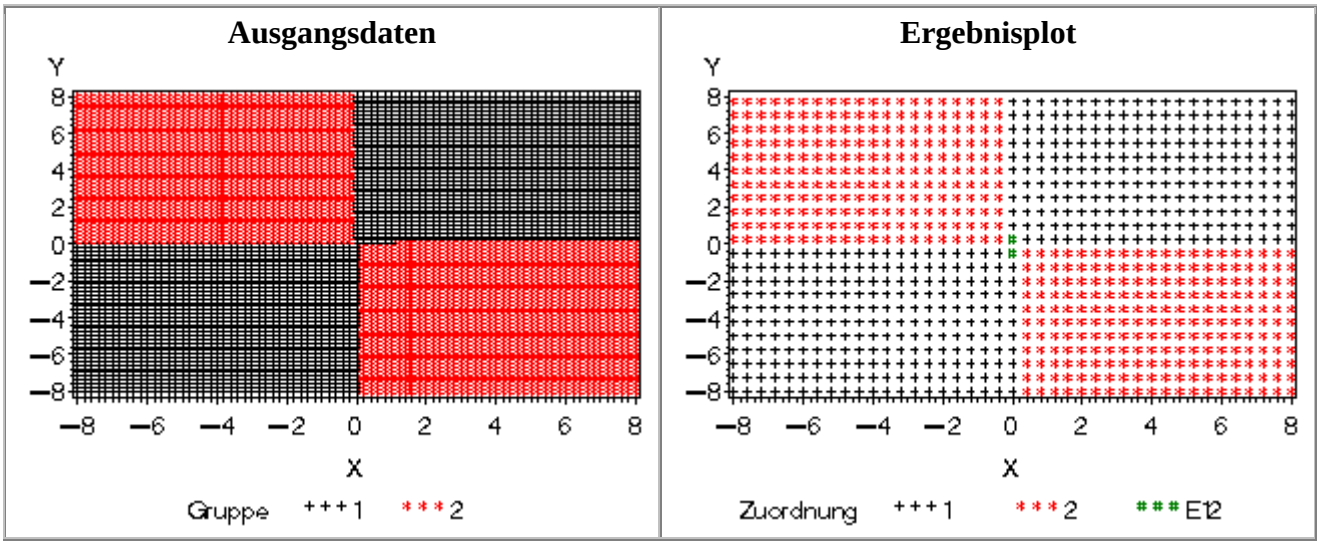


Die nichtparametrische Analyse liefert mit insgesamt 4 Fehlklassifikationen eine praktische fehlerfreie Gruppierung.

6.3 2 Gruppen in 4 Feldern (korrelierte Indikatoren)

Das Beispiel mit den korrelierten Indikatoren wird hier vorgezogen, weil das Ergebnis besser und einfacher zu interpretieren ist.

Mit einer quadratischen Diskriminanzanalyse lassen sich die beiden Gruppen fast vollständig reproduzieren. Lediglich die schmale, und schon in der obigen Analyse eher zufällige, Verbindung zwischen den Segmenten der Gruppe 1 lässt sich nicht reproduzieren und führt zu 11 Fehlzuordnungen (0,33 %).

