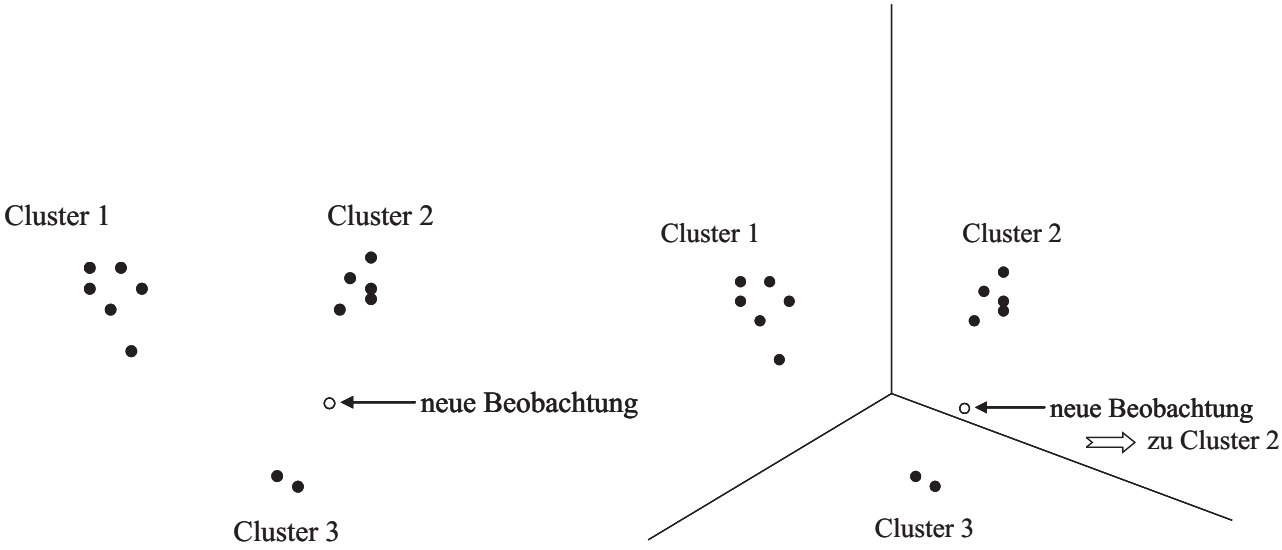
9 Diskriminanzanalyse

Ziel einer Diskriminanzanalyse: Bereits bekannte Objektgruppen (Klassen/Cluster) anhand ihrer Merkmale charakterisieren und unterscheiden sowie neue Objekte in die Klassen einordnen.
Nötig: Lernstichprobe von Objekten mit bekannter Klassenzugehörigkeit, um Abgrenzung der verschiedenen Gruppen anhand der beobachteten Merkmale zu lernen.
Die Diskriminanzanalyse fällt unter die so genannten Klassifikations- oder auch Mustererkennungsmethoden (engl. *pattern recognition*).



Beispiel 9.1. Kreditscoring
Die Vergabe von Krediten hängt von der Bereitschaft und Fähigkeit der Kunden ab, die anfallenden Zins- und Tilgungsraten zu bezahlen. Banken stufen potenzielle Kunden vor Kreditvergabe entweder als problemlos oder als Problemfall ein. Problemfälle werden genauer geprüft und der Kredit ggf. abgelehnt. Die Einstufung erfolgt auf Basis von charakterisierenden Merkmalen der Kunden hinsichtlich ihrer persönlichen, wirtschaftlichen und rechtlichen Situation.

Weitere Beispiele: Unterscheidung von

- handgeschriebenen Ziffern (Postleitzahlen),
- Käufern und Nicht-Käufern eines neuen Produktes,
- Texte verschiedener Autoren,
- und natürlich echten und falschen Schweizer Banknoten.

23. 1. 2014
23.Vorlesung

Definition 9.2. Modell der Diskriminanzanalyse
Eine Grundgesamtheit Ω bestehe aus mehreren Klassen (Gruppen) $C_1, \ldots, C_k$, so dass jedes Element (Objekt) $\omega \in \Omega$ zu genau einer Gruppe gehört. Für die Zerlegung $C_1, \ldots, C_k$ von Ω gelte also $C_i \cap C_j = \emptyset$ für $i \neq j$ und $\cup_{j=1}^{k} C_j = \Omega$.
Ziel ist es, für ein Objekt $\omega \in \Omega$ mit unbekannter Klassenzugehörigkeit anhand eines beobachteten Merkmalsvektors $\mathbf{x}$ die zugehörige Klasse C_j zu ermitteln.

Bemerkung 9.3. (Lernstichprobe)

- In der Diskriminanzanalyse werden in der Regel nicht die Klassen selbst, sondern nur bestimmte Merkmale der Objekte beobachtet, anhand derer die Klassenzugehörigkeit festzustellen ist.
- Um typische Werte der Merkmale für die verschiedenen Klassen zu ermitteln, steht eine Lernstichprobe von Objekten zur Verfügung, für welche die Merkmalsausprägungen und die Klassenzugehörigkeit bekannt sind.
- Lernstichprobe: $(\mathbf{x}_1', y_1)', \ldots, (\mathbf{x}_n', y_n)'$ mit $y_i = j \iff$ Objekt i gehört zu Klasse C_j, $i = 1, \ldots, n$, die ZVe Y gebe die Klassenzugehörigkeit an.

