Multivariate Analysemethoden

Multivariate Distanz – Multivariate Normalverteilung
Minimum Distance Classifier – Bayes Classifier

Günter Meinhardt
Johannes Gutenberg Universität Mainz

Multivariate Klassifikation

Ziele

- Einordnen von Fällen (Versuchspersonen, Beobachtungen) in Gruppen aufgrund ihrer Werte in mehreren Meßvariablen.
- Maßgeblich für die Zuordnung zu eine Gruppe ist a) die Wahrscheinlichkeit des Auftretens des Falles in der Zielgruppe (falls ermittelbar) oder b) die Distanz des Falles vom charakteristischen Wert der Gruppe (Prototyp, Zentroid)

Methoden

- Deskriptive Methoden:
  - Bestimmung von Distanzen und Wahrscheinlichkeiten auf dem Set der beobachteten Meßvariablen

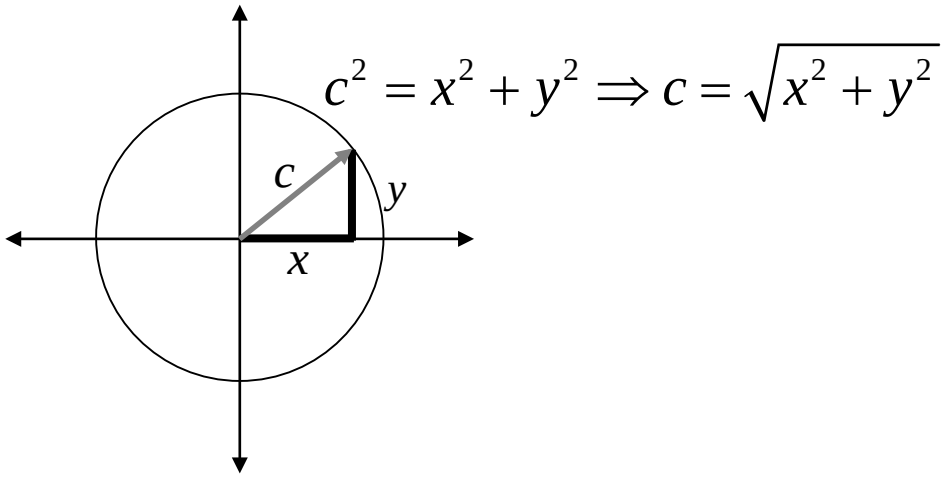
- Analytische Methoden:
  - Bestimmung von Distanzen und Wahrscheinlichkeiten auf transformierten Meßvariablen mit dem Ziel, die Separation von Gruppen zu maximieren (Diskriminanzanalytische Methoden)

- Weitere Kriterien sind Kosten von Fehlklassifikationen und die a-priori Wahrscheinlichkeit von Gruppen (Allg. Likelihood-Ratio und Bayes-Klassifikation)

Kreis

Iso-Distanz-Konturen in 2D

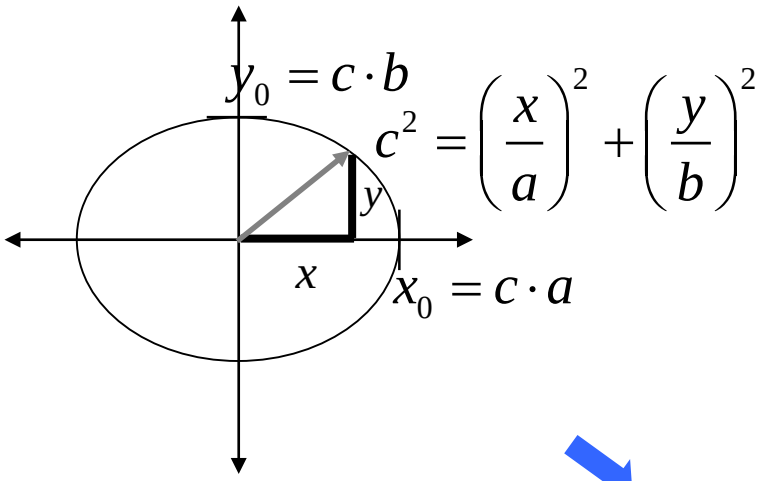
Kreis mit Radius *c*: Alle Punkte auf dem Kreisbogen haben euklidischen Abstand *c* zum Kreismittelpunkt



- Der Kreis ist die Grundform der Iso-Distanz Kontur im zweidimensionalen Raum (p = 2).
- Er entspricht im Variablenraum einer Iso-Distanz-Kontur für 2 unkorrelierte (orthogonale) Variablen mit derselben Skalierung.