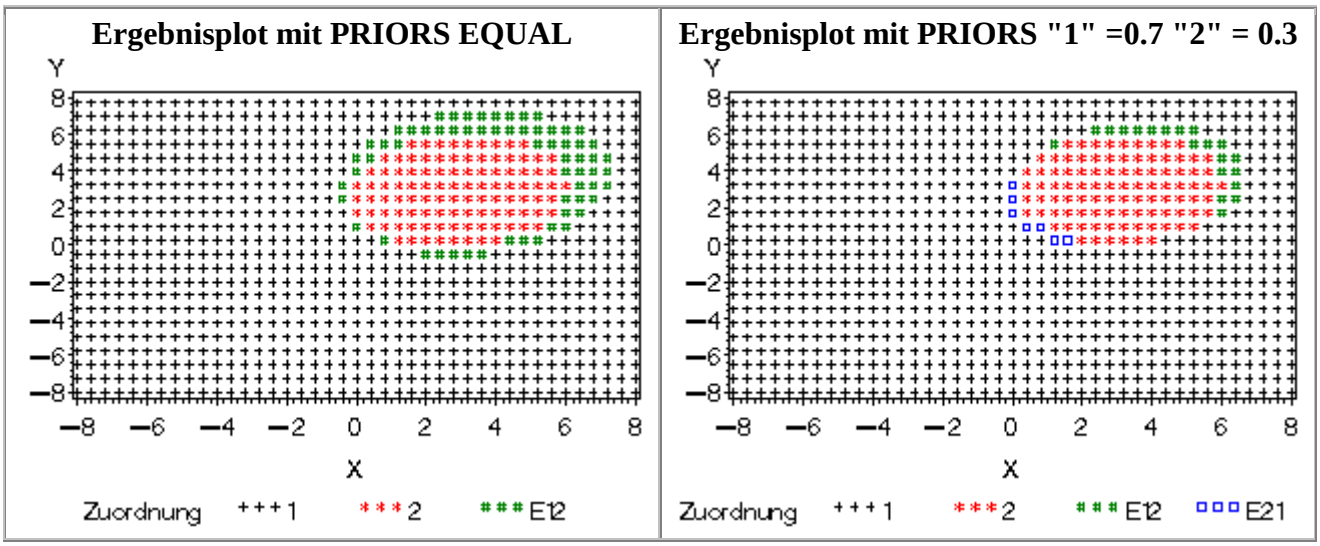


Die nicht-parametrische Analyse liefert auch hier eine fast perfekte Reproduktion der beiden Gruppen. Allerdings ist der Vorteil mit nur 2 gegenüber 11 Fehlzuordnungen bei insgesamt 6500 Fällen eher marginal. Etwas deutlicher ist der Vorteil bei der zweiten Gruppierung mit der Priorisierung der ersten Gruppe.

6.4 2. Gruppe als kreisförmiges Segment

Dieses Beispiel weicht sehr stark von der ursprünglichen Problemstellung ab. In der ersten Gruppe befinden sich 5635 Beobachtungen und in der zweiten 926 - prozentual ausgedrückt beträgt das Verhältnis 86% zu 14%.

Mit PRIORS EQUAL kommt es zu 448 (rund 8 %) Fehlzuordnungen aus der ersten (großen) Gruppe, die zudem nicht symmetrisch um die zweite Gruppe verteilt sind. Mit explizit gesetzten a-priori-Wahrscheinlichkeiten zeigt sich eine deutliche Verschiebung: 139 (2,5 %) Beobachtungen mit großen x- und y-Werten aus der ersten Gruppe werden falsch zur zweiten Gruppe zugeordnet, während 68 (8,5 %) Beobachtungen mit kleinen x- und y-Werten aus der zweiten Gruppe zur ersten Gruppe zugeordnet werden. Mit PRIORS PROP wäre diese Verschiebung noch extremer.



Erklärbar ist dieses Ergebnis, wenn man berücksichtigt, dass bei multivariater Normalverteilung im Bereich der kleineren x- und y-Werte die Dichte der Gruppe 1 höher wäre und einen Teil der kleinen Dichte aus der Gruppe 2 überlagern würde. Umgekehrt wäre die Dichte der Gruppe 1 bei großen x- und y-Werten kleiner. (Ein ähnlicher Effekt wäre auch im ursprünglichen Beispiel bei gleicher Gruppengröße zu beobachten gewesen.)

Die nicht-parametrische Analyse (nearest neighbours) mit PRIORS EQUAL liefert 116 Fehlzuordnungen aus der Gruppe 1, die symmetrische um die Gruppe 2 angeordnet sind. Mit PRIORS PROP läßt sich die Gruppierung genau reproduzieren.

In diesem Beispiel kann das Ergebnis der parametrischen Analyse durch eine nicht-lineare Transformation der Indikatoren verbessert werden. Mit einer solchen Transformation kann die Kurtosis ("Gipfligkeit") der Gruppe 2 erhöht werden, um die Dichteverteilung dieser Gruppe an die Normalverteilung anzunähern. Ein einfaches, wenn auch unzulängliches Verfahren wäre, die Mittelwerte der zweiten Gruppe von den Indikatoren abzuziehen und die Indikatoren dann unter Beibehaltung des Vorzeichens zu quadrieren. Im folgenden Programm werden die quadrierten Mittelwerte anschließend wieder dazuaddiert, was aber nur kosmetischen Charakter hat:

```
* Parameter;
%let exp = 2;
%let xmean = 2.92;
%let ymean = 2.90;

DATA work.test;
 SET work.test;
 * Transformation von x;
 x = x - &xmean;
 if x LT 0 then x = -((-x) ** &exp);
           else x = x ** &exp;
 x = x + &xmean ** &exp;
 * Transformation von y;
 y = y - &ymean;
 if y LT 0 then y = -((-x) ** &exp);
           else y = y ** &exp;
 y = y + &mean ** &exp;
RUN;
```

Nach dieser Transformation erhalten wir mit PRIORS EQUAL noch 389 (5 %) Fehlzuordnungen aus der ersten (großen) Gruppe und mit explizit gesetzten a-priori-Wahrscheinlichkeiten ("1" =0.8 "2" = 0.2 - also fast proportional zur Häufigkeit) 145 (2,5 %) Fehlzuordnungen aus der ersten Gruppe bei nur 3 Fehlzuordnungen (mit kleinen x- und y-Werten) aus der zweiten Gruppe zur ersten.

