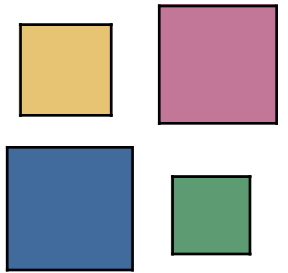


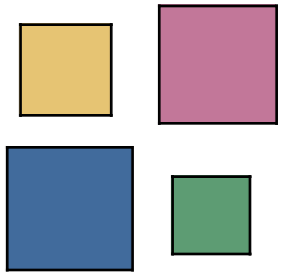
Klassifikation

- Grundsätzlich kann man die Komplexität eines Datensatzes über eine Reduktion der Variablen (*alle bisherigen Verfahren*), oder eine Reduktion der Beobachtungen erreichen.
- Die Reduktion der Information der Beobachtungen geschieht meist durch die Klassifikation der Daten in verschiedene Gruppen.
- Bei den Klassifikationen unterscheidet man **supervised** (überwachtes), und **unsupervised** (nicht-überwachtes) Lernen.
- Supervised Learning kennt die Gruppen a-priori, und sucht eine möglichst fehlerfreie Darstellung. Dazu gehören z.B.
 - Lineare und quadratische Diskriminanz Analyse (LDA, QDA)
 - Klassifikations- (und Regressions) Bäume (CART)
 - Neuronale Netze
- Unsupervised Learning versucht Gruppen in den Daten zu finden
 - Multi Dimensionale Skalierung (MDS)
 - Cluster Analyse



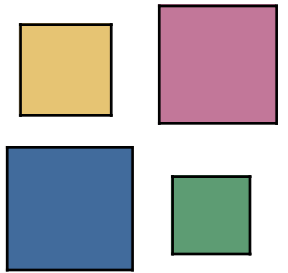
Klassifikation: Anwendungen

- Kundensegmentierung
 - Welche Kunden Typen habe ich in der Datenbank?
 - Woran erkenne ich profitable Kunden -> gezielt bewerben
 - Kreditrisiko
- Gesundheitsrisiko
 - welche Behandlung ist nötig? / Überlebenschancen?
- Hurricanes
 - wird Land getroffen?
- Fußballergebnisse ?
- Rasterfahndung
- Immer qualitative Entscheidung keine quantitative



Klassifikation: Graphik

- Gruppen können in verschiedenen Farben und Formen markiert werden.
- Darstellung in Graphiken für multivariate Daten.
- Graphik kann sehr nützlich sein, unnütze Variablen zu einem frühen Zeitpunkt aus einer Analyse zu entfernen.
- Vorteil der Graphik liegt mehr in der Diagnose der Klassifikationsergebnisse und Bewertung und Entdeckung von Anomalien als in der Aufstellung der Klassifikationsregeln selber.
- Viele Klassifikationsmethoden machen keine Verteilungsannahmen, und sind teilweise rein “algorithmisch” motiviert.



Lineare Diskriminanz Analyse (LDA)

LDA basiert auf der Annahme, daß die Daten aus einer multivariaten Normalverteilung mit gleicher Varianz-Covarianz Σ entstammen.

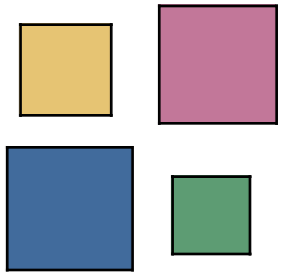
Dadurch reicht es die Dichten der verschiedenen Gruppen zu vergleichen, und die Entscheidungsregel reduziert sich zu:

Ordne eine neue Beobachtung, \mathbf{x}_0 in Gruppe 1 falls

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \mathbf{x}_0 - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq 0$$

sonst gehört sie in Gruppe 2.

Diese Entscheidungsfunktion ist linear in \mathbf{x} .



LDA – Herleitung

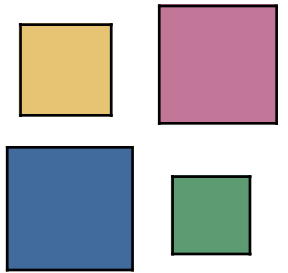
Aus dem Satz von Bayes folgt für die bedingte Wahrscheinlichkeit, dass eine Beobachtung in die 1. Gruppe fällt

$$\Pr(G = 1|X = x) = \frac{f_1(x)\pi_1}{\sum_{i=1}^2 f_i(x)\pi_i},$$

wobei π_i die a-priori Wahrscheinlichkeit ist, zur Klasse i zu gehören.