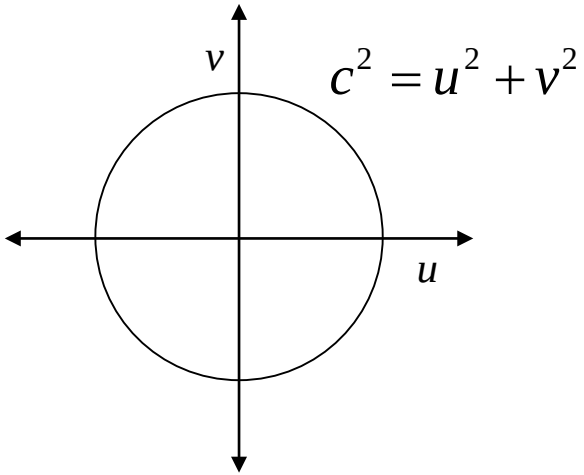
Ellipse: Skalierung

Ellipse mit Ellipsenradius *c*: Alle Punkte auf dem Ellipsenbogen haben, auf Standardskala normiert, denselben Abstand *c* zum Mittelpunkt



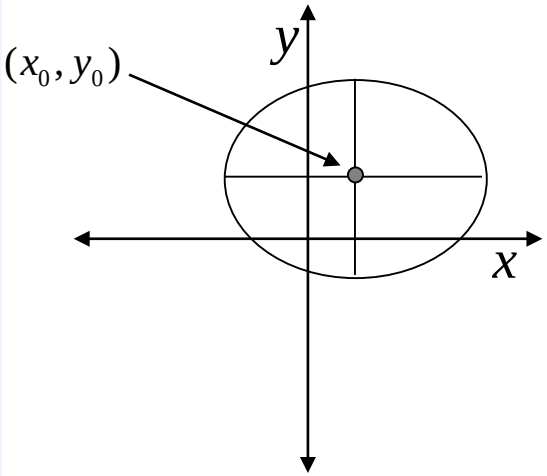
Standardskala:

$$u = \frac{x}{a} \qquad v = \frac{y}{b}$$



Ellipse
Translation

Translation zum Punkt (x_0,y_0) ändert an dieser Eigenschaft nichts:

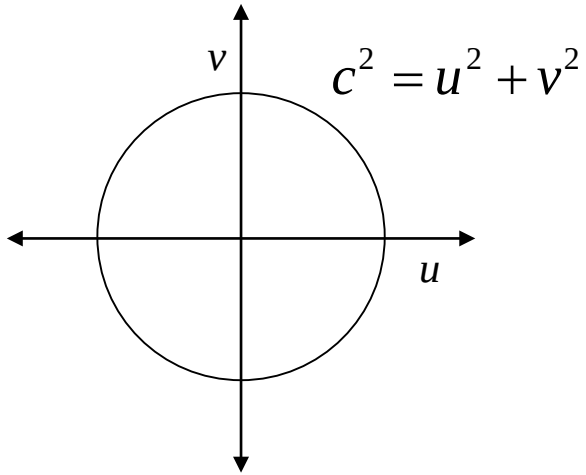


$$c^2 = \left(\frac{x - x_0}{a}\right)^2 + \left(\frac{y - y_0}{b}\right)^2$$

Standard-Transformation

Standardskala:

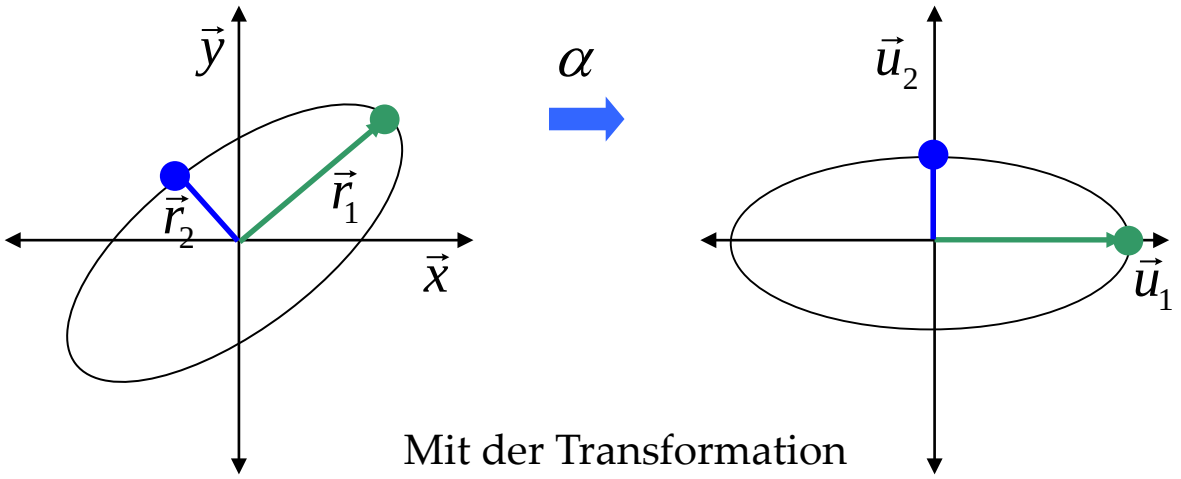
$$u = \frac{x - x_0}{a} \qquad v = \frac{y - y_0}{b}$$



