CI-Vorlesung 15.05.2012

Überwachtes Lernen: $X = \{ \langle \bar{x}^1, t^1 \rangle \ldots \langle x^N, t^N \rangle \}$ (Target → t)

↳ NN

SVM

$\boxed{\text{Bayes Klassifikator}}$

N … # samples

$t^n \in \mathbb{N} \Rightarrow$ Klassifikation

$t^n \in \mathbb{R} \Rightarrow$ Regression

Unüberwachtes Lernen: $X = \{ \bar{x}^1, \ldots, \bar{x}^N \}$

↳ •) Explorative Datenanalyse: Haupt Komponenten t. (PCA)

•) $\boxed{\text{Schätzung von Wahrschk. Verteilungen (ML, MAP)}}$

•) ….

(Reinforcement Learning:)

•) in dieser Vorlesung

$$P(X = x) = \lim_{N \to \infty} \frac{n_x}{N}$$

N… # Experimente

n_x… # $X = x$

$$1 \geq P(x) \geq 0$$

$$\sum_{x=1}^{|X|} P(X=x) = 1 \Rightarrow \text{diskret}$$

$$\int P(x)\,dx = 1 \Rightarrow \text{kontin.}$$

$f_X(x)$ oder $P(X)$ X… Zufallsvariable

Ist also W! das ZV $X = x$ ist also z.B. 1 gewürfelt

Verbund W! (Joint Prob.)

$$P(X, Y) = P(X|Y) \cdot P(Y) = P(Y|X) \cdot P(X)$$

⇒ Produktregel Kettenregel

$$P(X_1, \ldots, X_N) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_2 X_1) \ldots P(X_N|X_{N-1} \ldots X_1)$$

Bayes : $P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$

$$\left.\begin{aligned} P(x) &= \sum_{y=1}^{|Y|} P(X, Y=y) \\ P(y) &= \sum_{x=1}^{|X|} P(X, Y) \end{aligned}\right\} \text{Summenregel}$$

$$P(X,Y) = P(X|Y) \cdot P(Y)$$

$$\left.\begin{aligned} &= P(X) \cdot P(Y) \\ P(X|Y) &= P(X) \end{aligned}\right\} \text{Wenn } X \text{ statist. unabh. von } Y$$

Schätzen von W! Verteilungen

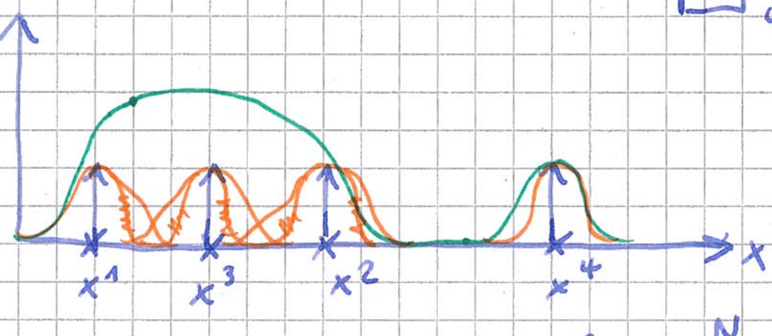
geg.: $X \{\vec{x}^1 \dots \vec{x}^N\}$ $\quad \vec{x}^n \in \mathbb{R}^d$

ges.: Wahrscheinlichkeitsverteilung $P(x)$

1) → parametrisches Modell : Bsp. Gaußverteilung
2) → n. parametrisches Modell

2) Kern basierter Schätzer :

geglättete Dichtefkt.



Empirische Dichtefunktion: $P^e(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x^n)$

↑ Dirac

geglättete Dichtefunktion

$$p^{gl}(x) = h(x) * p^{e}(x)$$

$h(x)$ ... Gauß-Kern
⇒ Glättungskern

$$= \int_{-\infty}^{\infty} h(x-\xi) \cdot p^{e}(\xi) \cdot d\xi$$

$$= \int_{-\infty}^{\infty} h(x-\xi) \frac{1}{N} \sum_{n=1}^{N} \delta(\xi - x^{n})\, d\xi$$

$$= \frac{1}{N} \sum_{n=1}^{N} \int_{-\infty}^{\infty} h(x-\xi) \cdot \delta(\xi - x^{n})\, d\xi$$

$$= \frac{1}{N} \sum_{n=1}^{N} h(x - x^{n})$$

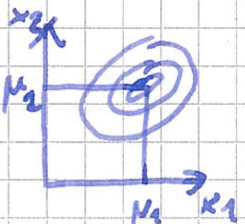
Auf den Impulsen sitzt Gauß-Kern

Wenn Verianz zu klein dann erst wieder Direc ähnlich, wenn zu groß zu ungenau... Tradeoff

1) Multiveriete Gaußverteilung $x \in \mathbb{R}^{d}$

$$N(\bar{x}|\theta) = p(\bar{x}|\Theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\bar{x}-\bar{\mu})^{T} \cdot \Sigma^{-1}(\bar{x}-\bar{\mu})\right)$$

$\Theta \ldots \{\bar{\mu}, \Sigma\}$

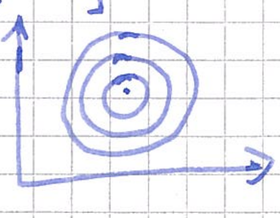


$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$

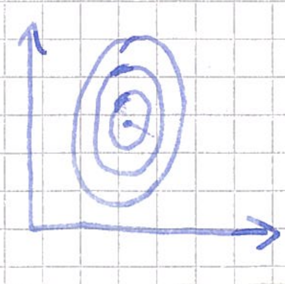
↳ symmetr. $\Sigma = \Sigma^{T}$
↳ semi positiv definit

$\Sigma = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$



$\Sigma = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$

$b > a$



WICHTIG für Prüfung

<u>ML-Schätzer</u> (Maximum Likelihood)

geg.: $X = \{\bar{x}^1 \dots \bar{x}^N\}$ ges.: Θ

Likelihood:

$$P(X|\Theta) = P(x_1 \dots x_N|\Theta) = P(\bar{x}_1|\Theta)\, P(\bar{x}_2|\cancel{\bar{x}_1}\,\Theta) \cdot P(\bar{x}_3|\cancel{\bar{x}_2 \bar{x}_1}\,\Theta) \dots$$
$$\cdot P(\bar{x}_N|\cancel{\bar{x}_{N-1} \dots \bar{x}_1}\,\Theta) = $$ iid... independent & indetically distributed
$$= \prod_{n=1}^{N} P(x_n|\Theta)$$

Likelihood $P(X|\Theta) \overset{iid}{=} \prod_{n=1}^{N} P(\bar{x}_n|\Theta)$

log Likelihood $L(X|\Theta) = \log \prod_{n=1}^{N} p(\bar{x}_n|\Theta) = \sum_{n=1}^{N} \log p(\bar{x}_n|\Theta)$

ML-Schätzer $\frac{\partial L(X|\Theta)}{\partial \Theta} \overset{!}{=} 0$ $\Theta_{ML} = \arg\max_{\Theta} L(X|\Theta)$

Bsp. Gauß: $L(X|\Theta) = \sum_{n=1}^{N} \log N(\bar{x}_n|\Theta)$

$$\frac{\partial L(X|\Theta)}{\partial \bar{\mu}} \overset{!}{=} 0 \qquad \boxed{\bar{\mu} = \frac{1}{N}\sum_{n=1}^{N} \bar{x}_n}$$

$$\frac{\partial L(X|\Theta)}{\partial \Sigma} \qquad \boxed{\Sigma = \frac{1}{N}\sum_{n=1}^{N} (\bar{x}_n - \bar{\mu})(\bar{x}_n - \bar{\mu})^T}$$

Bayes'sche Schätzer

$ML \rightarrow \Theta$ fix und unbekannt

↳ Θ wird als RV (random variable) verwendet

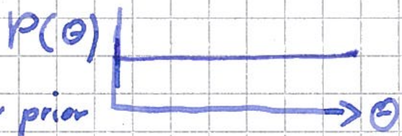
$$\underset{\text{posterior W!}}{P(\Theta|X)} = \frac{P(X|\Theta) \cdot \overset{\text{prior W!}}{P(\Theta)}}{P(X)}$$