Natürlich ließe sich das Ergebnis mit einer besser gewählten Transformation, die zu einer kreis- statt rautenförmigen Verteilung der zweiten Gruppe führt noch verbessern, aber die Wahl einer optimalen Transformation ist abhängig von der konkreten Verteilung und kann nicht Gegenstand dieses Referats sein. Wichtig war mir hier zu zeigen, dass sich das Ergebnis der Gruppierung auch mit einer suboptimalen Transformation deutlich verbessern lässt.

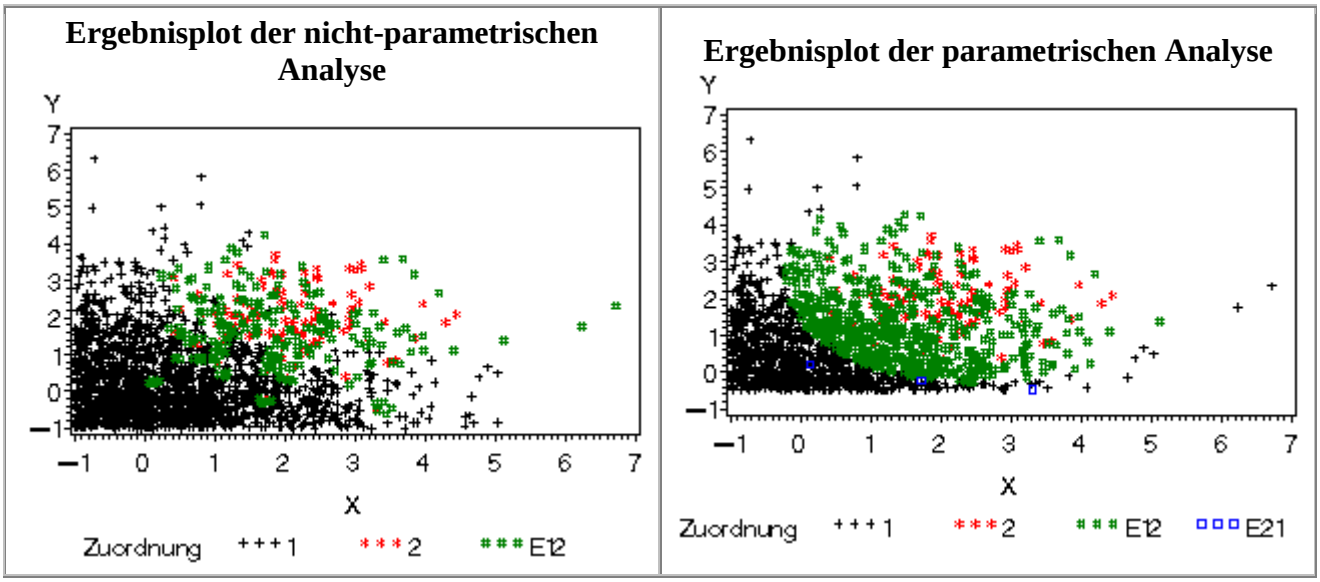
7. Nicht-parametrische Analyse auf Beispieldaten

Angesichts der guten Ergebnisse der nicht-parametrischen Analyse mit der nearest-neighbour-Methode, könnte man auf die Idee kommen, diese sei bei nicht-normalverteilten Daten grundsätzlich besser. Tatsächlich sind die guten Ergebnisse aber eine Folge der eindeutigen Separierung der beiden Gruppen.

Würde man nur die Zahlen betrachten, könnte man auch bei den Beispielen mit normalverteilten Indikatoren auf die Idee kommen, die nichtparametrische Methode sei zumindest nicht schlechter. Die graphische Kontrolle und Interpretation des Ergebnisses zeigt aber, dass die "scheinbar guten Zahlen" nicht verallgemeinerbar sind. Besonders deutlich wird dies am Beispiel mit dem kreisförmigen Segment, d.h. der kleinen Gruppe mit der geringeren Varianz.