Unter der Annahme, dass die Daten aus einer multivariaten Normalverteilung mit identischer Varianz Covarianz Matrix stammen gilt für den Log-Ratio

$$\begin{aligned} \log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} &= \log \frac{f_1(x)}{f_2(x)} + \log \frac{\pi_1}{\pi_2} \\ &= \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \\ &\quad + x' \Sigma^{-1}(\mu_1 - \mu_2) \end{aligned}$$



LDA – 1-dimensional

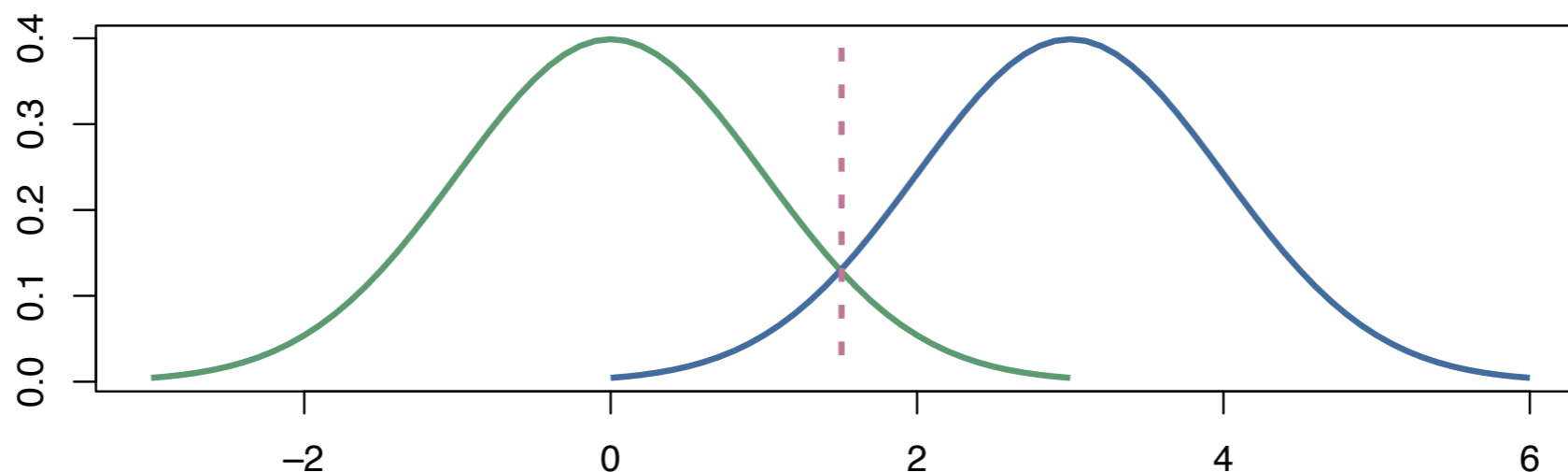
Im 1-dimensionalen Fall reduziert sich die Entscheidungsregel zu

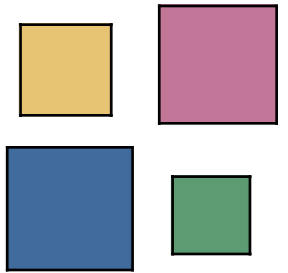
$$x_0 \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma_{pooled}} \geq \frac{1}{2} \frac{(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 + \bar{x}_2)}{\sigma_{pooled}}$$