MAP-Schätzer (Maximum-a-posterior)

$$\Theta_{MAP} = \arg\max_{\Theta} P(\Theta|X) = \arg\max_{\Theta} \left[P(X|\Theta)\, P(\Theta)\right]$$

Wenn $P(\Theta) \Rightarrow$ uniform

↳ non informativ prior



$$\Theta_{MAP} = \Theta_{ML}$$

CI VO 22.05.2012

Bayes Klassifikator

$$X = \{ \langle \bar{x}^1, t^1 \rangle, \dots, \langle \bar{x}^N, t^N \rangle \}$$

$\bar{x}^i \in \mathbb{R}^d$ $t \in N$ ..... # Samples

$t^n \in \{1, 2, \dots, C\}$ C ... # Klassen

Bayes: $$\underbrace{P(t|\bar{x})}_{\text{posterior prob}} = \frac{\overbrace{P(\bar{x}|t)}^{\text{likelihood}} \cdot \overbrace{P(t)}^{\text{prior prob}}}{\sum_{t^i} P(\bar{x}|t^i) \cdot P(t^i)}$$

$\sum_{t^i} P(\bar{x}|t^i) \cdot P(t^i) = P(\bar{x})$

Klassifikation:

$f: \bar{x} \Rightarrow t^*$

parametrisches Modell $N(\bar{x}|\Theta$
oder nicht parametrisch

$$t^* = \arg\max_t P(t|\bar{x}) = \arg\max_t \left[P(\bar{x}|t) \cdot P(t) \right]$$

$$= \arg\max_t \left[\ln P(\bar{x}|t) + \ln P(t) \right]$$

$$= \arg\max_t g_t(\bar{x})$$

$g_t(\bar{x})$ .... Entscheidungsfkt.

Bsp.: 2 Klassen $t \in \{1, 2\}$

$d = 2$ $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

geg.: X

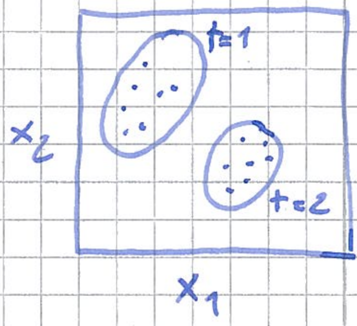
Klasse $t = 1$: $X^1 = \{ x^n \mid t^n = 1 \}$

$t = 2$: $X^2 = \{ x^n \mid t^n = 2 \}$

$p(\bar{x}|t) = N(\bar{x}|\Theta_t)$

$X^1 \Rightarrow \quad \Theta_1 = \{\bar{\mu}_1, \Sigma_1\}$

$X^2 \Rightarrow \quad \Theta_2 = \{\bar{\mu}_2, \Sigma_2\}$



$P(t=i) = \frac{|X^i|}{|X|}$

Gauß pdf:

$$N(\bar{x}|\Theta_t) = p(\bar{x}|t) = \frac{1}{(2\pi)^{d/2}\,|\Sigma_t|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\bar{x}-\bar{\mu}_t)^T \Sigma_t^{-1} (\bar{x}-\bar{\mu}_t)\right)$$

$$g_t(\bar{x}) = -\frac{1}{2}(\bar{x}-\bar{\mu}_t)^T \Sigma_t^{-1}(\bar{x}-\bar{\mu}_t) - \underbrace{\cancel{\frac{d}{2}\ln(2\pi)}}_{\text{const.}} - \frac{1}{2}\ln(|\Sigma_t|) + \ln(P(t))$$

$\boxed{\text{3 Cases:}}$

1) $\Sigma_t = \sigma^2 I$ ... Einheitsmatrix

$$g_t(\bar{x}) = -\frac{\|\bar{x}-\bar{\mu}_t\|^2}{2\sigma^2} + \ln(P(t))$$

$\|\bar{x}-\bar{\mu}_t\|^2 \rightarrow (\bar{x}-\bar{\mu}_t)^T \cdot (\bar{x}-\mu_t)$

$\ln(P(t))$: Vernachlässigbar wenn $P(t)$ uniform

$\boxed{\text{Lineare Entscheidungsgrenze}}$

"Entscheidungsgrenze"

$$g_t(\bar{x}) = g_i(\bar{x}) \quad i \neq t \quad \Rightarrow \quad P(t|\bar{x}) = P(i|\bar{x})$$

2) $\Sigma_t = \Sigma$

$$g_t(\bar{x}) = \underbrace{-\frac{1}{2}(\bar{x}-\bar{\mu}_t)^T \Sigma_t^{-1}(\bar{x}-\bar{\mu}_t)}_{\text{Mahalanobis Distance}} + \ln(P(t))$$

$\boxed{\text{Lineare Entscheidungsgrenze}}$

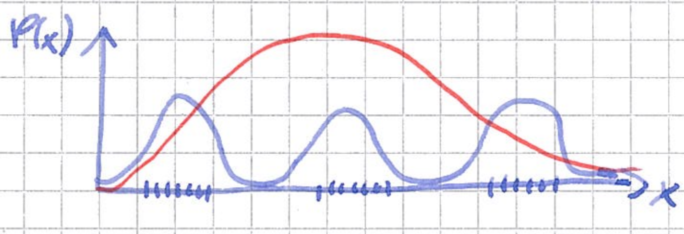
3) Σ_t beliebig

$$g_t(\bar{x}) = -\frac{1}{2}(\bar{x}-\bar{\mu}_t)^T \Sigma_t^{-1} (\bar{x}-\bar{\mu}_t) - \frac{1}{2}\ln(|\Sigma_t|) + \ln(P(t$$

$\Downarrow$

hyper quadratische Entscheidungsgrenze

Falls Daten verteilt ist eine Gauß Fkt. nicht ideal



Besser Summe mehrerer Gaußfunktionen

$\Rightarrow$ Gaußsche Mischverteilung

$$p(\bar{x}_n | \Theta) = \sum_{m=1}^{M} \alpha_m N(\bar{x}_n | \underbrace{\bar{\mu}_m \Sigma_m}_{\Theta_m})$$

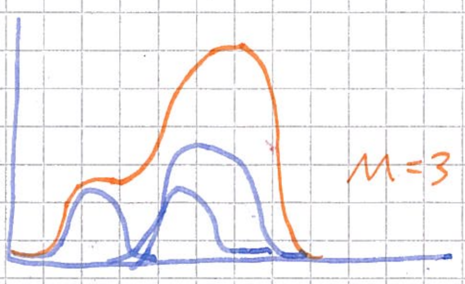
M... # Komponenten

α... Gewichtungsfaktor

Bed.: $\sum_{m=1}^{M} \alpha_m = 1 \quad 0 \leq \alpha_m \leq 1$

$$\int_{-\infty}^{\infty} p(\bar{x}|\Theta)\, d\bar{x} = 1$$

$$\Theta = \{\alpha_1, \dots, \alpha_M, \Theta_1, \dots, \Theta_M\}$$



Schätzproblem: geg.: $X = \{\bar{x}_1, \dots, \bar{x}_n\}$ $\bar{x} \in \mathbb{R}^d$

ges.: Θ

ML-Schätzer:

$$\Theta_{ML} = \arg\max_{\Theta} \{ \log \{ P(X|\Theta) \} \}$$

$$\frac{\partial \log(P(X|\Theta))}{\partial \Theta} \overset{!}{=} 0$$

$$\log(P(X|\Theta)) \overset{iid}{=} \ln \prod_{n=1}^{N} \underbrace{p(\bar{x}_n|\Theta)}_{\text{gaußsche Mischverteilung}} = \sum_{n=1}^{N} \ln \left[\sum_{m=1}^{M} \alpha_m N(\bar{x}_n | \bar{\mu}_m \Sigma_m \right.$$