Beispiel 9.4. Kredite (Fortsetzung Beispiel 9.1)
Eine süddeutsche Großbank benutzt zur Einschätzung des Kreditrisikos ihrer Kunden eine Lernstichprobe von 1000 ehemaligen Kreditnehmern. 300 dieser ehemaligen Kunden zahlten den Kredit nicht vereinbarungsgemäß zurück. Es wurden folgende Merkmale erfasst:

- Kredit zurückgezahlt (ja; nein)
- bestehendes laufendes Konto bei der Bank (nein; ja, aber im Minus; ja, mit geringem Betrag; ja, als Gehaltskonto oder in beträchtlicher Höhe)
- Laufzeit des Kredits (in halben Jahren, bis zu 5 Jahren)
- bisherige Zahlungsmoral (von schlecht bis sehr gut)
- Verwendungszweck des Kredits (PKW; Möbel; Radio/Fernsehen; Haushalt; Reparaturen; Ausbildung; Urlaub; Umschulung; Betrieb; Sonstiges)
- Darlehenshöhe (in insgesamt 10 Kategorien von < 500 bis $> 20\,000$ DM)
- Sparkonto oder Wertpapiere vorhanden (nach Anlagehöhen gestaffelt)
- Dauer der Beschäftigung bei derzeitigem Arbeitgeber
- Ratenhöhe in % des verfügbaren Einkommens
- Familienstand und Geschlecht
- weitere Schuldner / Bürgen beteiligt
- in der jetzigen Wohnung seit ... Jahren
- Vermögen vorhanden (Haus- und Grundbesitz; Bausparvertrag, Lebensversicherung; PKW, Sonstiges; keins)
- Alter (in Altersklassen)
- weitere Ratenkredite anderswo (andere Bank; Kauf-/Versandhaus; keine)
- Art der Wohnung (Miete; Eigentum; kostenlos überlassen)
- Anzahl bisheriger Ratenkredite einschl. des laufenden

- Beruf (nicht beschäftigt, ungelernt nicht sesshaft; ungelernt sesshaft; Facharbeiter, gelernte Angestellte, Beamte bis mittlerer Dienst; Führungskraft, selbstständig, Beamter höherer Dienst)
- Anzahl unterhaltsberechtigter Personen, die zu versorgen sind
- Telefon (nein; ja, unter dem Namen des Kunden)
- Gastarbeiter (ja; nein)

Aus den Charakteristika der Kunden in der Lernstichprobe und der Kenntnis über die Rückzahlung ihrer Kredite wurden Regeln abgeleitet, nach denen künftige potenzielle Kunden als unproblematisch oder als Risikofall eingestuft werden.

Eine Regel, nach der Objekte zu den einzelnen Klassen zugeordnet werden, basiert auf einer so genannten **Diskriminanzfunktion**.

Definition 9.5. (Diskriminanzfunktion, Diskriminanzregel)
Betrachtet wird ein Modell der Diskriminanzanalyse wie in Definition 9.2. Zu einem Objekt ω werde ein Merkmalsvektor $\mathbf{x}$ beobachtet. Eine Funktion D, die dem Beobachtungsvektor $\mathbf{x}$ für jede Gruppe C_i der Grundgesamtheit einen charakterisierenden Wert $D(\mathbf{x}, C_i)$ zuordnet, heißt **Diskriminanzfunktion**. Eine Regel, die anhand von $D(\mathbf{x}, C_1), \ldots, D(\mathbf{x}, C_k)$ entscheidet, welcher Gruppe C_i das Objekt ω zugeordnet wird, heißt **Diskriminanzregel**.

Nachfolgend meist Betrachtung metrisch skalierter stetiger Merkmale.

9.1 Lineare Diskriminanzanalyse nach Fisher

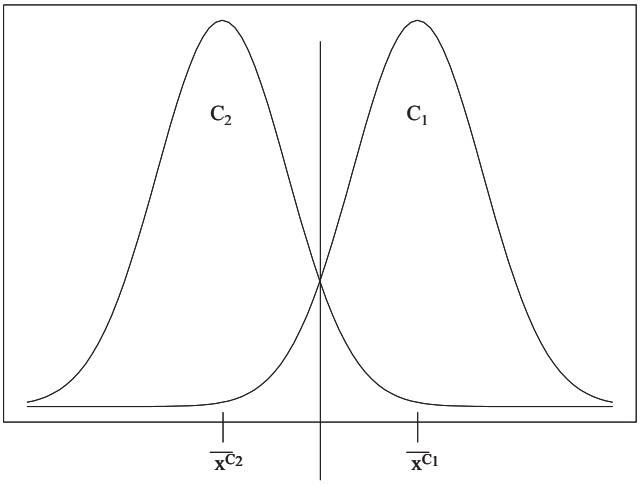
Idee einer Diskriminanzregel bei nur einem Merkmal und zwei Klassen

Die Grundgesamtheit Ω zerfalle in die Klassen C_1 und C_2.
Pro Objekt werde ein eindimensionales Merkmal X beobachtet.
Die Lernstichprobe enthalte n_1 Objekte aus C_1 und n_2 Objekte aus C_2:

$$\{x_1, \ldots, x_n\} = \{y_1, \ldots, y_{n_1}\} \cup \{z_1, \ldots, z_{n_2}\}, \quad n = n_1 + n_2,$$

$\overline{y} = \overline{x}^{C_1}$, $\overline{z} = \overline{x}^{C_2}$: arithmetische Mittel der Klassen C_1 bzw. C_2.