Standard-Ellipse
Neigung
Korrelation ρ

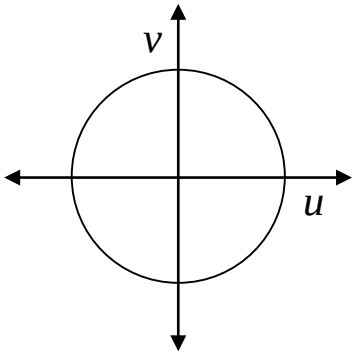
Die Invarianz der Distanz im neuen Koordinatensystem mit geneigten Achsen (Korrelation der Variablen) ist über eine Rotation der Koordinaten (anticlock) erklärt:

$$c^2 = x^2 + y^2 - 2\rho xy \qquad \rho = \cos\alpha$$



Mit der Transformation

$$u = \frac{u_1}{a} \qquad v = \frac{u_2}{b}$$



erfüllen alle Ellipsenpunkte:

$$c^2 = u^2 + v^2$$

Koordinaten Korrelierte Achsen

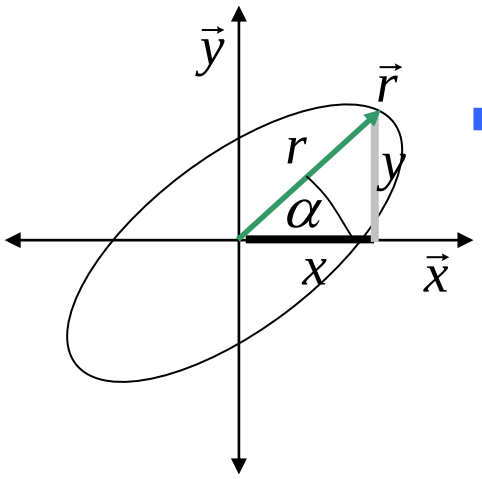
$$c^2 = u^2 + v^2$$

[Tafel: cos α]

Standard-Ellipse:
Zeichen-Routine

Ellipsen sind in kartesischen Koordinaten unpraktisch zu zeichnen.
Man geht über zur Darstellung in Polarkoordinaten.

$$c^2 = x^2 + y^2 - 2\rho xy$$



kartesisch polar

$$r = (x, y) \triangleq (r, \alpha)$$

Es gelten die Transformationen:

polar ⟶ kartesisch

$$x = r\cos\alpha$$
$$y = r\sin\alpha$$

kartesisch ⟶ polar

$$r^2 = x^2 + y^2$$
$$\alpha = \tan^{-1}\left(\frac{y}{x}\right)$$

Zum Zeichnen muß die Ellipsengleichung als Gleichung in Polarkoordinaten (Vektorlänge in Abhängigkeit des Winkels α) umgeschrieben werden

Iso-Distanz Konturen in 2D

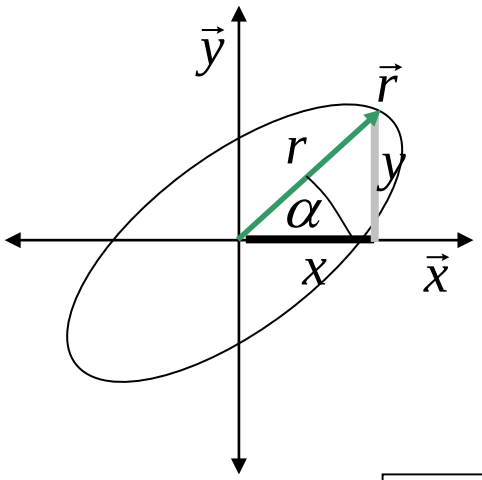
Standard-Ellipse:
Zeichen-Routine

Von der Darstellung in Polarkoordinaten kann einfach in kartesische Koordinaten zurückgerechnet werden (Setzen der Ellipsenpunkte)