für $\bar{x}_1 > \bar{x}_2$ also

$$x_0 \geq \frac{1}{2}(\bar{x}_1 + \bar{x}_2)$$

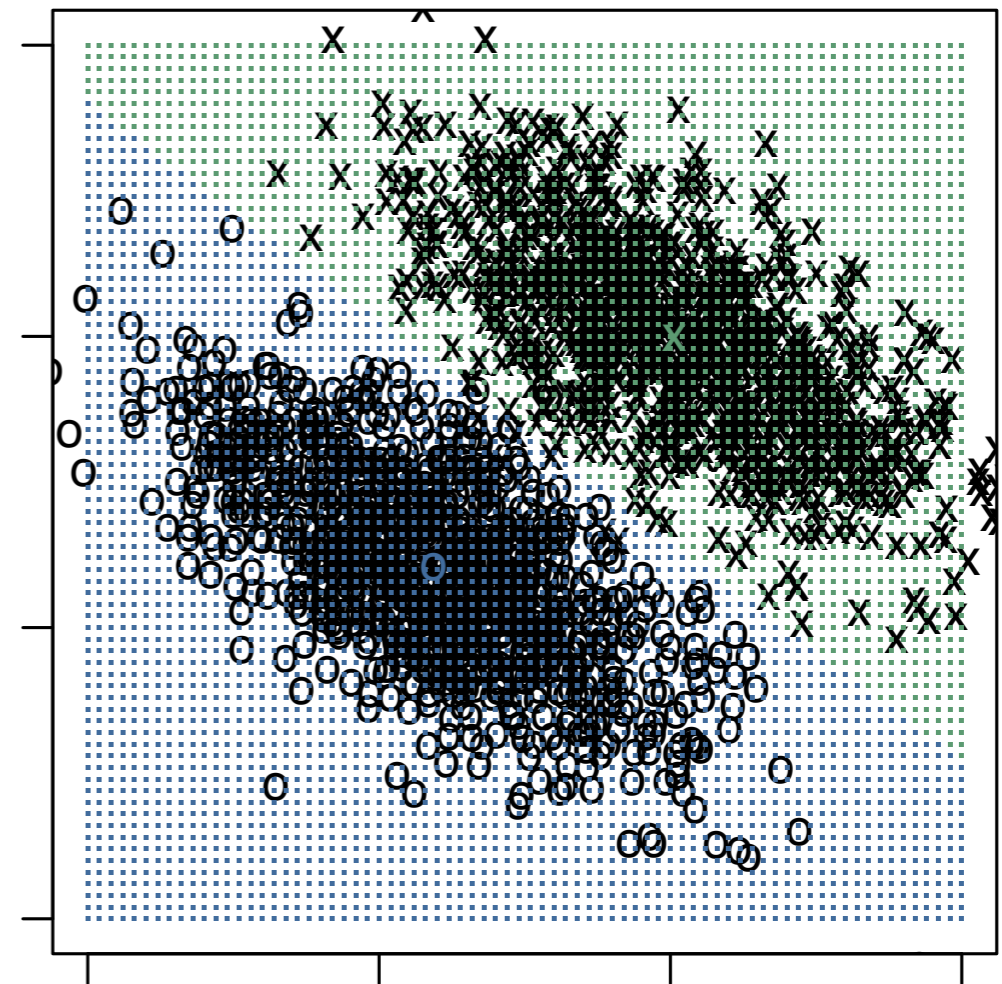
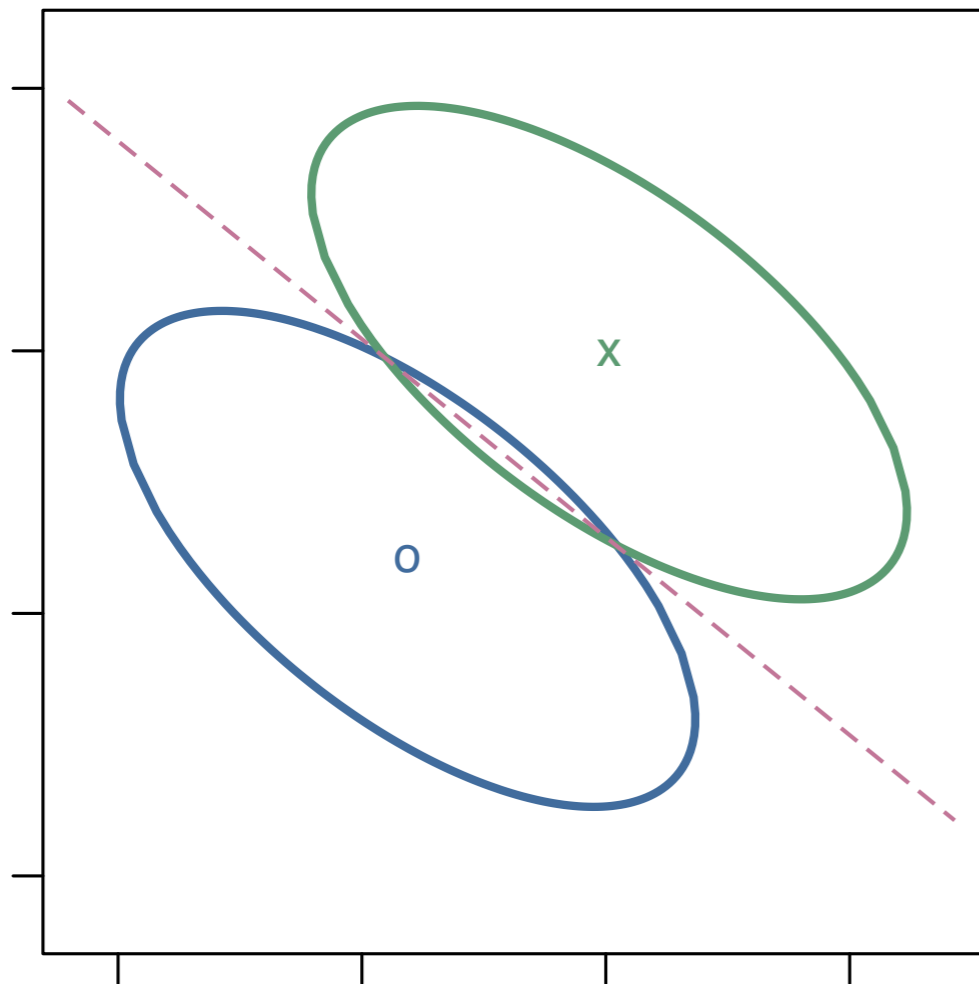
D.h. anhängig davon, ob ein zu klassifizierender Punkt x_0 näher bei \bar{x}_1 oder \bar{x}_2 liegt, wird die neue Beobachtung in die eine oder andere Klasse verteilt.