Mittelwert:

$$\frac{\partial \ln(P(X|\Theta))}{\partial \bar{\mu}_m} = \sum_{n=1}^{N} \frac{1}{\sum_{m'=1}^{M} \alpha_{m'} N(\bar{x}_n | \bar{\mu}_{m'} \Sigma_{m'})} \cdot \frac{\partial \sum_{m=1}^{M} \alpha_m N(\bar{x}_n | \bar{\mu}_m \Sigma_m)}{\partial \bar{\mu}_m}$$

$$= \sum_{n=1}^{N} \underbrace{\frac{\alpha_m N(\bar{x}_n | \bar{\mu}_m \Sigma_m)}{\sum_{m'=1}^{M} \alpha_{m'} N(\bar{x}_n | \bar{\mu}_{m'} \Sigma_{m'})}}_{= \Gamma_m^n} \cdot \frac{\partial \ln [\alpha_m \cdot N(\bar{x}_n | \bar{\mu}_m \Sigma_m)]}{\partial \bar{\mu}_m}$$

$$\frac{\partial \ln(N(\bar{x}_n | \bar{\mu}_m \Sigma_m))}{\partial \bar{\mu}_m} = \partial - \frac{1}{2}(x_n - \bar{\mu}_m)^T \Sigma_m^{-1} (\bar{x}_n - \bar{\mu}_m) - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln(|\Sigma_m$$

$$\left(\text{Bem.: } \frac{\partial (\bar{a} - \bar{x})^T C (\bar{a} - x)}{\partial \bar{x}} = (C + C^T) \cdot (\bar{a} - \cancel{\bar{x}}) \right) = -\frac{1}{2} \left(\Sigma_m^{-1} + \underbrace{(\Sigma_m^{-1})^T}_{\text{symmetr.}} \right) \cdot (\bar{x}_n - \bar{\mu}_m)$$

$$= -\Sigma_m^{-1} (\bar{x}_n - \bar{\mu}_m)$$

$$= -\sum_{n=1}^{N} \Gamma_m^n \cdot \Sigma_m^{-1} (\bar{x}_n - \bar{\mu}_m)$$

$$\frac{\partial}{\partial \bar{\mu}_m} \overset{!}{=} 0$$

$$-\sum_{n=1}^{N} \Gamma_m^n \cdot \Sigma_m^{-1} (\bar{x}_n - \bar{\mu}_m) = 0 \qquad / \cdot \Sigma_m$$

$$-\sum_{n=1}^{N} \Gamma_m^n \cdot \bar{x}_n + \bar{\mu}_m \sum_{n=1}^{N} \Gamma_m^n = 0$$

$$\boxed{\bar{\mu}_m = \frac{\sum_{n=1}^{N} \Gamma_m^n \bar{x}_n}{\sum_{n=1}^{N} \Gamma_m^n}}$$

$\sum_{n=1}^{N} \Gamma_m^n = N_m$ ...... Effektive Anzahl von Datenpunkten die von der m-ten Komponente modelliert werden

$\Gamma_m^n = P(m \mid \bar{x}^n \, \Theta) \Rightarrow$ posterior W! $\quad$ (W! = prob.)

CI Vorlesung 5. 6. 2012

Gaußsche Mischverteilung

$$p(\bar{x}_n | \Theta) = \sum_{m=1}^{M} \alpha_{m} \cdot N(\bar{x}_n | \bar{\mu}_m \Sigma_m)$$

$M \ldots$ # Komponenten

$\bar{x}_n \in \mathbb{R}^d$

$\alpha_m = P(m$ $\quad 0 \leq \alpha_m \leq 1 \quad \sum_{m=1}^{M} \alpha_m = 1$

$\Theta = \{\alpha_m, \bar{\mu}_m, \Sigma_m\}|_{m=1 \ldots M}$

↳ ML- Schätzer

ML - Schätzer

$X = \{\bar{x}_1, \ldots, \bar{x}_N\}$ $\quad N \ldots$ # Samples

$\Theta_{ML} = ?$

$\Theta_{ML} = \underset{\Theta}{\arg\max} [\log P(X|\Theta)]$

Lsg.: $\frac{\partial \log P(X|\Theta)}{\partial \Theta} \overset{!}{=} 0$

Standard ML-Schätzer bei Prüfung

$\log P(X|\Theta) \overset{iid}{=} \log \prod_{n=1}^{N} p(\bar{x}_n | \Theta)$

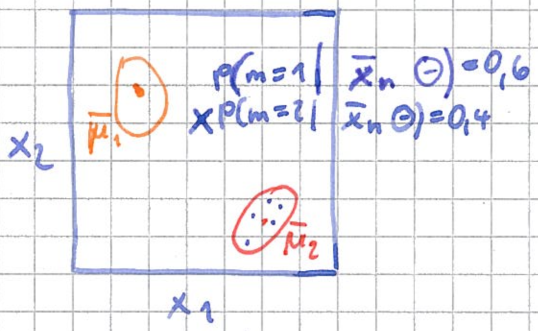
$$\boxed{\log P(X|\Theta) = \sum_{n=1}^{N} \log \sum_{m=1}^{M} \alpha_m N(\bar{x}_n | \bar{\mu}_m \Sigma_m)}$$

$$\frac{\partial \log P(X|\Theta)}{\partial \bar{\mu}_m} \stackrel{!}{=} 0 \xrightarrow[\text{Mal}]{\text{Letztes}} \bar{\mu}_m = \frac{\sum_{n=1}^{N} \Gamma_m^n \, \bar{x}_n}{N_m}$$

$N_m = \sum_{n=1}^{N} \Gamma_m^n$ ... Effektive # von Datenpunkten die von Komponente m modelliert werden

Posterior

$$\Gamma_m^n = P(m \mid \bar{x}_n \, \Theta) = \frac{\alpha_m \cdot N(\bar{x}_n \mid \bar{\mu}_m \, \Sigma_m)}{\sum_{m'=1}^{M} \alpha_{m'} \cdot N(\bar{x}_n \mid \bar{\mu}_{m'} \, \Sigma_{m'})}$$



$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \bar{x}$$

Henne Ei Problem ⇒ deshalb iterativer Algorithmus später

$$\frac{\partial \log P(X|\Theta)}{\partial \Sigma_m} \stackrel{!}{=} 0 \qquad \Sigma_m = \frac{1}{N_m} \sum_{n=1}^{N} \Gamma_m^n (\bar{x}_n - \bar{\mu}_m)(\bar{x}_n - \bar{\mu}_m)^T$$

$$\frac{\partial \log P(X|\Theta)}{\partial \alpha_m} \stackrel{!}{=} 0 \qquad \sum_{m=1}^{M} \alpha_m = 1$$ ⇒ Bedingung

Optimierung mit Lagrange Multiplikator

$$J(m) = \log P(X|\Theta) + \lambda \cdot \left(\sum_{m=1}^{M} \alpha_m - 1\right)$$

$$\frac{\partial J(m)}{\partial \alpha_m} \stackrel{!}{=} 0$$

$$\frac{\partial \log P(X|\Theta)}{\partial \alpha_m} = \sum_{n=1}^{N} \frac{1}{\sum_{m'=1}^{M} \alpha_{m'} N(\bar{x}_n | \bar{\mu}_{m'} \Sigma_{m'})} \cdot N(\bar{x}_n | \bar{\mu}_m \Sigma_m)$$

$$\frac{\partial J(m)}{\partial \alpha_m} = \sum_{n=1}^{N} \underbrace{\boxed{\frac{\alpha_m \cdot N(\bar{x}_n | \bar{\mu}_m \Sigma_m)}{\sum_{m'=1}^{M} \alpha_{m'} \cdot N(\bar{x}_n | \bar{\mu}_{m'} \Sigma_{m'})}}}_{=\Gamma_m^n} + \lambda \cdot \alpha_m \stackrel{!}{=} 0 \qquad \Big/ \cdot \alpha_m$$

$$\sum_{m=1}^{M} \underbrace{\sum_{n=1}^{N} \Gamma_m^n}_{N_m} + \sum_{m=1}^{M} \lambda \cdot \alpha_m \stackrel{!}{=} 0 \qquad \Big/ \sum_{m=1}^{M}$$

$$\left(\underbrace{\sum_{m=1}^{M} N_m}_{N} + \lambda \underbrace{\sum_{m=1}^{M} \alpha_m}_{1} \stackrel{!}{=} 0 \qquad \Rightarrow \boxed{\lambda = -N} \right)$$

$$\underbrace{\sum_{n=1}^{N} \Gamma_m^n}_{N_m} - N \cdot \alpha_m \stackrel{!}{=} 0 \qquad \Rightarrow \boxed{\alpha_m = \frac{N_m}{N}}$$