$$c^2 = x^2 + y^2 - 2\rho xy$$

Setze $q = \frac{y}{x}$

damit $c^2 = x^2\left(1 + q^2 - 2\rho q\right)$



$$|x| = \frac{c}{\sqrt{1 + q^2 - 2\rho q}}$$

$$r = \sqrt{x^2 + y^2} = |x|\sqrt{\left(1 + q^2\right)}$$

Verfahren

1. Variiere α von $-\pi$ bis π (= ein Kreisumlauf).
2. Für jeden Winkel α berechne $q = \tan^{-1}(\alpha)$.
3. Berechne dann $|x|$
4. Berechne damit r.
5. Berechne dann x,y: $x = r\cos\alpha$, $y = r\sin\alpha$

[Excel-Sheet]

1 D-Normal Verteilung

Die Funktion $f(x) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ hat Fläche $\sqrt{2\pi}\sigma$

Die auf die Fläche 1 normierte Funktion

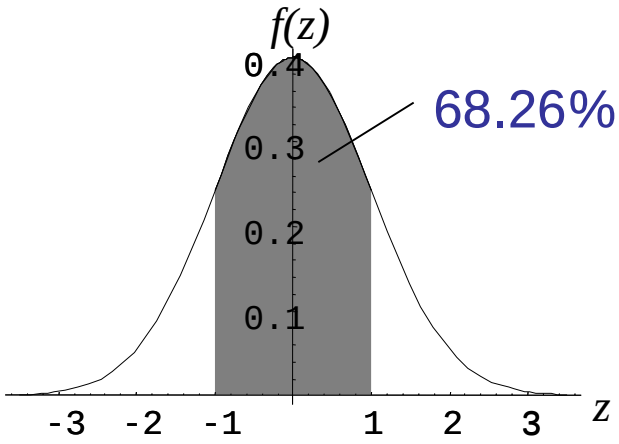
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

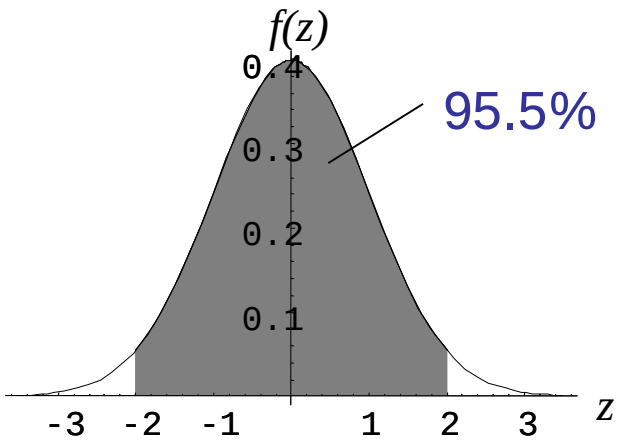
$$z = \frac{x-\mu}{\sigma}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

heißt Normalverteilung (Gauss-Verteilung). Mit ihr sind Wahrscheinlichkeiten als Flächen-Anteile für z - Standardvariablen definierbar.

(Standard-NV)





[Kurzübung]

p-variater Fall

Man bemerke daß $z^2 = \left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)\frac{1}{\sigma^2}(x-\mu)$ ist.

Man habe nun nicht eine, sondern m Variablen:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \text{ mit Zentroid } \vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}$$

(jeder Messpunkt ist ein m-dimensionaler Vektor und der Zentroid ist ein m-dimensionaler Vektor)

Mahalanobis-Distanz

Dann definiert

$$\Delta^2 = (\vec{x}-\vec{\mu})^t \, \mathbf{\Sigma}^{-1} (\vec{x}-\vec{\mu})$$

mit $\mathbf{\Sigma}^{-1}$ die Inverse der Varianz-Kovarianz Matrix $\mathbf{\Sigma}$.

die verallgemeinerte quadrierte Distanz im multivariaten Raum.
Sie heißt quadrierte Mahalanobis-Distanz.