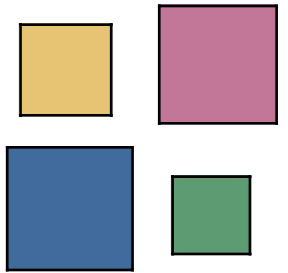


LDA – 2-dimensional

Im 2-dimensionalen Fall reicht die euklidische Distanz zu den beiden Gruppenmitteln \bar{x}_1 und \bar{x}_2 nicht aus um Punkte zu den zwei Gruppen zuzuordnen. Entscheidend ist die Dichte!



Auch hier ergibt sich eine lineare Entscheidungsgrenze (decision boundary).



LDA 2d cont.

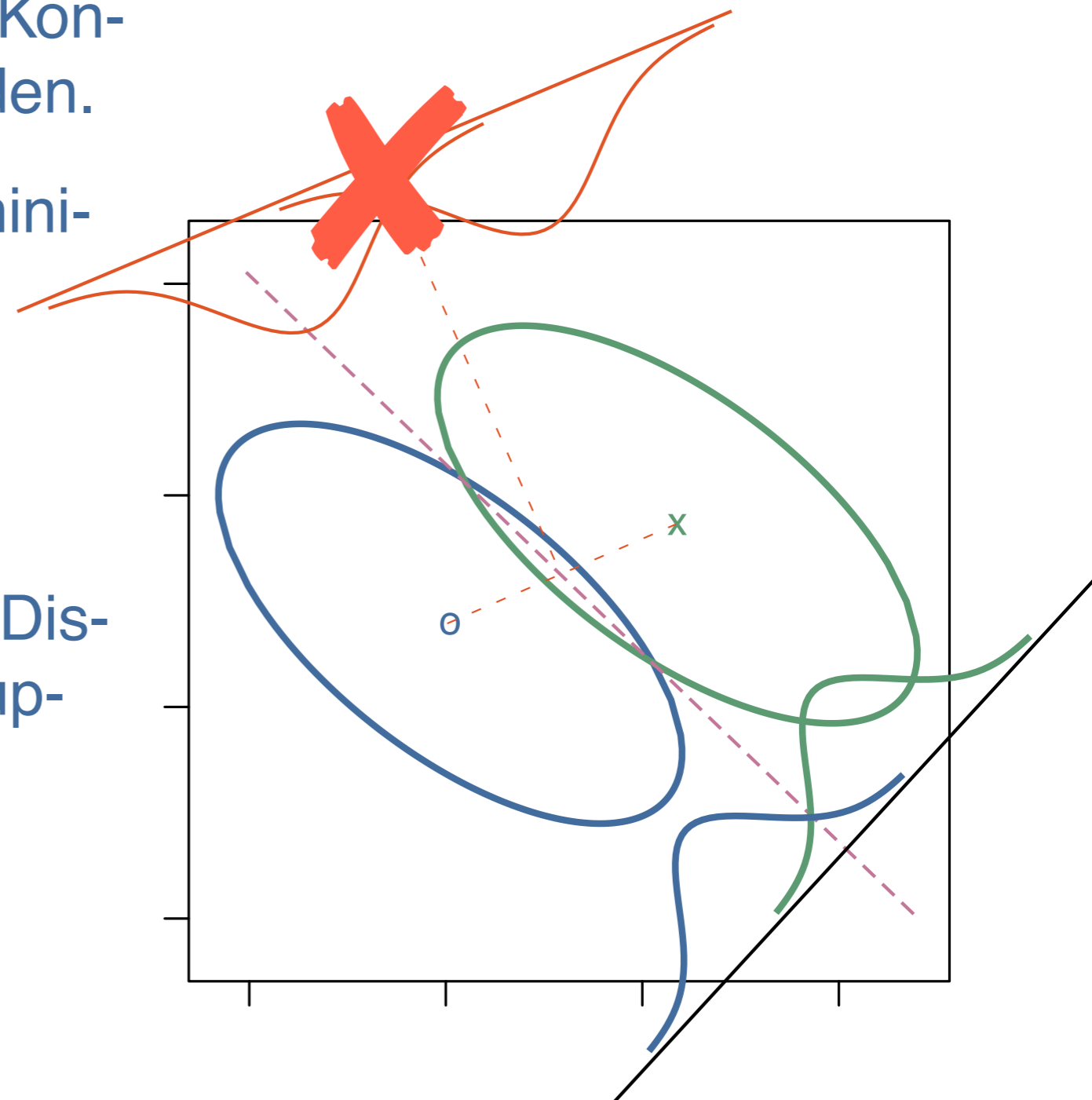
Die Entscheidungsgrenze kann über die Schnittpunkte zweier Konfidenzellipsen konstruiert werden.