Angewandt auf die Originaldaten liefert die nichtparametrische Analyse mit PRIORS EQUAL nur 231 Fehlzuordnungen aus der Gruppe 1, während wir mit der parametrischen Analyse 562 Fehlzuordnungen aus der Gruppe 1 und 3 Fehlzuordnungen aus der Gruppe 2 erhalten. Die Ergebnisplots zeigen aber, dass der Unterschied vor allem daher rührt, dass zufällige lokale Häufungen von Beobachtungen aus der Gruppe 1 bei der nichtparametrischen Analyse "richtig" zugeordnet werden. Ein solches Ergebnis ist mit einem zweiten Sample mit Sicherheit nicht reproduzierbar. Eine Test-Stichprobe würde diesen Fehler aufdecken.

Dagegen liefert die parametrische Analyse einen klar abgegrenzten Bereich, in dem die Beobachtungen der zweiten Gruppe zu erwarten sind.



Ähnlich sieht das Ergebnis im "Kreuzungsbereich" beim Beispiel mit den korrelierten Indikatoren aus.

8. Zusammenfassung

Insgesamt erweist sich die parametrische Diskriminanzanalyse als weitgehend robustes Verfahren zur Gruppierung, das stabile und nützliche Ergebnisse liefert. Mit der quadratischen Methode werden auch die Unterschiede in den Kovarianzmatrizen der Gruppen zur Gruppierung genutzt.

Grundsätzlich sollte zur Bestimmung der Parameter der gesamte, ggf. um Ausreißer bereinigte, Datenbestand verwendet werden. Präferenzen und Ziele lassen sich über explizit gesetzte a-priori-Wahrscheinlichkeiten einführen.

Lern- und Test-Stichproben liefern Anhaltspunkte für die Angemessenheit des Verfahrens. Um zufällige Ergebnisse zu vermeiden, sollte dieser Test aber mehrfach durchgeführt werden. Fehlerquellen und Möglichkeiten zur Optimierung lassen sich aber durch eine graphische Analyse und Interpretation der Ursprungsdaten und der Ergebnisse besser identifizieren. Bei identifizierten potentiellen Problemen kann das Verhalten der Methode mit künstlich erzeugten Datensätzen getestet werden.

Eine nichtlineare Datentransformation zur Manipulation der Verteilungen ist nur in Ausnahmefällen sinnvoll. Ihr Ergebnis muss nicht notwendigerweise die Anpassung der Verteilungen in allen Gruppen an die Normalverteilung sein, um eine deutliche Verbesserung der Gruppierung zu erzielen.

Von den nichtparametrischen Verfahren der Prozedur Discrim ist bei großen Fallzahlen nur die nearest-neighbour-Methode ausreichend performant. Dabei sollten die Gruppen keine "Überschneidungen" aufweisen. Bessere Ergebnisse in der "Lernstichprobe" können durchaus zufällig sein. Deshalb muss das Ergebnis graphisch kontrolliert und mit Lern- und Test-Samples geprüft werden.

9. Fragen und Antworten

In Diskussionen während und nach der Präsentation kamen einige Fragen auf, die in diesen Vortrag nicht mehr eingearbeitet werden konnten und deshalb hier abgehandelt werden:

Werden die a-priori-Wahrscheinlichkeiten hier "missbraucht"? Müssen sie den erwarteten relativen Häufigkeit der einzelnen Gruppen entsprechen?
Als a-priori-Wahrscheinlichkeiten werden solche Wahrscheinlichkeiten bezeichnet, die vor Ermittlung der Diskriminanzfunktion hinsichtlich der Gruppenzugehörigkeit gegeben sind oder geschätzt werden können. Wenn die Konsequenzen ("Kosten") der Fehlklassifikation zwischen den Gruppen differieren, kann dieses Konzept durch Anwendung der Bayes'schen Entscheidungsregel erweitert werden. Dabei wird der Erwartungswert eines Kostenkriteriums minimiert bzw. eines Nutzenkriteriums maximiert. (Vgl. Klaus Backhaus u.a.: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung, 4. Aufl., Springer-Verlag 1987, S. 190ff)

Bei verletzten Verteilungsannahmen werden die a-priori-Wahrscheinlichkeiten hier zusätzlich verwendet, um die Konsequenzen der Fehler, d.h. die Abweichungen der Verteilungen in den Grenzbereichen zwischen den Gruppen von der geschätzten Normalverteilung, zu korrigieren.
In der Kombination Einführung von Zielkriterien und Korrektur erfolgt die Bestimmung der a-priori-Wahrscheinlichkeiten iterativ an Hand des Gruppierungsergebnisses. D.h. die a-priori-Wahrscheinlichkeiten werden solange geändert, bis das Verhältnis der Fehlklassifikationen den Zielkriterien entspricht. Betrachtet man nur den Nutzen der Gruppierung, ist dieses Verfahren m.E. zulässig. Die Parameter des Modells selbst sind dann aber nur eingeschränkt interpretierbar.