Schätzung der Häufigkeitsverteilungen des Merkmals X in den beiden Klassen, (z. B. durch geglättete Histogramme oder Anpassen von Normalverteilungen):



Größere Werte von X sprechen tendenziell für die Zugehörigkeit zu C_1, während kleinere Werte von X für Klasse C_2 sprechen. Suche Trennpunkt t zwischen "großen" und "kleinen" Werten, vgl. senkrechten Trennstrich in Abb. Der Trennpunkt kann als Mitte zwischen den Gruppenmittelwerten der Lernstichprobe festgelegt werden,

$$t = \frac{\overline{y} + \overline{z}}{2}.$$

Für neue Beobachtung x würde man bei der vorliegenden Lernstichprobe entscheiden, dass x aus C_1 (C_2) stammt, wenn $x > t$ ($x < t$) ($x = t$ hat für stetige Merkmale Wahrscheinlichkeit Null und ist vernachlässigbar).

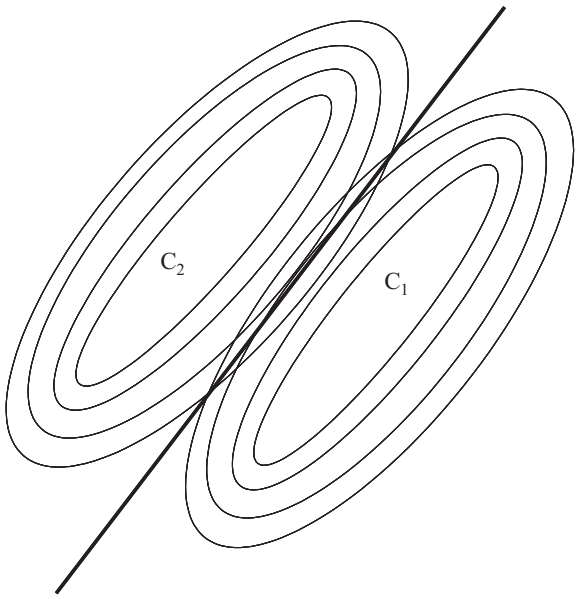
Resultierende Diskriminanzfunktion:

$$D(x, C_1) = \begin{cases} 1, & x > t, \\ 0, & x < t, \end{cases} \qquad D(x, C_2) = \begin{cases} 0, & x > t, \\ 1, & x < t. \end{cases}.$$

Diskriminanzregel: ordne neue Beobachtung x der Klasse C_i mit $D(x, C_i) = 1$ zu.

Diese Wahl des Trennpunkts t unterstellt implizit, dass das Merkmal X in beiden Gruppen dieselbe Varianz besitzt.

Idee der Diskriminanzregel bei zwei Merkmalen (und zwei Klassen)
Ein Merkmal alleine führt selten zu guten Trennungen zwischen den Gruppen: Suche bessere Unterscheidung anhand mehrerer Merkmale. Bei Vorliegen eines **zweidimensionalen Merkmalsvektors** $\mathbf{X} = (X_1, X_2)'$ ist es sinnvoll, X_1 und X_2 nicht einzeln zu betrachten, sondern eine Kombination aus beiden zur Trennung der Gruppen zu verwenden. Visualisierung der Häufigkeitsverteilungen in den Klassen über die Dichtekonturlinien der zugehörigen Verteilungen visualisieren.



Die Trennung zwischen den beiden Gruppen erfolgt dann nicht mit einem Trennpunkt, sondern mit einer Trenngeraden.

Verallgemeinerung auf p-dimensionale Merkmale $\mathbf{X}$:

- Zerlegung des R^p in zwei Klassen, Trennung mittel $(p-1)$-dim. Hyperebene.
- Suche Richtung $\mathbf{a}$ (Normalenvektor der Hyperebene), so dass Projektionen $\alpha_i = \mathbf{a}'\mathbf{x}_i = a_1 x_{i,1} + \ldots + a_p x_{i,p}$ im $\mathbb{R}^1$ möglichst gut trennt werden.

Lineare Diskriminanzanalyse mit p Merkmalen und k Klassen

Notation: Stichprobe $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^p$ zerfällt in k Cluster $C_1, \ldots, C_k$, wobei

$$C_j = \{\mathbf{x}_1^{C_j}, \ldots, \mathbf{x}_{n_j}^{C_j}\}, \quad j = 1, \ldots, k.$$

Die Clusterzentren sind $\overline{\mathbf{x}}^{C_1}, \ldots, \overline{\mathbf{x}}^{C_k}$. Seien $\mathbf{a} \in \mathbb{R}^p$ mit $\mathbf{a}'\mathbf{a} = 1$ und

$$\alpha_i = \mathbf{a}'\mathbf{x}_i,\ \overline{\alpha} = \mathbf{a}'\overline{\mathbf{x}}, \qquad i = 1, \ldots, n,$$

$$\alpha_i^{C_j} = \mathbf{a}'\mathbf{x}_i^{C_j},\ \overline{\alpha}^{C_j} = \mathbf{a}'\overline{\mathbf{x}}^{C_j}, \qquad i = 1, \ldots, n_j,\ j = 1, \ldots, k$$

die Projektionen der Beobachtungen auf die Achse mit Richtungsvektor $\mathbf{a}$.