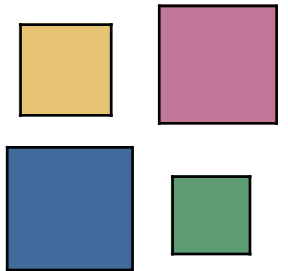
Die Projektion $\hat{a}'\mathbf{x}$ der Daten minimiert die Fehlklassifikation

$$\int_{G_1} f_2(x) dx + \int_{G_2} f_1(x) dx$$

und maximiert die statistische Distanz zwischen den beiden Gruppenmitteln, d.h.

$$\frac{(\hat{a}'\bar{\mathbf{x}}_1 - \hat{a}'\bar{\mathbf{x}}_2)^2}{\hat{a}'\mathbf{S}_{pooled}\hat{a}}$$





Kosten für Fehlklassifikation

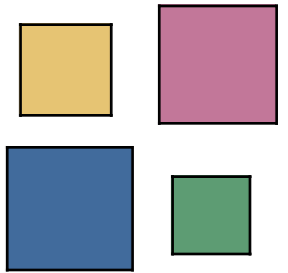
In vielen Fällen werden Fehler bei der Klassifikation nicht symmetrisch behandelt.

Kostenmatrix:

		Klassifikation	
		p_1	p_2
wahre Klasse	p_1	0	$c(2 1)$
	p_2	$c(1 2)$	0

In den meisten Fällen geht man entweder davon aus, dass die a-priori Wahrscheinlichkeit für beide Klassen gleich sind $\pi_1 = \pi_2$, oder man schätzt π_i aus der Größe n_1 und n_2 der Klassen aus der Stichprobe, d.h.

$$\pi_1 = \frac{n_1}{n_1 + n_2},$$
$$\pi_2 = \frac{n_2}{n_1 + n_2}$$



Klassifikation mit Kostenfaktor

Die Entscheidungsregel ändert sich bei Einbeziehung der Kosten entsprechend zu

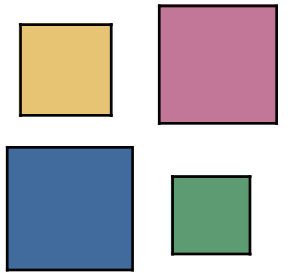
$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} \mathbf{x}_0 - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln \left(\frac{c(1|2) \pi_2}{c(2|1) \pi_1} \right)$$

Die zu erwartenden Kosten einer Klassifikation sind

$$c(2|1) \Pr(2|1) \pi_1 + c(1|2) \Pr(1|2) \pi_2$$

Als Spezialfälle treten auf:

- Gleiche a-priori Wahrscheinlichkeiten: $\pi_1 = \pi_2$
- Gleiche Kosten für Fehlklassifikation: $c(2|1) = c(1|2)$
- Gleiche Kosten und Priors: $\frac{\pi_2}{\pi_1} = \frac{c(1|2)}{c(2|1)} = 1$

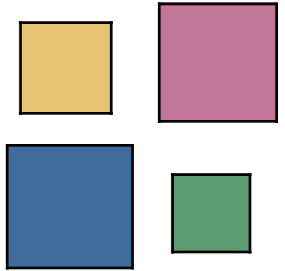


Bemerkungen

- Die Entscheidungsregion (für Gruppe 1) ergeben sich aus

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2) \pi_2}{c(2|1) \pi_1}$$
$$\left(\begin{array}{c} \text{Dichte-} \\ \text{quotient} \end{array} \right) \geq \left(\begin{array}{c} \text{Kosten-} \\ \text{quotient} \end{array} \right) \cdot \left(\begin{array}{c} \text{a-priori-} \\ \text{quotient} \end{array} \right)$$

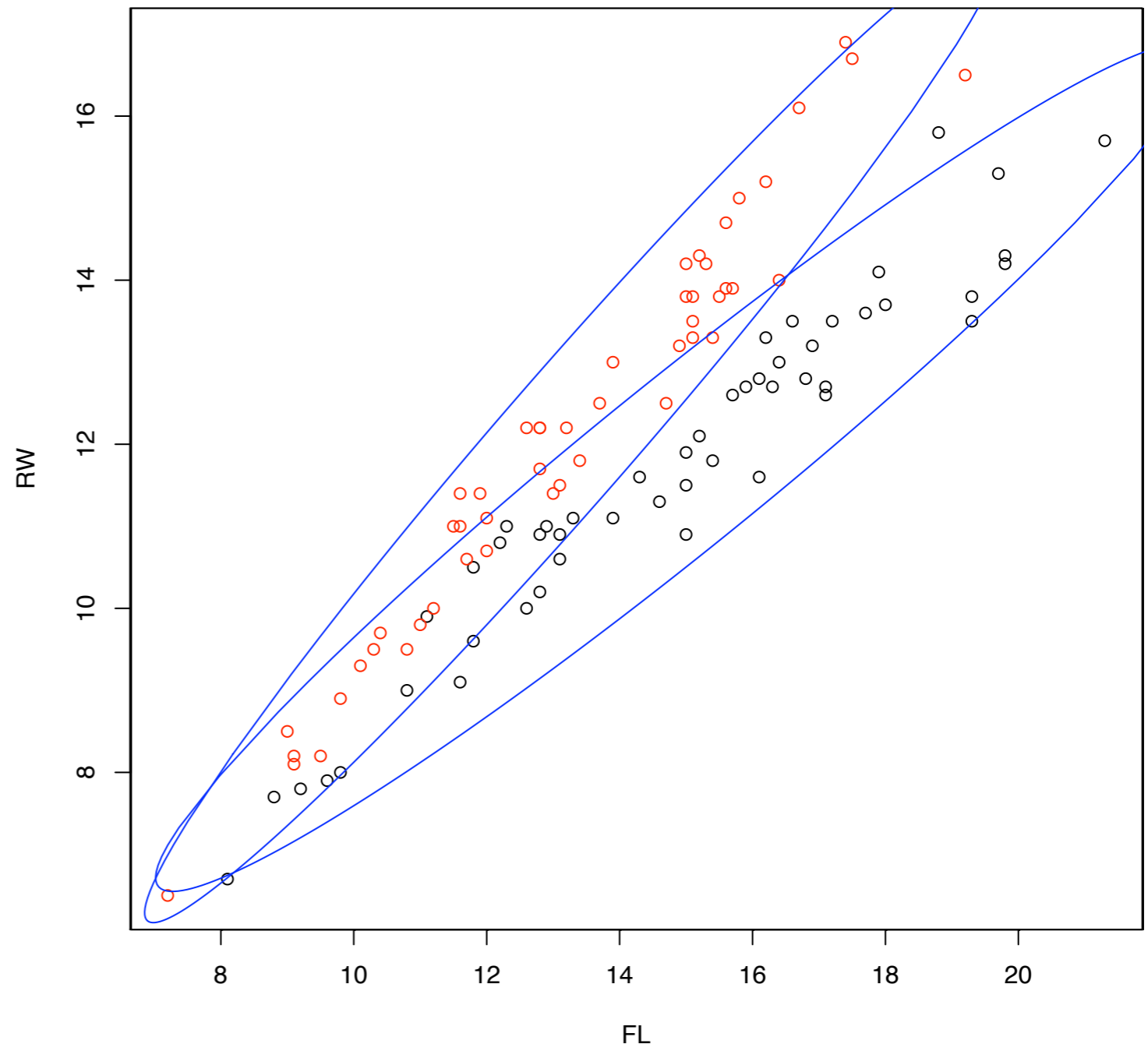
- Strenge Voraussetzungen der LDA sind
 - Die Daten sind multivariat normalverteilt, $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
 - Die Varianz Covarianz Matrix $\boldsymbol{\Sigma}$ beider Gruppen ist identisch
- Die Entscheidungsgrenze (decision boundary) zwischen den beiden Gruppen im p -dim. Raum ist in jedem Falle eine Hyper-ebene der Ordnung $(p - 1)$.