Keine geschlossenen Lösungen $\Rightarrow$ Iterativer Algorithmus

Algorithmus für $\hat{\Theta}_{ML}$ von Gaußschen Mischverteilungen

$\Rightarrow$ EM Algorithmus

Expectation Maximization

EM:

1) Init:

$$\underset{t=0}{\Theta^{(0)}} = \left\{ \bar{\mu}_m^{(0)}, \Sigma_m^{(0)}, \alpha_m^{(0)} \right\} \Big|_{m=1}^{M}$$

2) E-Step:

$$\Gamma_m^n = p(m \mid \bar{x}_n \, \Theta^{(t)}) \qquad \begin{matrix} m = 1 \dots M \\ n = 1 \dots N \end{matrix}$$

3) M-Step:

$$\bar{\mu}_m^{(t+1)} = \frac{1}{N_m} \sum_{n=1}^{N} \Gamma_m^n \cdot \bar{x}_n$$

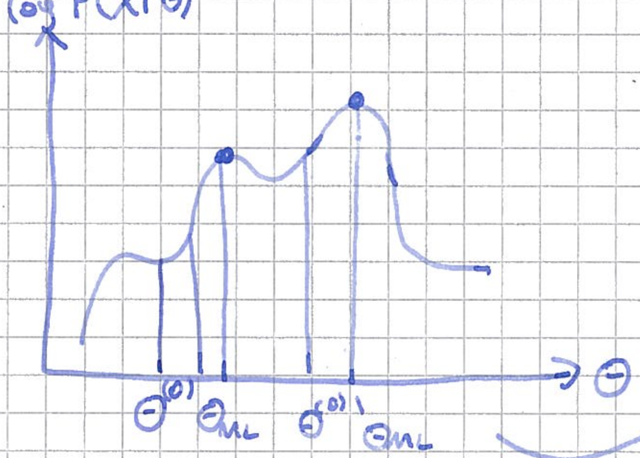
$$\Sigma_m^{(t+1)} = \frac{1}{N_m} \sum_{n=1}^{N} \Gamma_m^n (\bar{x}_n - \bar{\mu}_m^{(t+1)}) \cdot (\bar{x}_n - \bar{\mu}_m^{(t+1)})^T$$

$$\alpha_m^{(t+1)} = \frac{N_m}{N} \qquad m = 1 \dots M$$

4) Termination:

if $\log P(X \mid \Theta) \rightarrow$ konvergiert then $\Theta_{ML} = \Theta^{(t)}$

else $\rightarrow$ Step 2



Eigenschaften:

- $\log P(X \mid \Theta)$ wird mit jedem Iterationsschritt monoton besser
- Lösung von Init abhängig
  $\rightarrow$ konvergiert gegen lokales Optimum

<u>Init:</u> $\alpha_m \rightarrow$ uniform

$\Sigma_m = \Sigma \Rightarrow$ aus X berechnet

$\bar{\mu}_m =$ zufällig gewählte $\bar{x}_n$ (Samples)

$\hookrightarrow$ oder k-means Algorithmus

EM für GMV $\longrightarrow$ k-means

Annahme: a) $\Sigma_m = \sigma^2 I$

b) α_m ... vernachlässigen

c) $r_m^n \in \{0, 1\} \Leftarrow$ klassifikation

E-Step:

$$r_m^n = \frac{\alpha_m \cdot N(\bar{x}_n | \bar{\mu}_m \Sigma_m)}{\sum_{m'=1}^{M} \alpha_{m'} \cdot N(\bar{x}_n | \bar{\mu}_{m'} \Sigma_{m'})}$$

Bayes-Klassifikator Case I

a)

b)

$$g_m(\bar{x}_n) = -\frac{\|\bar{x}_n - \bar{\mu}_m\|^2}{2\sigma^2} + \cancel{\ln \alpha_m}$$

$$\Rightarrow g_m(\bar{x}_n) = -\underbrace{\|\bar{x}_n - \bar{\mu}_m\|^2}_{\text{Euklidische Distanz}}$$

c)

$$m^* = \underset{m}{\arg\max} \left(-\|\bar{x}_n - \bar{\mu}_m\|\right)$$

$$= \underset{m}{\arg\min} \left(\|\bar{x}_n - \bar{\mu}_m\|\right)$$

K jetzt gleich M

K-means:

1) Init: $\Theta^{(0)} = \{ \bar{\mu}_k^{(0)} \}_{k=1}^{K}$

$t = 0$

2) Step 1: Annahme a) b) c)

$Y_k = \{ \bar{x}_n \mid k = \arg\min_{k'} \| \bar{x}_n - \bar{\mu}_k^{(t)} \|^2 \}$ $\quad n = 1 \dots N$, $m = 1 \dots K$

3) Step 2: $\bar{\mu}_k^{(t+1)} = \frac{1}{|Y_k|} \sum_{x_n \in Y_k} \bar{x}_n \qquad t = t+1 \qquad k = 1 \dots K$

$|Y_k| \dots$ # samples in Y_k

4) Termination:

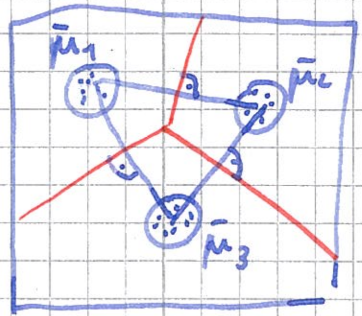
if kumulative Distanz $J = \sum_{k=1}^{K} \sum_{x \in Y_k} \| \bar{x} - \bar{\mu}_k^{(t)} \|$ konvergiert

then $\Theta = \Theta^{(t)}$

else Step 1

Eigenschaften:

- konvergiert gegen lokales Optimum
- abhängig von Initialisierung $\Theta^{(0)}$
- kumulative Distanz wird minimiert
- Entscheidungsgrenzen stückweise linear →



Anwendung:

- Vektorquantisierung

CI - Vorlesung 12.6.12

Markov Modell

Bsp.: Grammatik modell für natürliche Sprache

$q_n \in W$ W... Wörterbuch = {"ich", "du", "ist"}

n... Zeitpunkt, Position in Sequence

q... Zustand, State

Satz: "Ich gehe einkaufen" = $w_1 \dots w_n$

$P(q_1, \dots, q_n) = ?$

$P(q_1, \dots, q_n) = P(q_1) \cdot P(q_2 | q_1) \dots P(q_n | q_{n-1}, \dots, q_1)$

Parameter: $|W|$ $|W|^2$ ... $|W|^n$

⇒ Schwierig, lange Tabellen

Unigramm-Modell: $q_i \perp\!\!\!\perp q_j$ $i \neq j$ $\perp\!\!\!\perp$... stat. unabhängig

$$P(q_1, \dots, q_n) = \prod_{i=1}^{n} P(q_i)$$

Problem: P(Ich gehe einkaufen)
= P(gehe einkaufen ich)

Bigramm-Modell:

$$P(q_1, \dots, q_n) = P(q_1) \cdot \prod_{i=2}^{n} P(q_i | q_{i-1})$$

Bsp.: P(...) = P("ich") · P("gehe" | "ich") · P("einkaufen" | "gehe")

⇒ Markov-Modell 1ter Ordnung

Trigramm-Modell:

$$P(q_1, \ldots, q_n) = P(q_1) \cdot P(q_2 | q_1) \cdot \prod_{i=3}^{n} P(q_i | q_{i-1}\ q_{i-2})$$

↳ Markov-Modell 2ter Ordnung

Bsp.: Wettersequenzen:

$$q_n \in \{R, S, F\}$$

MM - 1ter Ordnung

$P(q_1)$ ... prior W!