[Excel-Beispiel 2D]

$$\vec{x} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \vec{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad \Delta^2 = 4$$

p D-Normal Verteilung

Die Funktion $f\left(\vec{x}\right) = e^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\vec{x}-\vec{\mu}\right)}$ hat Volumen $\left(2\pi\right)^{m/2}\left|\boldsymbol{\Sigma}\right|^{1/2}$

Die auf Volumen 1 normierte Funktion

$$f\left(\vec{x}\right) = \frac{1}{\left(2\pi\right)^{m/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

heißt multivariate Normalverteilung (multivariate Gauss-Verteilung). Mit ihr sind Wahrscheinlichkeiten als Anteile des Gesamtvolumens eines m-dimensionalen Ellipsoids definiert.

Die in ihrem Argument auftretende Mahalanobis-Distanz erfüllt die Bedingung:

$$\Delta^2 = \left(\vec{x}-\vec{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\vec{x}-\vec{\mu}\right) \leq \chi^2_m\left(\alpha\right)$$

mit α einem zu setzenden alpha-Fehler Niveau.

Alle Mahalanobisdistanzen Δ, die diese Bedingung erfüllen, erzeugen Konturen gleicher Wahrscheinlichkeit (iso-probability contours) mit P = 1- α in der multivariaten Normalverteilung.

2 D-Normal Verteilung

Die multivariate Normalverteilung mit m = 2 Variablen (bivariate Normalverteilung) hat die Form

$$f\left(x_1, x_2\right) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}} e^{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]}$$

Die im Argument auftretende Mahalanobis-Distanz definiert eine Ellipse im zweidimensionalen Raum für jede Konstante *c*:

$$c^2 = \Delta^2 = \left(\vec{x}-\vec{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\vec{x}-\vec{\mu}\right)$$
$$= \frac{1}{\left(1-\rho^2\right)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$

Diese ist eine Iso-Probability-Contour im obigen Sinne (s. multivariate NV, vorherige Folie)

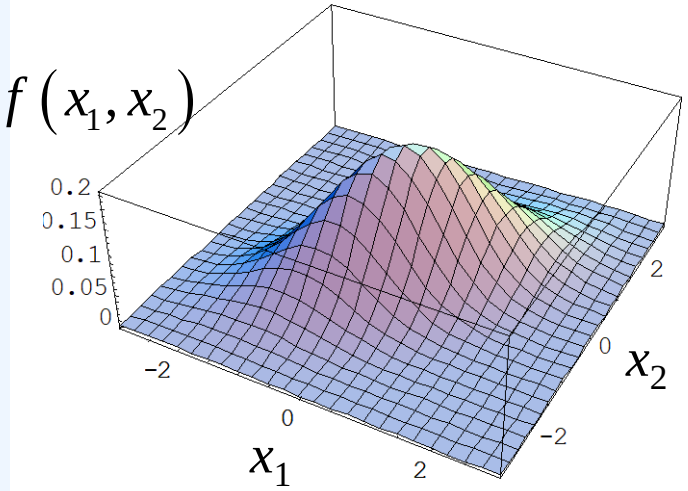
[Tafelbetrachtung]

2 D-Normal Verteilung

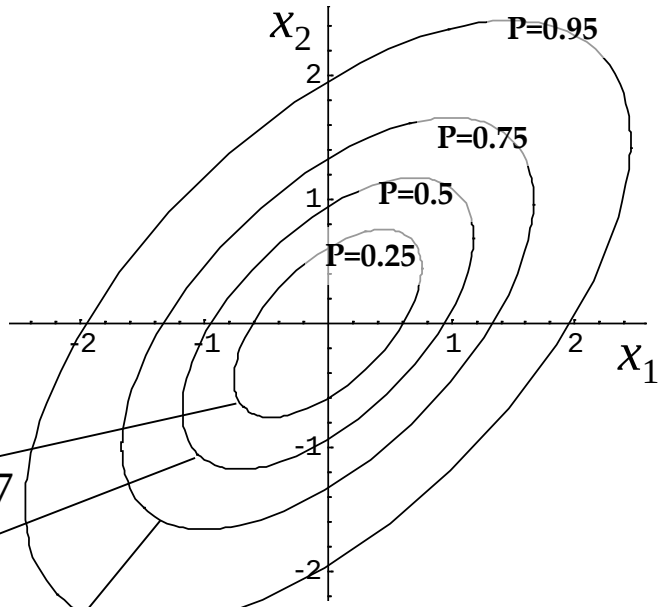
Bivariate Normalverteilung mit m = 2 Variablen und Korrelation $r = 0.6$