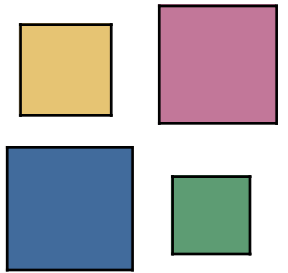


LDA – Beispiel: Australische Krabben

- **Variablen**

- Species (blue, orange)
 - Sex (male, female)
 - Index
 - Frontal Lobe
 - Rear Width
 - Car Length
 - Car Width
 - Body Depth
- 200 Fälle,
je 100 pro Gruppe
und Geschlecht
 - Diskriminierung der
Geschlechter der
blauen Species





Beispiel cont.

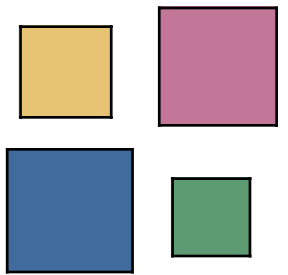
Australische Krabben

Diskiminierung der Geschlechter der blauen Species

$$\begin{aligned}
 \bar{\mathbf{x}}_{Male} &= (14.8 \ 11.7)' & \bar{\mathbf{x}}_{Fem} &= (13.3 \ 12.1)' \\
 n_{Male} &= 50, & n_{Fem} &= 50 \\
 \mathbf{S}_{Male} &= \begin{bmatrix} 10.3 & 6.5 \\ 6.5 & 5.5 \end{bmatrix} & \mathbf{S}_{Fem} &= \begin{bmatrix} 6.9 & 6.3 \\ 6.3 & 5.9 \end{bmatrix}
 \end{aligned}$$

Zusammengefasste Varianz Covarianz

$$\begin{aligned}
 \mathbf{S}_{pooled} &= \frac{(n_1 - 1)\mathbf{S}_1}{(n_1 - 1) + (n_2 - 1)} + \frac{(n_2 - 1)\mathbf{S}_2}{(n_1 - 1) + (n_2 - 1)} \\
 &= \begin{bmatrix} 8.6 & 6.4 \\ 6.4 & 5.2 \end{bmatrix} \\
 \mathbf{S}_{pooled}^{-1} &= \begin{bmatrix} 1.47 & -1.81 \\ -1.81 & 2.42 \end{bmatrix}
 \end{aligned}$$



Beispiel cont.

Erster Teil der Klassifikationsregel,

$$\begin{aligned}
 (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} &= \begin{bmatrix} 1.5 & -0.4 \end{bmatrix} \begin{bmatrix} 1.47 & -1.81 \\ -1.81 & 2.42 \end{bmatrix} \\
 &= \begin{bmatrix} 2.93 \\ -3.68 \end{bmatrix}
 \end{aligned}$$

mit $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' = (14.8 - 13.3 \quad 11.7 - 12.1)$.

Dies definiert einen 1d Vektor (Linie), durch 0, welche die Richtung der maximalen Separation zwischen den 2 Gruppen definiert, die durch die Gleichung $x_2 = \frac{-3.68}{2.93} x_1 = -1.26 x_1$ gegeben ist.

\mathbf{x}_0 , genau wie das Gesamt Mittel $\frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$, werden auf diesen Vektor projiziert.