$P(q_i | q_{i-1})$ ... Übergangs W! / Transition Probability

Hidden Markov Modell (HMM)

$$\Theta = \{\pi, A, B\}$$

HMM

MM

- $S = \{s_1, \ldots, s_n\}$ ... Menge der Zustände $N_s = |S|$
- $P(q_1 = s_i) = \pi_i$ ... Prior W! $\pi = \{\pi_1, \ldots, \pi_n\}$
- $P(q_n = s_j \mid q_{n-1} = s_i) = a_{ij} = a_{q_{n-1} q_n}$ ... Übergangs W!

$$A = [a_{ij}]_{N_s \times N_s}$$

- Beobachtungs W!

$$b_{i,x_n} = P(\overline{x}_n | q_n = s_i)$$

$\overline{x}_n \in \mathbb{N}^{d}$ (diskret)

$\overline{x}_n \in \mathbb{R}^{d}$ (Kontin.)

↳ Bei Sprechererkennung: Gauß bzw. GMV

$$B = [b_{i,x_n}]_{i=1 \ldots N_s}$$

$$P(x_n \mid q_1 \dots q_n, x_1 \dots x_{n-1}) = P(x_n \mid q_n)$$

$$\Rightarrow \quad x_n \perp\!\!\!\perp \{q_1 \dots q_{n-1}, x_1 \dots x_n\}$$ Annahme !

Normalisierungsbed.:

$$\sum_{i=1}^{N_s} \pi_i = 1 \qquad \sum_{j=1}^{N_s} a_{ij} = 1 \qquad \int_{-\infty}^{\infty} b_{i,x_n}\, dx_n = 1$$

$i = 1 \dots N_s$ $\qquad$ $i = 1 \dots N_s$

Zustandssequenz / Pfad

$$Q = \{q_1, \dots, q_n\}$$

Beobachtungssequenz

$$X = \{x_1, \dots, x_n\}$$

3 Probleme:

1) Evaluierungsproblem / klassifikation

geg.: X, Θ

ges.: $P(X \mid \Theta)$ ... likelihood / Produktions W!

Forward Backward Algorithmus

$$P(t \mid X) = \frac{P(X \mid t) \cdot P(t)}{P(x)}$$

$$t^* = \arg\max_t P(t \mid X) = \arg\max_t \underbrace{P(X \mid t) \cdot P(t)}_{P(X \mid \Theta_t)}$$

2) Dekodierungsproblem

Viterbi Algorithmus

geg.: X, Θ

ges.: $Q^* = \arg\max_Q P(Q \mid X \Theta)$

Q^*... Pfad der X bei Θ am besten erklärt

3) Schätzproblem

geg.: $X^{1:R} = \{x^1 \dots x^R\}$ R... # Beobachtungssequenzen

ges.: $\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\, P(X^{1:R} | \Theta)$

$= \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{R} \underbrace{P(x^i | \Theta)}_{\text{siehe 1)}}$

→ EM Algorithmus

Baum-Welch-Algorithmus

zu 1) $P(X|\Theta) = ?$

P für state sequenz Q ⇒ MM

$$P(Q|\Theta) = \pi_{q_1} \cdot \prod_{i=2}^{n} a_{q_{i-1} q_i} = P(q_1) \cdot \prod_{i=2}^{n} P(q_i | q_{i-1})$$

P für X bei geg Θ, Q

$$P(X|Q\,\Theta) = \prod_{i=1}^{n} P(x_i | q_i) = \prod_{i=1}^{n} b_{q_i x_i}$$

$$P(XQ|\Theta) = P(X|Q,\Theta) \cdot P(Q|\Theta)$$

$$P(X|\Theta) = \sum_{Q} P(XQ|\Theta)$$

↳ $\in \tilde{Q}$ ... Menge von Pfaden $|Q| = (N_s)^n$

von Rechenoperationen : $O(2n\,(N_s)^n)$

Effizienter : Forward / Backward Alg.

Forward / Backward Algorithmus:

Ausnützen des Distributivgesetzes

$\boxed{a\cdot(b+c)} = \boxed{ab+ac} \xrightarrow[\text{Naiv}]{} O(2n(N_s)^n)$

$\Rightarrow O(2N_s^2 n)$

2) geg.: Θ, X

ges.: Q^*

$$Q^* = \arg\max_Q P(Q|X\Theta) = \arg\max_Q \frac{P(XQ|\Theta)}{P(X|\Theta)}$$

$$= \arg\max_{Q\in\tilde{Q}} P(XQ|\Theta)$$

$$P(Q^* X|\Theta) = \max_Q P(XQ|\Theta) \overset{!}{=} P^*(X|\Theta)$$

<u>Viterbi:</u>

max. erzielbare W! für Beob. $x_1 \dots x_n$ entlang eines einzigen Pfades $q_1 \dots q_{n-1}$ der zum Zeitpunkt n im state $q_n = s_i$ mündet

$$\delta_n(i) \overset{!}{=} \max_{q_1\dots q_{n-1}} P(q_1 \dots q_{n-1}\, q_n = s_i, x_1 \dots x_n | \Theta)$$

$$\delta_1(i) = \max_{q_1} P(q_1 = s_i, x_1 | \Theta) = \max_{q_1} (\pi_{q_1} \cdot b_{q_1 x_1})$$

rekursiv

$\vdots$

$\delta_2(i)$

$\vdots$

$\delta_n(i)$

Rekursionsformel $\boxed{\delta_n(j) \overset{!}{=} \max_{1\le i\le N_s} [\delta_{n-1}(i)\, a_{ij}] \cdot b_{j x_n}}$

$$\Psi_n(i) \overset{!}{=} \arg\max_{q_1\dots q_{n-1}} P(q_1 \dots q_{n-1}\, q_n = s_i, x_1 \dots x_n | \Theta)$$

↳ Welches $[\delta_{n-1}(i) \cdot a_{ij}]$ hat zu Max geführt (Pfad)

Algorithmus:

Init: $\delta_1(i) = \pi_i \cdot b_{i x_n}$ $\quad i = 1 \dots N_s$

$\psi_1(i) = 0$

Rekursion: $\delta_n(j) = \max\limits_{1 \leq i \leq N_s} \left[\delta_{n-1}(i) \cdot a_{ij} \right] \cdot b_{j x_n}$ $\quad j = 1 \dots N_s$, $n = 2 \dots N$

$\psi_n(j) = \operatorname{argmax}\limits_{1 \leq i \leq N_s} \left[\delta_{n-1}(i) \cdot a_{ij} \right]$ $\quad N \dots$ Sequenzlänge

Termination: $P^*(X|\Theta) = \max\limits_{1 \leq i \leq N_s} \delta_N(i)$

$q_N^* = \operatorname{argmax}\limits_{1 \leq i \leq N_s} \delta_N(i)$

Wir hätten jetzt zwar das beste Ergebnis wissen aber den Pfad der hinführte nicht $Q^* = ?$ Deshalb

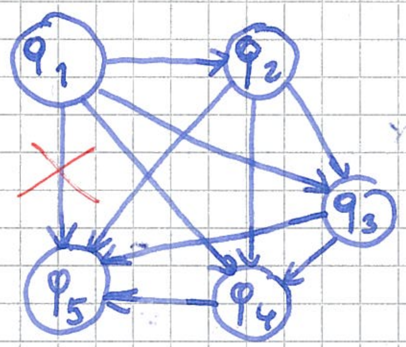
Backtracking: Extrahieren von $Q^* = \{q_1^* \dots q_N^*\}$ aus $\psi_n(i)$