Im Vortrag wird vollständig auf die Erläuterung der Methode selbst, also der Algorithmen verzichtet. Ist es sinnvoll eine Methode anzuwenden, ohne sie vollständig zu verstehen?

Im Prinzip: Nein. Aber ich habe bisher nur wenige Menschen kennen gelernt, die z.B. in der Lage sind anzugeben, an welcher Stelle der Herleitung eines Verfahrens welche Annahme eingeführt wird, und welche Konsequenzen es hätte, wenn auf diese Annahme verzichtet wird. Bei den immer kürzeren Studienzeiten ist ein solches tiefes Verständnis der Methoden auch nicht mehr vermittelbar. Deshalb wurde hier ein Zugang versucht, der auf ein grundlegendes Verständnis der Prinzipien (Verteilungen, Varianzen, Kovarianzen) aufsetzt und das Verfahren selbst möglichst anschaulich und leicht verständlich darstellt. Ich verbinde damit die Hoffnung, dass eine solche anschauliche Darstellung insbesondere auch den Mitarbeitern in Unternehmen hilft, ein Verfahren besser anzuwenden und anderen die Ergebnisse zu erläutern.

In der Einleitung wurde das Ziel angeführt, Kreditausfallrisiken zu erkennen. Waren die Indikatoren intervallskaliert oder nur ordinal?

Da ich in dem Projekt selbst nicht beteiligt war, kann ich diese Frage letztlich nicht beantworten. In der Regel lassen sich aber konstruierte Indikatoren, die aus mehreren Quellen gebildet werden, als intervallskaliert behandeln.

Betrachtet man nur das Gruppierungsergebnis, ist die Frage des Skalenniveaus (intervallskaliert oder ordinal) zweitrangig, weil die individuellen Werte der Diskriminanzfunktionen nicht interpretiert werden. Wichtig ist die möglichst große Anzahl der möglichen Ausprägungen eines Indikators und die Verteilung der Werte.

Ist die Diskriminanzanalyse bei dieser Problemstellung (Kreditausfallrisiken) überhaupt das geeignete Verfahren?

Diese Frage lässt sich nur pragmatisch beantworten. Etwas vereinfacht und bildlich: Mit der Diskriminanzanalyse werden die Daten von unabhängigen Gruppen verglichen, um anschließend die Gruppenzugehörigkeit einer Beobachtung aufgrund dieser Daten zu schätzen. Essentiell ist dabei die diskrete Trennung der Gruppen. Kreditengagements sind aber nicht einfach "gut" oder "schlecht" sondern mehr oder weniger risikobehaftet. Es handelt sich also nicht um zwei unabhängige Gruppen. Selbst wenn das Risiko eines Engagements nur dichotom (0,1) ermittelt werden kann, würde es sich anbieten, eine Funktion in der Form

```
risiko = const. + b1*i1 + b2*i2 + b12*i1*i2 ... +err.
```

zu schätzen.

Allerdings befinden sich in der zweiten Gruppe nur die Fälle, bei denen das besondere Risiko a posteriori durch einen Schadensfall erkannt wurde. Alle Fälle mit großem Risiko, in denen ein Schaden durch gutes Krisenmanagement vermieden wurde, bleiben unerkannt.

Vor diesem Hintergrund halte ich die Diskriminanzanlyse mit dem Ziel, möglichst alle erkannten Risikofälle richtig zuzuordnen, für ein geeignetes Verfahren. Dabei wäre zu vermuten, dass ein großer Teil der Fehlklassifikationen auf abgewendete Schäden bei gleicher Ausgangssituation zurückzuführen ist.

Probleme die sich aus verletzten Verteilungsannahmen ergeben, lassen sich nur am konkreten Fall diskutieren.

10. Kontakt

Wilfried Schollenberger

WS Unternehmensberatung und Controlling-Systeme GmbH
Bergstraße 7
69120 Heidelberg

Tel: 06221 / 401 409
EMail: wisch @ ws-unternehmensberatung.de
WEB: www.ws-unternehmensberatung.de

Hinweis: Unter http://www.ws-unternehmensberatung.de/KSFE2001 finden Sie weitere Programme, Ergebnisse und Grafiken, die hier aus Platzgründen nicht aufgenommen wurden.