Die klassische ANOVA-Varianzzerlegung (der projizierten Daten):

$$\sum_{i=1}^{n}(\alpha_i - \overline{\alpha})^2 \;=\; \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\alpha_i^{C_j} - \overline{\alpha}^{C_j})^2 + \sum_{j=1}^{k} n_j(\overline{\alpha}^{C_j} - \overline{\alpha})^2$$

Maximiere Verhältnis von *Between-Group-Sum-of-Squares* zu *Within-Groups-Sum-of-Squares*:

$$f(\mathbf{a}) = \frac{\sum_{j=1}^{k} n_j(\overline{\alpha}^{C_j} - \overline{\alpha})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\alpha_i^{C_j} - \overline{\alpha}^{C_j})^2} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}, \tag{7}$$

wobei

$$\mathbf{B} = \sum_{j=1}^{k} n_j (\overline{\mathbf{x}}^{C_j} - \overline{\mathbf{x}})(\overline{\mathbf{x}}^{C_j} - \overline{\mathbf{x}})', \qquad \mathbf{W} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{x}_i^{C_j} - \overline{\mathbf{x}}^{C_j})(\mathbf{x}_i^{C_j} - \overline{\mathbf{x}}^{C_j})'.$$

$\frac{1}{n}\mathbf{W}$ ist eine gepoolte Kovarianzschätzung: $\mathbf{W} = \sum_{j=1}^{k} n_j \hat{\mathbf{\Sigma}}^{C_j}$, wobei

$$\hat{\mathbf{\Sigma}}^{C_j} \;=\; \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{x}_i^{C_j} - \overline{\mathbf{x}}^{C_j})(\mathbf{x}_i^{C_j} - \overline{\mathbf{x}}^{C_j}) \;=\; \frac{1}{n_j} \left(\mathbb{X}^{C_j}\right)' \mathcal{H}_{n_j} \mathbb{X}^{C_j}$$

die empirische Kovarianzmatrix des Clusters C_j ist, $\mathbb{X}^{C_j} = (\mathbf{x}_1^{C_j}, \ldots, \mathbf{x}_{n_j}^{C_j})'$ die Datenmatrix des Clusters C_j und $\mathcal{H}_{n_j}$ die Zentrierungsmatrix.

Die Lösung des OP's (7) ist bekannt (vgl. Lemma 6.2): Lösungsvektor $\mathbf{a}$ ist der Eigenvektor zum größten Eigenwert von $\mathbf{W}^{-1}\mathbf{B}$.

Klassifikationsregel: Gruppiere $\mathbf{x}$ in die Klasse j mit

$$j = \arg\min_i |\mathbf{a}'(\mathbf{x} - \overline{\mathbf{x}}^{C_i})|.$$

Spezialfall $k = 2$**.** (führt zu etwas übersichtlicherer Notation.)
Grundgesamtheit Ω zerfalle in zwei Klassen C_1 und C_2. Die Lernstichprobe enthalte n_1 Objekte aus C_1 und n_2 Objekte aus C_2:

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}\} \cup \{\mathbf{z}_1, \ldots, \mathbf{z}_{n_2}\}, \qquad n = n_1 + n_2$$

Dann ergibt sich für das zu maximierende Zielkriterium (7):

$$f(\mathbf{a}) \propto \frac{\left[\mathbf{a}'(\overline{\mathbf{y}} - \overline{\mathbf{z}})\right]^2}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

und für den Lösungsvektor $\mathbf{a}$ (muss noch auf Länge 1 normiert werden):

$$\mathbf{a} = \mathbf{W}^{-1}(\overline{\mathbf{y}} - \overline{\mathbf{z}}).$$

Einordnung einer neuen Beobachtung $\mathbf{x}$:

- Diskriminanzfunktion:

$$D(\mathbf{x}, C_1) = \begin{cases} 1, & \mathbf{a}'\mathbf{x} > t, \\ 0, & \mathbf{a}'\mathbf{x} < t, \end{cases}, \qquad D(\mathbf{x}, C_2) = \begin{cases} 0, & \mathbf{a}'\mathbf{x} > t, \\ 1, & \mathbf{a}'\mathbf{x} < t, \end{cases}$$

  wobei

$$t = \frac{\overline{\alpha}^{C_1} + \overline{\alpha}^{C_2}}{2} = \mathbf{a}' \left(\frac{\overline{\mathbf{y}} + \overline{\mathbf{z}}}{2} \right)$$

- Diskriminanzregel: Ordne Beobachtung $\mathbf{x}$ der Klasse C_j zu, für die $D(\mathbf{x}, C_j) = 1$ ist.