$\hookrightarrow q_n^* = \psi_{n+1}(q_{n+1}^*)$ $\quad n = N-1, \dots, 1$

CI - Vorlesung 19.06.2012

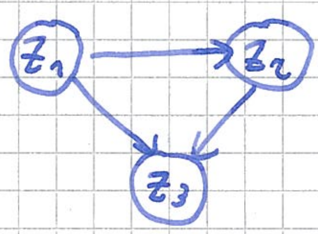
Graph Models

$$P(q_1, \dots, q_N) = P(q_1) \cdot P(q_2 | q_1) \cdot \dots \cdot P(q_N | q_{N-1}, \dots, q_1)$$



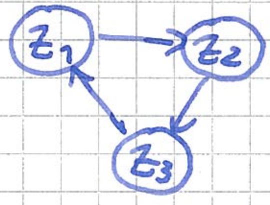
Falls Variablen stat. unabhängig sind fallen Kanten weg

Azyklisch



$$P(z_1, z_2 z_3) = P(z_1) \cdot P(z_2 | z_1) \cdot P(z_3 | z_2 z_1)$$

Zyklisch



$$P(z_1, z_2 z_3) \overset{?}{=} P(z_1 | z_3) \cdot P(z_2 | z_1) \cdot P(z_3 | z_2)$$

schlecht...

stat. independence $\quad X_i \perp X_j$

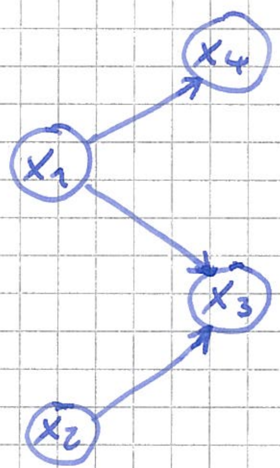
$$P(X_i\, X_j) = P(X_i) \cdot P(X_j) \Rightarrow P(X_i | X_j) = P(X_i)$$
$$P(X_j | X_i) = P(X_j)$$

cond. independence $\quad X_i \perp X_j | X_Z$

$$P(X_i\, X_j | X_Z) = P(X_i | X_Z) \cdot P(X_j | X_Z)$$

$$\Rightarrow P(X_i | X_j\, X_Z) = P(X_i | X_Z)$$
$$P(X_j | X_i\, X_Z) = P(X_j | X_Z)$$

Bsp.:



$X_1 \perp X_2$

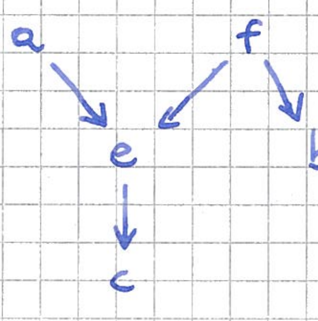
$X_4 \perp X_3 | X_1$

$X_2 \perp X_1$

$X_2 \perp X_4 | X_1$

⋮

Bsp.:



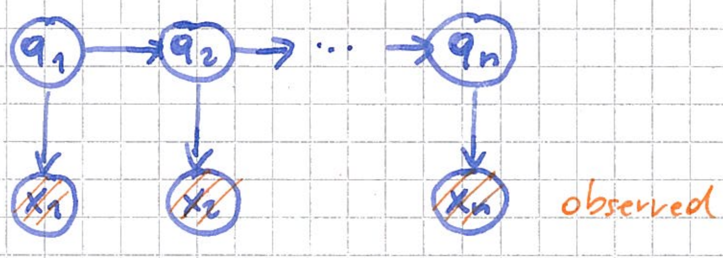
? $a \perp b | f$ ✓

? $a \perp b | e$ ✗

? $a \perp b | c$ ✗

Bsp.: HMM - revisited

$$P(XQ) = \underbrace{P(q_1)}_{\pi_{q_1}} \cdot \underbrace{P(x_1 | q_1)}_{b_{q_1 x_1}} \cdot \prod_{n=2}^{N} \underbrace{P(q_n | q_{n-1})}_{a_{q_{n-1} q_n}} \cdot \underbrace{P(x_n | q_n)}_{b_{q_n x_n}}$$



$$q_1 \perp q_3 \,|\, q_2 \quad \Rightarrow \quad q_{n+1:N} \perp q_{1:n-1} \,|\, q_n$$

$$x_2 \perp x_1 \,|\, q_2$$

Produktions W!: $P(X) = \sum_{Q \in \tilde{Q}} P(XQ)$

$\Downarrow$ Inferenz

Bsp.: GMM - revisited

$$p(x) = \sum_{m=1}^{M} \underbrace{P(m)}_{\alpha_m} \cdot \underbrace{\mathcal{N}(x | \Theta_m)}_{P(X|m)}$$

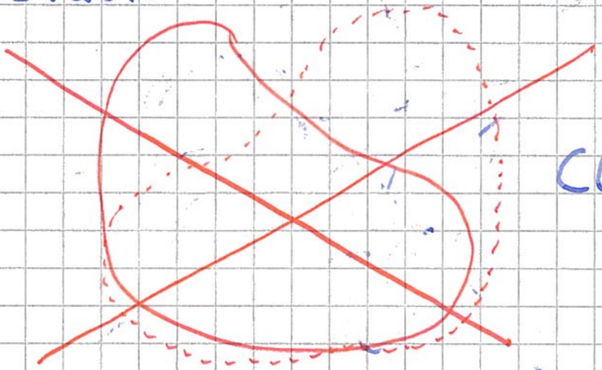
$$P(X) = \sum_{m=1}^{M} \underbrace{P(m) \cdot P(X|m)}_{P(m,X)}$$

Inferenz:

$P(m|X) = ?$ $\hat{=}$ E-Step im EM-Alg

$$p(m|X) = \frac{P(m,X)}{P(X)} = \frac{P(m) \cdot P(X|m)}{\sum_{m'} \underbrace{P(m')}_{\alpha_{m'}} \cdot \underbrace{P(X|m')}_{\mathcal{N}(x|\Theta_m)}}$$

Bsp.: CLIQUE

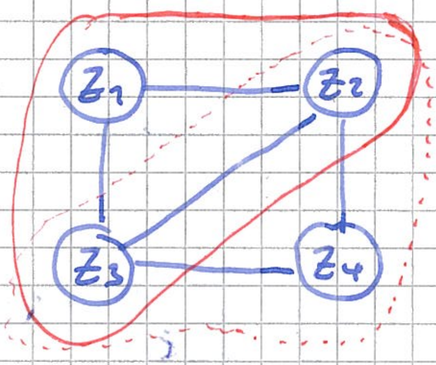


Clique: $(z_1 - z_3), (z_1 - z_2)$

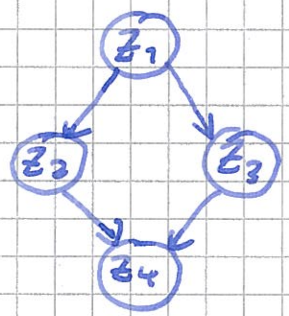
$(z_3 - z_4), (z_2 - z_4)$

$\underline{(z_1 - z_3 - z_2)}, \underline{(z_2 - z_3 - z_4)}$

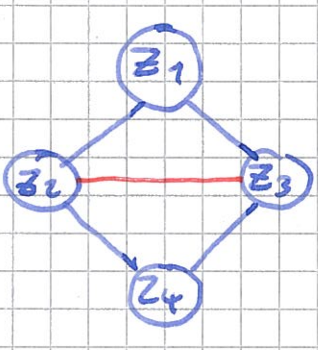
max Clique



Convert DGM to UGM



$\Rightarrow$



$$P(z_1 \dots z_4) = \underbrace{P(z_1) \cdot P(z_2 \mid z_1) \cdot P(z_3 \mid z_1)} \cdot \underbrace{P(z_4 \mid z_2\, z_3)}$$

$$P(z_1 \dots z_4) = \frac{1}{w} \quad \Psi_{123}(z_1\ z_2\ z_3) \cdot \Psi_{234}(z_2\ z_3\ z_4)$$

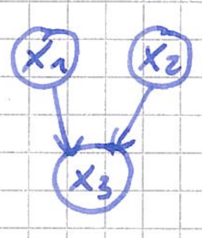
Bayes Netz

↓

DGM / BN : $P(z_1,\dots,z_4) = P(z_1)\cdot P(z_2|z_1)\cdot P(z_3|z_2)\cdot P(z_4|z_3)$