Density-Plot

$$\vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$



Contour-Plot



$$\Delta^2 = \chi_2^2(0.75) = 0.57$$

$$\Delta^2 = \chi_2^2(0.5) = 1.38$$

$$\Delta^2 = \chi_2^2(0.05) = 5.99$$

$$\Delta^2 = \chi_2^2(0.25) = 2.77$$

Ellipsen gleicher Wahrscheinlichkeit und zugehöriges Distanzmaß (quadrierte Mahalanobis-Distanz)

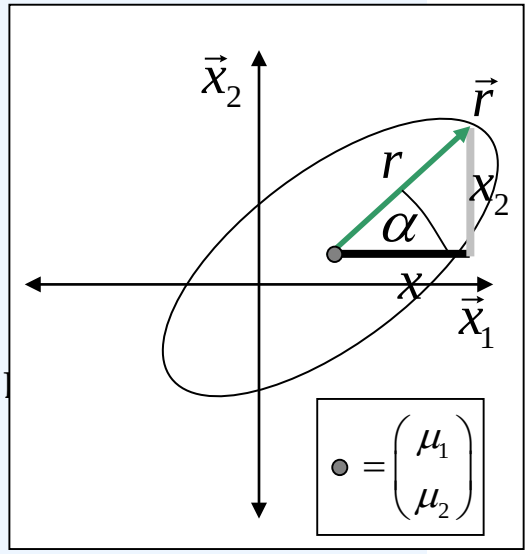
[Excel-Übung]

Iso-Distanz Konturen in 2D

NV-2D-Ellipse: Zeichen-Routine

$$c^2 = \frac{1}{(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$ **(NV-Ellipse)**

Setze $q = \frac{y}{x}$ **und temporär** $\mu_1 = 0, \mu_2 = 0$



$$c^2 = \frac{1}{1-\rho^2}\left(\left(\frac{x}{\sigma_1}\right)^2 + \left(\frac{qx}{\sigma_2}\right)^2 - 2\rho\frac{x}{\sigma_1}\frac{qx}{\sigma_2}\right)$$

$$c^2 = \frac{1}{1-\rho^2}x^2\left(\frac{1}{\sigma_1^2} + \frac{q^2}{\sigma_2^2} - 2\frac{\rho q}{\sigma_1\sigma_2}\right)$$

$$|x| = \frac{c}{\sqrt{\frac{1}{1-\rho^2}\left(\frac{1}{\sigma_1^2} + \frac{q^2}{\sigma_2^2} - 2\frac{\rho q}{\sigma_1\sigma_2}\right)}}$$

Verfahren

[Excel-Sheet]

Und es gilt:

a) α läuft von –π bis π (= ein Kreisumlauf)

b) $r = \sqrt{x_1^2 + x_2^2} = |x|\sqrt{1+q^2}$

c) $x_1 = r\cos\alpha + \mu_1$

$x_2 = r\sin\alpha + \mu_2$

m-dim-Normal Verteilung

Die Ellipsen der Form

$$c^2 = (\vec{x} - \vec{\mu})^t \, \boldsymbol{\Sigma}^{-1} (\vec{x} - \vec{\mu}) \leq \chi^2_m(\alpha)$$

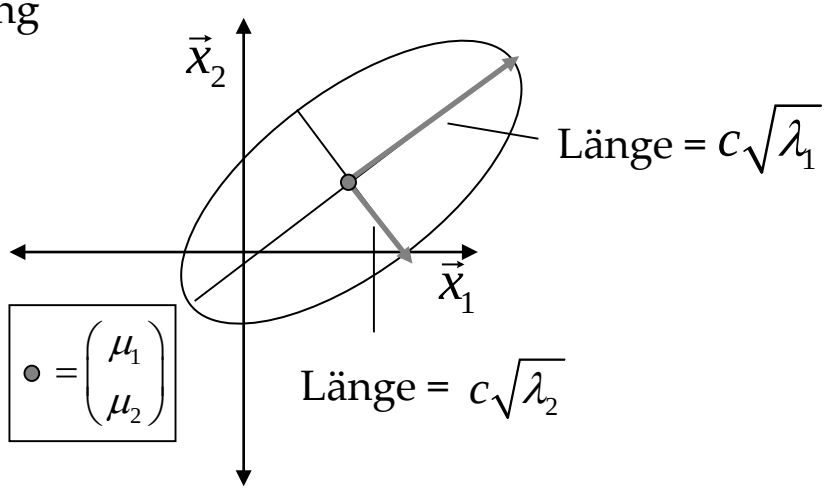
sind zentriert in $\vec{\mu}$ und haben Hauptachsen $\pm c\sqrt{\lambda_i}\vec{e}_i$

mit Eigenwertbedingung

$$\boldsymbol{\Sigma}\vec{e}_i = \lambda_i \vec{e}$$

Eine Eigenwertzerlegung der Varianz-Kovarianz Matrix liefert somit die Hauptachsen des m-variaten Ellipsoids der multivariaten Normalverteilung