Koeffizienten a_i der Diskriminanzfunktion geben Auskunft darüber, welche Variablen wie stark zur Trennung der Gruppen beitragen. Damit die Koeffizienten direkt vergleichbar sind, müssen sie geeignet standardisiert werden.

Die Lineare Diskriminanzfunktion stellt die einfachste Struktur zur Trennung von Gruppen dar (zweidimensional: Geraden, dreidimensional: Ebenen, höherdimensional: Hyperebenen als trennende Mengen), unterstellt implizit, dass alle Cluster die gleiche Kovarianzstruktur haben. Andere Trennfunktionen sind denkbar: Quadratische Funktionen führen zur so genannten **quadratischen Diskriminanzanalyse** (QDA).

Beispiel 9.6. Schweizer Banknoten:

$$\mathbf{a} = (-0.002, -0.327, 0.334, 0.439, 0.463, -0.612)'$$

Das sollte verglichen werden mit den ersten beiden Hauptkomponen-

ten, vgl. Bsp. 4.14:

$$\mathbf{a}_1 = (-0.044, 0.112, 0.139, 0.768, 0.202, -0.579)'$$

$$\mathbf{a}_2 = (0.011, 0.071, 0.066, -0.563, 0.659, -0.489)'$$

Die Hauptkomponenten $\mathbf{a}_1$ und $\mathbf{a}_2$ geben die beiden Richtungen mit der höchsten Gesamtvariabilität (über beide Gruppen hinweg) an, der Vektor $\mathbf{a}$ gibt die Richtung an, in der Zwischen-Gruppen-Variabilität relativ zur Vergleich zur Intra-Gruppen-Variabilität am größten ist.
Also wird eine Banknote $\mathbf{x}$ als falsch klassifiziert, falls

$$\mathbf{a}'\mathbf{x} > \mathbf{a}' \left(\frac{\overline{\mathbf{x}}_f + \overline{\mathbf{x}}_e}{2} \right) = -76.4967$$

Beispiel 9.7. Fishers berühmter Iris-Datensatz ('iris' in R), Fisher, 1936: The use of multiple measurements in taxonomic problems, Ann. Eugen. 7, 179-188.
Unterscheidung der Irisarten iris setosa (C_1) und iris versicolor (C_2) anhand von Länge und Breite des Kelchblattes ($p = 2$). Lernstichprobe mit $n_1 = n_2 = 50$ Pflanzen jeder Art.

$$\overline{\mathbf{x}}^{C_1} = (5.006, 3.428)', \qquad \overline{\mathbf{x}}^{C_2} = (5.936, 2.770)'$$

$$\mathbf{S}_x = \frac{1}{98}(49 \cdot \mathbf{S}_1 + 49 \cdot \mathbf{S}_2)$$

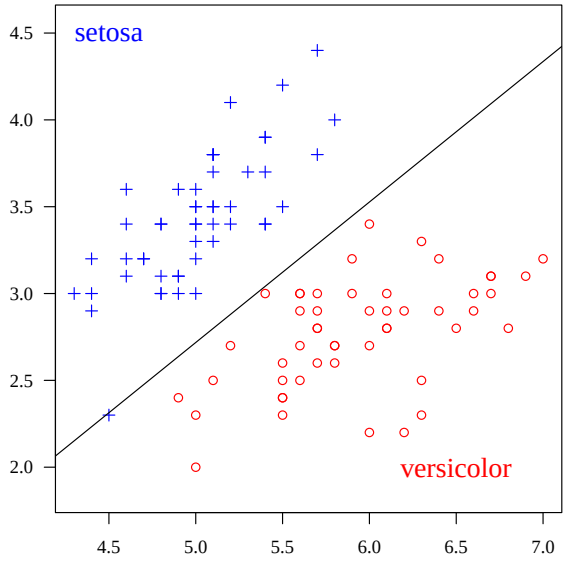
$$= \frac{1}{98} \left[\begin{pmatrix} 5.9682 & 4.7628 \\ 4.7628 & 6.8992 \end{pmatrix} + \begin{pmatrix} 12.7939 & 4.0915 \\ 4.0915 & 4.7825 \end{pmatrix} \right]$$

$$= \frac{1}{98} \begin{pmatrix} 18.7621 & 8.8543 \\ 8.8543 & 11.6277 \end{pmatrix}$$

$$\mathbf{S}_x^{-1} = \begin{pmatrix} 8.153 & -6.209 \\ -6.209 & 13.156 \end{pmatrix}, \qquad \rightsquigarrow \ \mathbf{a} = (-11.673, 14.431)'$$