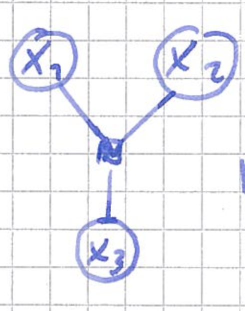
UGM : $P(z_1,\dots,z_4) = \frac{1}{W}\,\Psi_{12}(z_1 z_2)\,\Psi_{23}(z_2 z_3)\,\Psi_{34}(z_3 z_4)$

Bsp.: BN

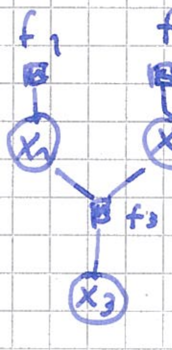


$$P(x_1, x_2, x_3) = \boxed{\underbrace{P(x_1)}_{f_1(x_1)}\ \underbrace{P(x_2)}_{f_2(x_2)}\ \underbrace{P(x_3|x_1 x_2)}_{f_3(x_1, x_2 x_3)}}$$

Faktor Graph:



$$P(x_1 x_2 x_3) = \frac{1}{W} f(x_1 x_2 x_3)$$



$$P(x_1, x_2, x_3) = \frac{1}{W} f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_1 x_2 x_3)$$

Lernen: ⇒ DGM / BN

↳ Was: •) Parameter ⇒ $P(z_i | z_{\pi_i})$ ⇒ ML - Schätzer
Bayes - Schätzer

•) Graph - Struktur ⇒ Faktorisierung

↳ Heuristiken (Greedy) ⇒ Score

CI - Vorlesung 26.06.2012

Lineare Transformationen - Dimensionsreduktion

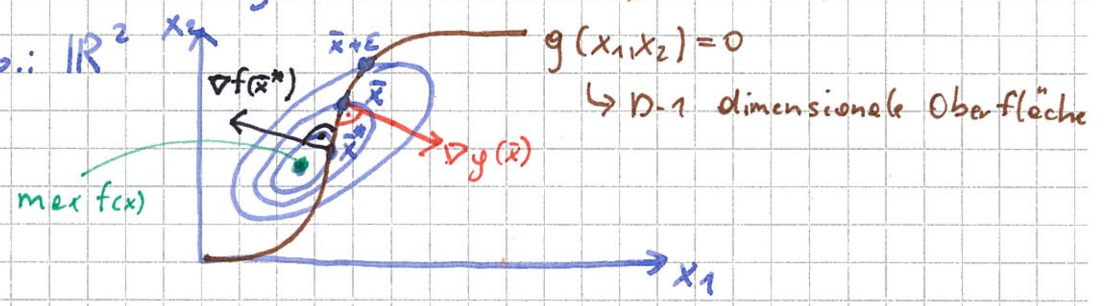
kein Prüfungs-Stoff

Exkurs: Lagrange - Multiplikatoren

→ maximieren einer Funktion mit Bedingungen (Constraints)

Problem: $\max f(\bar{x})$ (Objective) $\bar{x} = [x_1, \ldots, x_D]^T$
s.t. $g(\bar{x}) = 0$ (Constraints)

Bsp.: $\mathbb{R}^2$



Taylor von $g(\bar{x})$

$$g(\bar{x} + \bar{\varepsilon}) \approx g(\bar{x}) + \bar{\varepsilon}^T \nabla g(\bar{x})$$

$$g(\bar{x}) = g(\bar{x} + \bar{\varepsilon}) = 0$$

$$\Rightarrow \bar{\varepsilon}^T \nabla g(\bar{x}) \approx 0 \qquad \left(\lim_{\|\bar{\varepsilon}\| \to 0} \bar{\varepsilon}^T \nabla g(\bar{x}) = 0 \right)$$

$\Rightarrow \bar{\varepsilon}$ parallel zu $g(\bar{x})$

$\nabla g(\bar{x})$ normal auf $g(\bar{x})$

$\nabla f(\bar{x}^*)$ normal auf $g(\bar{x})$

$\Rightarrow \nabla f$ und ∇g parallel an $\bar{x} = \bar{x}^*$

$$\Rightarrow \nabla f(\bar{x}) + \lambda \nabla g(\bar{x}) = 0 \qquad (\lambda \neq 0)$$

→ Lagrange Funktion

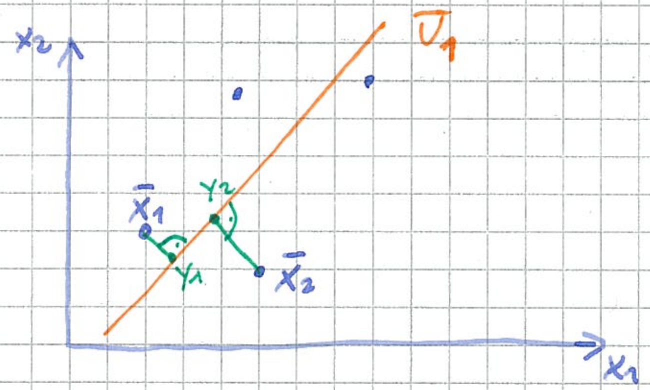
$$\mathcal{L}(\bar{x}, \lambda) = f(\bar{x}) + \lambda g(\bar{x})$$

↳ Lagrange - Mult.

$$\nabla_{\bar{x}} \mathcal{L} = \nabla_{\bar{x}} f(\bar{x}) + \lambda \nabla_{\bar{x}} g(\bar{x}) \stackrel{!}{=} 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = g(\bar{x}) \stackrel{!}{=} 0$$ (Constreints von vorher)

Bsp.: geg.: $X = \{\bar{x}_n\}_{n=1}^{N}$ $\quad \bar{x}_n \in \mathbb{R}^D$



Projektion auf:

$M = 1$-dim. Raum

$$\underset{(1)}{Y_n} = \underset{(1\times D)}{\bar{U}_1^T} \cdot \underset{(D\times 1)}{\bar{x}_n}$$

$$\Rightarrow \bar{U}_1^T = ?$$

$M \leq D$ dim. Raum

$$\underset{(M\times 1)}{\bar{Y}_n} = \underset{(M\times D)}{U} \; \underset{(D\times 1)}{\bar{x}} = \begin{bmatrix} \bar{U}_1^T \\ \vdots \\ \bar{U}_M^T \end{bmatrix} \bar{x}_n$$

Stat. Eigenschaften der transf. Daten:

Mittelwert: $m_y = \frac{1}{N} \sum_{n=1}^{N} \bar{u}_1^T \bar{x}_n = \bar{u}_1^T \cdot \bar{m}_x$

Varianz: $\sigma_y^2 = \frac{1}{N} \sum_{n=1}^{N} (\bar{u}_1^T \bar{x}_n - \bar{u}_1^T \bar{m}_x)^2$

$$= \frac{1}{N} \sum_{n=1}^{N} (\bar{u}_1^T (\bar{x}_n - \bar{m}_x))^2$$

$$= \bar{u}_1^T \underbrace{\frac{1}{N} \sum_{n=1}^{N} (\bar{x}_n - \bar{m}_x)(\bar{x}_n - \bar{m}_x)^T}_{S_x} \cdot \bar{u}_1$$

$$= \bar{u}_1^T S_x \bar{u}_1$$

Bestimmen von $\bar{u}_1^T$

Annahme: Varianz (Leistung) $\hat{=}$ Information

$\bar{u}_1 = \arg\max_{\bar{u}_1} \bar{u}_1^T S_x \bar{u}_1$ würde zu $\|\bar{u}_1\| \to \infty$

s.t. $\bar{u}_1^T \cdot \bar{u}_1 = 1$ $(:= \bar{u}_1^T \cdot \bar{u}_1 - 1 = 0)$ deshalb