Beispiel 2D



MDC

Mit der Mahalanobisdistanz für eine Beobachtung $\vec{x}$ zum Zentroid der Gruppe c_j

$$\Delta_j^2 = \left(\vec{x} - \vec{\mu}_j\right)^t \boldsymbol{\Sigma}_j^{-1} \left(\vec{x} - \vec{\mu}_j\right)$$

definiere die Regel:
Gruppiere $\vec{x}$ in Gruppe c_i, wenn gilt

MDC-Regel

$$\Delta_i^2 = \min\left(\Delta_1^2, \Delta_2^2, \ldots, \Delta_j^2, \ldots \Delta_k^2\right)$$

Die Performance des MDC läßt sich mit großen Stichproben für die k – Gruppen mit einer Konfusions-Matrix bewerten:

Häufigkeit zur Einordnung von Fall (Zeile) in Gruppe (Spalte)

allocated to group

Case $\vec{x}$ is group	c_1	c_2	…	c_j	…	c_k
c_1	h_{11}	h_{12}	…	h_{1j}	…	h_{1k}
c_2	h_{21}	h_{22}	…	h_{2j}	…	h_{2k}
⋮	⋮	⋮	⋱	⋮	⋮	⋮
c_j	h_{j1}	h_{j2}	…	h_{jj}	…	h_{jk}
⋮	⋮	⋮	⋮	⋮	⋱	⋮
c_k	h_{k1}	h_{k1}	…	h_{kj}	…	h_{kk}

Confusion-matrix

	c_1	c_2	…	c_j	…	c_k
c_1	h_{11}	h_{12}	…	h_{1j}	…	h_{1k}
c_2	h_{21}	h_{22}	…	h_{2j}	…	h_{2k}
⋮	⋮	⋮	⋱	⋮	⋮	⋮
c_j	h_{j1}	h_{j2}	…	h_{jj}	…	h_{jk}
⋮	⋮	⋮	⋮	⋮	⋱	⋮
c_k	h_{k1}	h_{k1}	…	h_{kj}	…	h_{kk}

Hits

Korrekte Klassifizierungen sind die Häufigkeiten auf der Diagonalen:

$$h_o = \sum_{i=1}^{k} h_{ii}$$

Erwartete Häufigkeiten bei Zufall (anteilige Gleichverteilung)

Mit den Zeilensummen h_i und N der Summe aller Häufigkeiten gilt

$$\hat{h}_{ij} = p_j \cdot h_i \quad \text{(erwartete Zellhäufigkeit)}$$

mit p_j der A-priori Wahrscheinlichkeit der Gruppe c_j

p_j kann ggf. aus den empirischen Gruppenstärken über $p_j = h_i/N$ geschätzt werden, wenn keine Information über die A-priori Wahrscheinlichkeiten vorliegt.

Minimum Distance Classifier

Erwartete Confusion-matrix

Hits

Erwartete Häufigkeiten bei Zufall (anteilige Gleichverteilung)

Dann ist

$$\hat{h} = \sum_{i=1}^{k} \hat{h}_{ii}$$

die erwartete Hit-Häufigkeit.

	c_1	c_2	…	c_j	…	c_k
c_1	$\hat{h}_{11}$	$\hat{h}_{12}$	…	$\hat{h}_{1j}$	…	$\hat{h}_{1k}$
c_2	$\hat{h}_{21}$	$\hat{h}_{22}$	…	$\hat{h}_{2j}$	…	$\hat{h}_{2k}$
⋮	⋮	⋮	⋱	⋮	⋮	⋮
c_i	$\hat{h}_{i1}$	$\hat{h}_{i2}$	…	$\hat{h}_{ij}$	…	$\hat{h}_{ik}$
⋮	⋮	⋮	⋮	⋮	⋱	⋮
c_k	$\hat{h}_{k1}$	$\hat{h}_{k1}$	…	$\hat{h}_{kj}$	…	$\hat{h}_{kk}$

Mit

$$p_e = \frac{\hat{h}}{N} \qquad \sigma^2 = N \cdot p_e \cdot (1 - p_e)$$

ist h_o normalverteilt über die Approximation der Binomialverteilung

wenn $N \cdot p_e \cdot (1 - p_e) > 9$ gilt.