Resultierende Diskriminanzregel: Neue Iris mit Merkmalsvektor $\mathbf{x} = (x_1, x_2)'$ in Gruppe iris setosa (C_1) einordnen, falls

$$-11.673x_1 + 14.431x_2 > -19.141$$



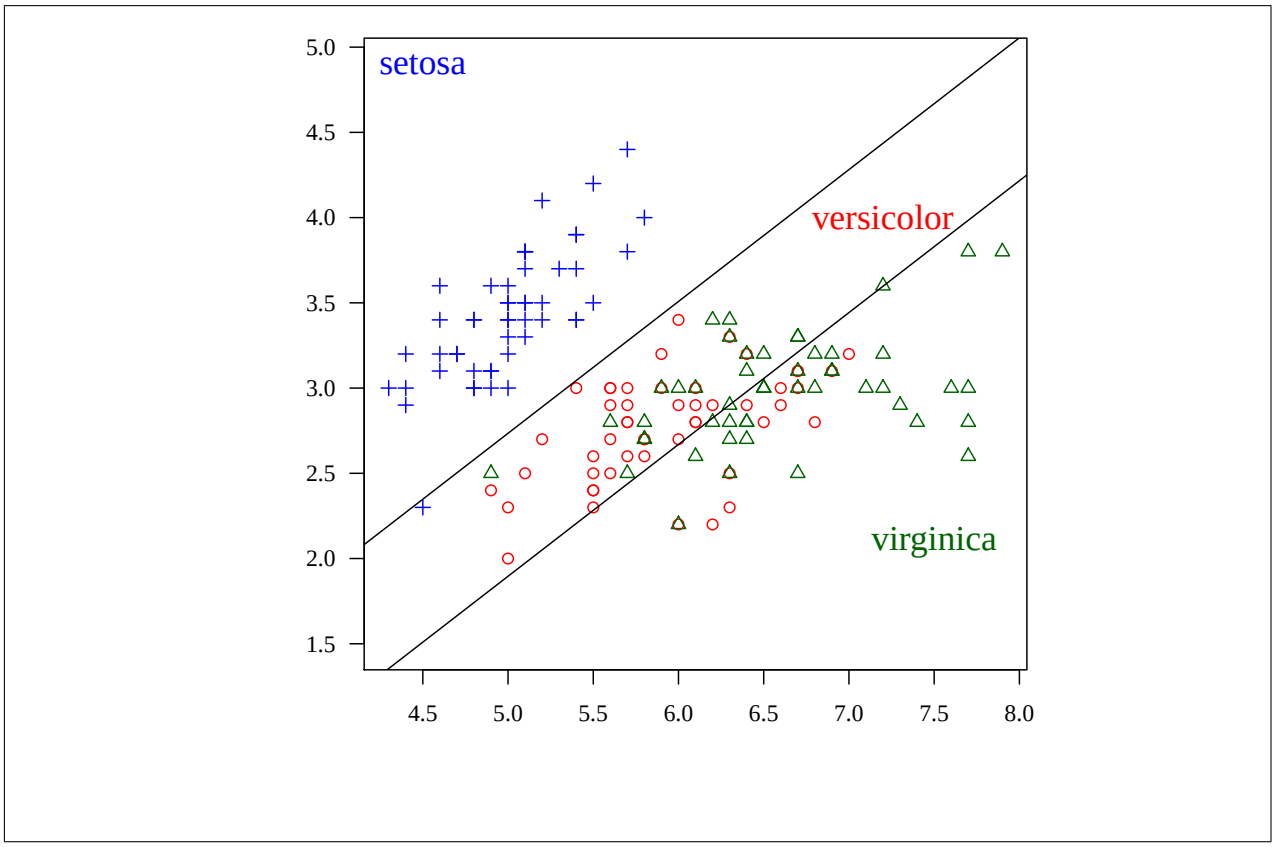
Bei $k = 3$ mit dritter Irisart iris virginica, $n_3 = 50$:

$$\begin{aligned} \overline{\mathbf{x}}^{C_3} &= (6.588, 2.974)' \\ \mathbf{S} &= \frac{1}{147}(49 \cdot \mathbf{S}_1 + 49 \cdot \mathbf{S}_2 + 49 \cdot \mathbf{S}_3) \\ \mathbf{W} &= \begin{pmatrix} 38.96 & 13.63 \\ 13.63 & 16.96 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 63.21 & -19.95 \\ -19.95 & 11.35 \end{pmatrix} \\ \mathbf{a} &= (1, -1.293)' \end{aligned}$$

Diskriminanzregel:

Vergleich von $|\mathbf{a}'\mathbf{x} - 0.5736|$, $|\mathbf{a}'\mathbf{x} - 2.3544|$ und $|\mathbf{a}'\mathbf{x} - 2.7426|$ führt zu

$$\begin{aligned} R_1 &= \{\mathbf{x} : \mathbf{a}'\mathbf{x} < 1.4640\}, \qquad R_2 = \{\mathbf{x} : 1.4640 < \mathbf{a}'\mathbf{x} < 2.5485\}, \\ R_3 &= \{\mathbf{x} : 2.5485 < \mathbf{a}'\mathbf{x}\} \end{aligned}$$



27. 1. 2014
24.Vorlesung

9.2 Maximum-Likelihood- und Bayes-Regeln

Notation: $f_j(\mathbf{x}) = f(\mathbf{x}|Y = j)$ Dichte von $\mathbf{X}$ unter $Y = j$, $j = 1, \ldots, k$,
$\hat{Y}$ Vorhersage von Y aus der erlernten Diskriminanzregel.
$R_j \subset \mathbb{R}^p$ Menge aller $\mathbf{x}$, die Klasse C_j zugeordnet werden, $j = 1, \ldots, k$.