↳ weil nur Richtung interessant

$(\|\bar{u}_1\|^2 = 1 \Rightarrow \bar{u}_1^T \cdot \bar{u}_1 \overset{!}{=} 1)$

Lagrange: $\mathcal{L}(\bar{u}_1, \lambda) = \bar{u}_1^T S_x \bar{u}_1 + \lambda (1 - \bar{u}_1^T \bar{u}_1)$

$$\frac{\partial \mathcal{L}(\bar{u}_1, \lambda)}{\partial \bar{u}_1} = 2 S \bar{u}_1 + 2 \lambda \bar{u}_1 \overset{!}{=} 0$$

$$S \cdot \bar{u}_1 = \tilde{\lambda} \bar{u}_1$$

(Eigenwert)
($\bar{u}_1$ ist EV von S_x zu EW $\tilde{\lambda}$)

$$\boxed{\text{oder: } \bar{u}_1^T \cdot S_x \bar{u}_1 = \tilde{\lambda} \Rightarrow \text{größter EW}}$$

damit max. Varianz

→ Principal Component Analysis (PCA)

$$y_n = \bar{u}_1^T \cdot \bar{x}_n \quad \text{(1. Hauptkomponente)}$$

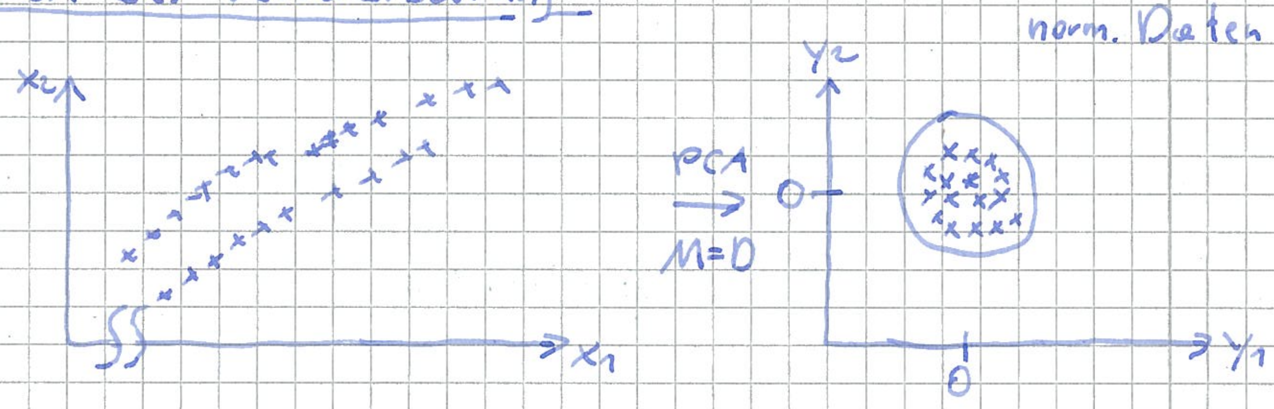
→ M Hauptkomponenten

$$\bar{y}_n = \begin{bmatrix} y_1 \\ \vdots \\ y_D \end{bmatrix} = \begin{bmatrix} \text{---} \; \bar{u}_1^T \; \text{---} \\ \vdots \\ \text{---} \; \bar{u}_D^T \; \text{---} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

M Komp. (Teil von oben)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \text{---} \; \bar{u}_1^T \; \text{---} \\ \vdots \\ \text{---} \; \bar{u}_M^T \; \text{---} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

PCA als Vorverarbeitung



$$L = \begin{bmatrix} \lambda_1 & & \emptyset \\ & \ddots & \\ \emptyset & & \lambda_D \end{bmatrix} \qquad U = \begin{bmatrix} \bar{u}_1^T \\ \vdots \\ \bar{u}_D^T \end{bmatrix} \qquad S \cdot U = U \cdot L$$

(mehrere EW)

PCA (modifiziert)

$$\bar{y}_n = L^{-1/2} \, U^T (\bar{x}_n - \bar{m}_x)$$

1) $\bar{m}_y = 0$

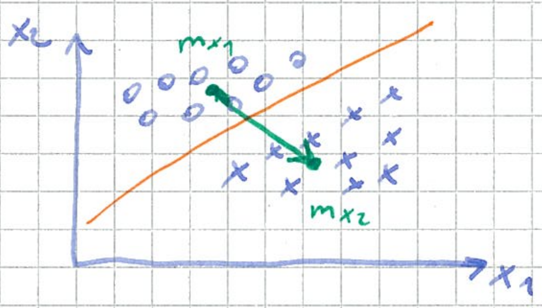
2) $S_y = \frac{1}{N}\sum_{n=1}^{N} \bar{y}_n \bar{y}_n^T = \frac{1}{N}\sum_{n=1}^{N} L^{-1/2} U^T \underbrace{(\bar{x}_n - \bar{m}_x)(\bar{x}_n - \bar{m}_x)^T}_{S_x} \cdot U \cdot L^{-1/2}$

$$= L^{-1/2} \underbrace{U^T S_x U}_{L} L^{-1/2} = I \quad \text{(Einheitsmatrix)}$$

$\Rightarrow y_i, y_j$ paarweise dekorrelliert da keine Werte außerhalb der Diagonale (S_y)

Linear Discriminant Analysis (LDA) (klassif. Problem)

$$X = \{\underbrace{\bar{x}_1, \dots, \bar{x}_{N_1}}_{T_1}, \underbrace{\bar{x}_{N_1+1}, \dots, \bar{x}_N}_{T_2}\}$$



o... T_1

X... T_2

$m_{x,1} = \frac{1}{N} \sum_{\bar{x} \in T_1} \bar{x}_n \qquad m_{x,2} = \frac{1}{N} \sum_{x \in T_2} \bar{x}_n$

Projektion: $y = \bar{U}^T \bar{x}$

Idee: $\bar{U} = \arg\max_{\bar{U}} (m_{y,2} - m_{y,1})$

$$= \arg\max_{\bar{U}} \bar{U}^T (\bar{m}_{x,2} - \bar{m}_{x,1})$$

$$\text{s.t. } \bar{U}^T \bar{U} = 1$$

$$\mathcal{L}(\bar{v}, \lambda) = \bar{v}^T (\bar{m}_{x,2} - \bar{m}_{x,1}) + \lambda (1 - \bar{v}^T \bar{v})$$

$$\boxed{\bar{v} \propto \bar{m}_{x,2} - \bar{m}_{x,1}}$$ → nicht optimal! (da nur MW verwendet)

↓ prop. zu

→ Varianzen einbeziehen

$$\sigma_{y,1}^2 = \sum_{n \in T_1} (y_n - m_{y,1})^2$$

Summen Varianz: $\sigma_y^2 = \sigma_{y,1}^2 + \sigma_{y,2}^2$ → sollte "klein" sein

Als $f(\bar{v}, \bar{x})$:

$$S_w = \sum_{n \in T_1} (\bar{x}_n - \bar{m}_{x,1})(\bar{x}_n - \bar{m}_{x,1})^T + \sum_{n \in T_2} (\bar{x}_n - \bar{m}_{x,2})(\bar{x}_n - \bar{m}_{x,2})^T$$

"Within - class - Cov."

$$\rightarrow \sigma_{y_1}^2 + \sigma_{y_2}^2 = \bar{v}^T S_w \bar{v}$$

maximieren: "Between - class - Cov."

$$(m_{y,2} - m_{y,1})^2 = \left(v^T (\bar{m}_{x,2} - \bar{m}_{x,1})\right)^2 = \ldots = \bar{v}^T \cdot S_B \cdot \bar{v}$$

$$\bar{v} = \arg\max_{\bar{v}} \frac{\bar{v}^T \cdot S_B \cdot \bar{v}}{\bar{v}^T S_W \bar{v}}$$

s.t. $\bar{v}^T \bar{v} = 1$

$$\boxed{\bar{v} \propto S_w^{-1} (\bar{m}_{x,2} - \bar{m}_{x,1}) \Rightarrow \text{LDA}}$$

$\bar{v} \propto \bar{m}_{x,2} - \bar{m}_{x,1}$ wenn $S_w \propto I$