Dann testet der z- Test

$$z = \frac{h_o - N p_e}{\sqrt{N \cdot p_e \cdot (1 - p_e)}}$$

die Hitrate des MDC gegen den Zufall.

A-priori Wahrscheinlichkeit der Gruppen

A-posteriori WK

Max-Aposteriori WKn Classifier

Normalverteilungsannahme

Man habe Information über die A-priori Wahrscheinlichkeiten der Gruppen c_j:

$$P(c_1), P(c_2), \ldots, P(c_j), \ldots P(c_k)$$

Dann liefert eine Klassifikation der Beobachtung $\vec{x}$ nach ihrer A-posteriori Wahrscheinlichkeit

$$P(c_j \mid \vec{x})$$

eine korrektere Zuordnung als nur nach der kürzesten Distanz zum Gruppenzentroid.

Regel:
Gruppiere $\vec{x}$ in Gruppe c_i, wenn gilt

$$P(c_i \mid \vec{x}) = \max\left(P(c_2 \mid \vec{x}), P(c_2 \mid \vec{x}), \ldots, P(c_j \mid \vec{x}), \ldots P(c_k \mid \vec{x})\right)$$

Um die A-posteriori WKn zu berechnen, muss für die Likelihood-Funktionen die Annahme der multivariaten Normalverteilung gelten.

Likelihoods

A-posteriori WK

Klassifikations-Raum

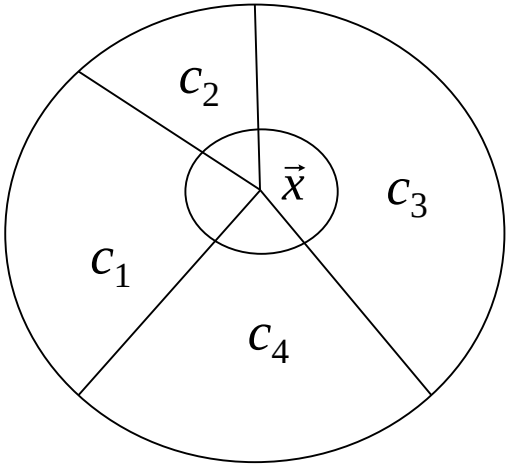
Normalverteil-ungsannahme

Mit der multivariaten Normalverteilung haben die Likelihoods die Form

$$P\left(\vec{x}\middle|c_j\right) = f\left(\vec{x}\middle|c_j\right) = \frac{1}{\left(2\pi\right)^{m/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} e^{-\frac{1}{2}\Delta_j^2}$$

mit $$\Delta_j^2 = \left(\vec{x}-\vec{\mu}_j\right)^t \boldsymbol{\Sigma}_j^{-1}\left(\vec{x}-\vec{\mu}_j\right)$$

der quadrierten Mahalanobisdistanz zum Gruppenzentroid $\vec{\mu}_j$



Der Klassifikationsraum ist durch alle Gruppen vollständig partitioniert.

Es gilt:

$$\vec{x} = \left(\vec{x}\cap c_1\right)\cup\left(\vec{x}\cap c_2\right)\cup\ldots\cup\left(\vec{x}\cap c_k\right)$$

Und wegen der Disjunktheit:

$$P\left(\vec{x}\right) = P\left(\vec{x}\cap c_1\right) + P\left(\vec{x}\cap c_2\right) + \ldots + P\left(\vec{x}\cap c_k\right)$$

Likelihoods

Da

$$P\left(\vec{x}\middle|c_j\right)=\frac{P\left(\vec{x}\cap c_j\right)}{P\left(c_j\right)}$$

(Def. der bedingten Wahrscheinlichkeit), folgt

Satz der totalen WK

$$P\left(\vec{x}\right)=P\left(\vec{x}\middle|c_1\right)P\left(c_1\right)+P\left(\vec{x}\middle|c_2\right)P\left(c_2\right)+\ldots+P\left(\vec{x}\middle|c_k\right)P\left(c_k\right)$$

Und damit

Satz von Bayes

$$P\left(c_i\middle|\vec{x}\right)=\frac{P\left(\vec{x}\middle|c_i\right)P\left(c_i\right)}{\sum_j P\left(\vec{x}\middle|c_j\right)P\left(c_j\right)}$$

der Satz von Bayes für die A-posteriori WK der Gruppe c_i, gegeben die multivariate Beobachtung

Normalverteilungsannahme

Die approximative Gültigkeit der multivariaten NV kann durch Q-Q-Plot Methoden überprüft werden.