Maximum Likelihood Diskriminanzregel:

$$R_j = \{\mathbf{x} : f_j(\mathbf{x}) > f_i(\mathbf{x}), i = 1, \ldots, k, i \neq j\}$$

Diese unterstellt implizit gleiche a-priori Wahrscheinlichkeiten für alle Klassen und gleiche Kosten aller möglichen Fehlklassifikationen, also gleiche Relevanz aller Gruppen.

Verallgemeinerung: Bayes-Regel

$p_j = P(Y = j)$ a-priori W'keit von C_j, $j = 1, \ldots, k$, mit $p_1 + \ldots + p_k = 1$.
Kosten von Fehlklassifikationen bei $k = 2$ Klassen

$y\|\hat{y}$	1	2
1	0	$c(\hat{Y} = 2\|Y = 1)$
2	$c(\hat{Y} = 1\|Y = 2)$	0

Kriterium: Minimiere erwartete gesamte Kosten

Erwartete gesamte Kosten (Expected Costs of Misclassification, ECM):

$$\begin{aligned} ECM &= c(\hat{Y} = 2|Y = 1)P(\hat{Y} = 2|Y = 1)P(Y = 1) \\ &\quad + c(\hat{Y} = 1|Y = 2)P(\hat{Y} = 1|Y = 2)P(Y = 2) \\ P(\hat{Y} = 2|Y = 1) &= P(\mathbf{X} \in R_2|Y = 1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} \\ P(\hat{Y} = 1|Y = 2) &= P(\mathbf{X} \in R_1|Y = 2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \end{aligned}$$

Bayes-Klassifikationsregel:

$$\begin{aligned} R_1: \ f_1(\mathbf{x})P(Y = 1) &\geq \frac{c(\hat{Y} = 1|Y = 2)}{c(\hat{Y} = 2|Y = 1)} f_2(\mathbf{x})P(Y = 2) \\ R_2: \ f_1(\mathbf{x})P(Y = 1) &< \frac{c(\hat{Y} = 1|Y = 2)}{c(\hat{Y} = 2|Y = 1)} f_2(\mathbf{x})P(Y = 2) \end{aligned}$$

Spezialfälle:

- **Gleiche Kosten von Fehlklassifikationen**
  $c(\hat{Y} = 2|Y = 1) = c(\hat{Y} = 1|Y = 2)$:

$$\begin{aligned} R_1: \ f_1(\mathbf{x})P(Y = 1) &\geq f_2(\mathbf{x})P(Y = 2) \\ R_2: \ f_1(\mathbf{x})P(Y = 1) &< f_2(\mathbf{x})P(Y = 2) \end{aligned}$$

- **Maximum-Likelihood-Regel**
  $P(Y=1) = P(Y=2)$ und $c(\hat{Y}=1|Y=2) = c(\hat{Y}=2|Y=1)$:

$$\begin{aligned} R_1 : f_1(\mathbf{x}) &\geq f_2(\mathbf{x}) \\ R_2 : f_1(\mathbf{x}) &< f_2(\mathbf{x}) \end{aligned}$$

Bayes-Regel unter Normalverteilungsannahmen

Fall zweier Gruppen mit **gleichen Kovarianzmatrizen**,
$\mathbf{X}|\{Y=j\} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$:

$$\begin{aligned} f_j(\mathbf{x}) &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right\}, \quad j=1,2 \\ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} \\ &= \exp\left\{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\} \\ &= \exp\left\{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]\right\} \end{aligned}$$

Klassifikationsregel:

$$R_1 : (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right] \geq \ln\left[\frac{c(\hat{Y}=1|Y=2)}{c(\hat{Y}=2|Y=1)}\frac{P(Y=2)}{P(Y=1)}\right]$$

Spezialfall ML-Regel:

Für $P(Y=1) = P(Y=2)$ und $c(\hat{Y}=1|Y=2) = c(\hat{Y}=2|Y=1)$:

$$R_1 : (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right] \geq 0$$

$$\iff (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} \geq (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\left[\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]$$

Maximum Likelihood Ansatz führt unter Annahme multivariater Normalverteilung von $\mathbf{X}$ mit gleichen Kovarianzmatrizen zur linearen Diskriminanzanalyse von Fisher (LDA). Vorige Herleitung über Projektionen benötigt keine Verteilungsannahmen, man kann von der LDA daher auch ohne Normalverteilung gute Ergebnisse erhoffen.

Verallgemeinerungen:
ML-Diskriminanzregel bei k normalverteilten Gruppen mit gleicher Kovarianzmatrix:

$$R_j : (\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)' = \min_{i=1,\ldots,k} (\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)'$$

Quadratische Diskriminanzanalyse (QDA):
Fall zweier Normalverteilungen mit verschiedenen Kovarianzmatrizen, $\mathbf{X}|\{Y = j\} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$:

$$\begin{aligned}
f_j(\mathbf{x}) &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}_j|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\right\} \\
\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{(2\pi)^{-p/2}|\boldsymbol{\Sigma}_1|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right\}}{(2\pi)^{-p/2}|\boldsymbol{\Sigma}_2|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right\}} \\
&= \exp\left\{-\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}\right\} \\
&\quad \cdot \frac{|\boldsymbol{\Sigma}_1|^{-1/2}}{|\boldsymbol{\Sigma}_2|^{-1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right\} \\
R_1 : &\quad \left\{\mathbf{x} \in \mathbb{R}^p : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(\hat{Y} = 1|Y = 2)}{c(\hat{Y} = 2|Y = 1)} \frac{P(Y = 2)}{P(Y = 1)}\right\}
\end{aligned}$$

9.3 Bewertung der Klassifikationsgüte

Zur Bewertung, wie gut die durch die Diskriminanzregel erreichte Trennung ist, gibt es verschiedene Methoden.

1. **Analytische Berechnung der Fehlklassifikationsraten:**
   Für ML Regel bei zwei Klassen mit $\mathbf{X}|\{Y = j\} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$
   Mit $\mathbf{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ und $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ gilt

   $$\mathbf{a}'\left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] \Bigg| \{Y = 2\} \sim N(-0.5\delta^2, \delta^2)$$

   und somit

   $$\begin{aligned} P(\hat{Y} = 1|Y = 2) &= P\left(\mathbf{a}'\left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] > 0 \Bigg| Y = 2\right) \\ &= 1 - \Phi(0.5\delta^2/\delta) \\ &= \Phi(-0.5\delta) \;=\; P(\hat{Y} = 2|Y = 1) \end{aligned}$$

   was aus den Daten geschätzt werden kann.

2. **In sample Fehlklassifikationsraten (apparent error rate):**
   Diskriminanzregel auf die Objekte in der Lernstichprobe anwenden und die Anzahl der Falschklassifikationen bestimmen. Die entstehenden Fehlklassifikationsraten geben einen ersten Hinweis auf die Trennungsgüte, sind aber zu optimistisch.

3. **Out of sample Fehlklassifikationsraten:**
   Bei vielen Beobachtungen mit bekannter Klassenzugehörigkeit kann die Menge dieser Objekte in eine Lern- und eine Validierungsstichprobe unterteilt werden. Die Diskriminanzregel wird anhand der Lernstichprobe und die Fehlklassifikationsraten anhand der Validierungsstichprobe ermittelt.

4. **Kreuzvalidierung (cross validation, leave one out):**
   Unter Auslassung einer Beobachtung aus dem Datensatz wird die Diskriminanzfunktion auf Basis der restlichen $n-1$ Beobachtungen bestimmt und die ausgelassene Beobachtung mit dieser Diskriminanzfunktion klassifiziert. Dies wird sukzessive für alle Beobachtungen durchgeführt und die Fehlklassifikationsraten als Gütekriterium berechnet.

Ziel der Diskriminanzanalyse: möglichst gute Trennung der Klassen auf Basis der beobachteten Merkmale. Ein weiteres Gütekriterium für eine Diskriminanzregel ist daher, dass sie eine bessere Zuordnung zu den Klassen vornimmt als eine rein zufällige Zuordnung.

Bemerkung 9.8. (Test gegen zufällige Zuordnung für LDA)
Im Modell der Diskriminanzanalyse gemäß Definitionen 9.2 und 9.5 werde unterstellt, dass die multivariate Variable $\mathbf{X}$ einer p-dimensionalen Normalverteilung folgt. Dann ist ein Test für das Problem

$$\begin{aligned} & H_0: && \text{keine der betrachteten Variablen verbessert die Klassifikation} \\ & && \text{im Vergleich zu einer zufallsbasierten Zuordnung} \\ \textit{vs.}\quad & H_1: && \text{mindestens eine Variable verbessert die Klassifikation} \end{aligned}$$

gegeben durch folgende Entscheidungsregel (mit der Notation aus Abschnitt 9.1): H_0 wird zum Niveau α verworfen, falls

$$\frac{n-p-1}{p} \cdot \frac{n_1 \cdot n_2}{n} \cdot \frac{\left(\overline{\alpha}^{C_1} - \overline{\alpha}^{C_2}\right)^2}{WSS(\alpha)} > F_{p,n-p-1;1-\alpha},$$

wobei n_1, n_2 Stichprobenumfänge zu den Gruppen C_1 und C_2 und

$$WSS(\alpha) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\alpha_i^{C_j} - \overline{\alpha}^{C_j})^2$$

die Within-Group-Sum-of-Squares der projizierte Werte α_i ist.

Dies entspricht dem F-Test im Varianzanalysemodell der einfachen Varianzanalyse (H_0: Die Mittelwerte der p Variablen sind in beiden Gruppen gleich).

Bemerkung 9.9.

- **Variablenselektion:** Vergleichbar zur Regressionsanalyse werden die Variablen nicht alle in einem Schritt zur Klassifikation herangezogen, sondern sukzessive aufgenommen oder entfernt, wobei jedesmal getestet wird, ob die Hinzunahme / Entfernung die Klassifikation verbessert. Dies erlaubt Identifikation "nutzloser" Variablen, die für neue Objekte nicht mehr erhoben werden müssen.
- **Andere Variablentypen:** Falls die beobachteten Merkmale nur ordinale oder nominale Skalierung besitzen, gibt es Verfahren der Diskriminanzanalyse, die auf entsprechenden Modellannahmen beruhen. Man verwendet hierzu ein so genanntes Multinomialmodell.