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

↳ NN
SVM
Bayes Klassifikator

N ... # Samples
t^n ... $\in \mathbb{N}$ ⇒ Klassifikation
t^n ... $\in \mathbb{R}$ ⇒ Regression

unüberwachtes Lernen: $X = \{\bar{x}^1, \dots, \bar{x}^N\}$
(Reinforcement Learning :)

- Explorative Datenanalyse: Hauptkomponententransformation (PCA)
- Schätzung von W! - Verteilung (ML Schätz, MAP Schätz)
- ...

In der VO

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

N ... # ~~Atome~~ Experimente

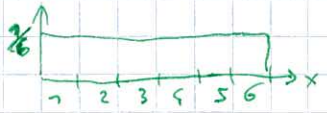
n_x ... # $X = x$

X ... Zufallsvariable.

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

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

$$\int P(x)\,dx = 1 \rightarrow \text{kontinuierlich}$$



Verbundwahrscheinlichkeit / Joint Prob.

$$P(X, Y) = P(X|Y) \cdot P(Y) \quad \Leftarrow \text{Produktregel}$$
$$= P(Y|X) \cdot P(X)$$

Kettenregel ↓

Bsp: $P(X_1, \dots, X_N) = P(X_1)\, P(X_2|X_1)\, P(X_3|X_2X_1) \dots P(X_N|X_{N-1})$

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

$$P(X) = \sum_{Y=1}^{|Y|} P(X, Y=y)$$
$$P(Y) = \sum_{X} P(X, Y)$$
} Summenregel

$$P(X, Y) = P(X|Y)\,P(Y)$$
$$= P(X)\,P(Y)$$
$$P(X|Y) = P(X)$$
} wenn X statistisch unabhängig von Y

Schätzen von Wahrscheinlichkeitsverteilungen

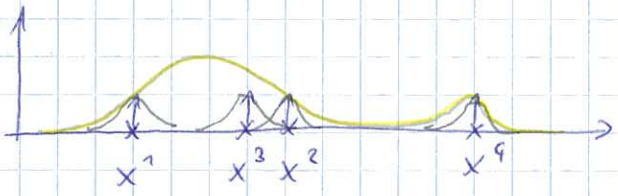
geg.: $X = \{\bar{x}^1, \dots, \bar{x}^N\}$ $\quad \bar{x}^n \in \mathbb{R}^d$

ges.: W!-Verteilung $P(X)$

I) → Parametrisches Modell: Bsp: Gaußverteilung

II) → Nicht parametrisches Modell:

II) Kernbasierter Schätzer:



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

$$\int P^e(x)\,dx = 1$$

geglättete Dichtefkt:

$h(x)$ ... Gaußkern ⇒ Glättungskern

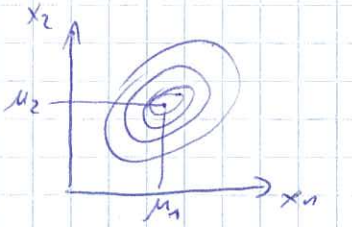
$$P^{gl}(x) = h(x) * P^e(x)$$
$$= \int_{-\infty}^{\infty} h(x-\xi)\, P^e(\xi)\, d\xi$$
$$= \int_{-\infty}^{\infty} h(x-\xi)\, \frac{1}{N}\sum_{n=1}^{N} \delta(\xi - x^n)\, d\xi$$
$$= \frac{1}{N}\sum_{n=1}^{N} \int_{-\infty}^{\infty} h(x-\xi)\, \delta(\xi - x^n)\, d\xi$$
$$= \frac{1}{N}\sum_{n=1}^{N}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

I) Multivariate Gaußverteilung $\bar{x} \in \mathbb{R}^d$

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

Θ ... $\{\bar{\mu}, \Sigma\}$

$\Theta = ?$



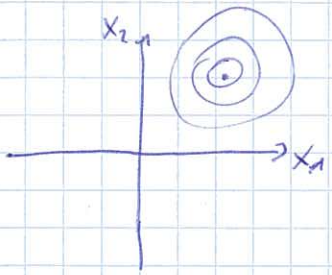
$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

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

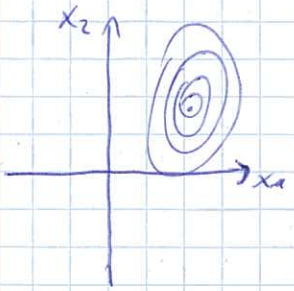
↳ symmetrisch $\Sigma = \Sigma^T$

↳ semi positiv definit

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



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



Maximum Likelihood (ML) Schätzer

geg: $X = \{\bar{x}^1, \dots, \bar{x}^N\}$

ges: Θ

Likelihood: $p(X \mid \Theta) = p(\bar{x}_1 \dots \bar{x}_N \mid \Theta) =$

$= p(\bar{x}_1 \mid \Theta)\, p(\bar{x}_2 \mid \bar{x}_1, \Theta)\, p(\bar{x}_3 \mid \bar{x}_2 \bar{x}_1 \Theta) \dots p(\bar{x}_N \mid \bar{x}_{N-1} \dots \bar{x}_1 \Theta)$

$\overset{\text{i.i.d}}{=}$ independent & identically distributed

$= \prod_{n=1}^{N} p(\bar{x}_n \mid \Theta)$

Likelihood: $p(X \mid \Theta) \overset{\text{i.i.d}}{=} \prod_{n=1}^{N} p(\bar{x}_n \mid \Theta)$

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

ML-Schätzer: $\frac{\partial L(X \mid \Theta)}{\partial \Theta} \overset{!}{=} \emptyset$

$\Theta_{ML} = \underset{\Theta}{\operatorname{argmax}}\, L(X \mid \Theta)$

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

$\frac{\partial L(X \mid \Theta)}{\partial \bar{\mu}} \overset{!}{=} \emptyset$

$\frac{\partial L(X \mid \Theta)}{\partial \Sigma} \overset{!}{=} \emptyset$

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

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (\bar{x}_n - \bar{\mu})(\bar{x}_n - \bar{\mu})^T$$

Bayes'sche Schätzer:

ML → Θ fix & unbekannt

↳ Θ wird als RV (random variable) definiert.

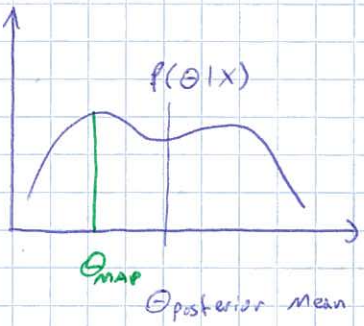
$$P(\Theta|X) = \frac{P(X|\Theta)\,P(\Theta)}{P(X)}$$

$P(\Theta)$: prior Wahrscheinlichkeit

$P(\Theta|X)$: posterior Wahrscheinlichkeit

MAP-Schätzer (Maximum-a-posteriori)

$$\Theta_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, P(\Theta|X) = \underset{\Theta}{\operatorname{argmax}}\, [P(X|\Theta)P(\Theta)]$$



Wenn $P(\Theta)$ → uniform

↳ non-informativ prior

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

Bayes Klassifikator

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

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

$t \in \mathbb{N}$ # ... samples

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

likelyhood prior prob

Bayes: $P(t|\bar{x}) = \frac{P(\bar{x}|t)\,P(t)}{\sum_{t'} P(\bar{x}|t')P(t')}$ (posterior)

$= P(\bar{x})$

$P(\bar{x})$ wurde vernachlässigt, da es die Funktion nur skaliert, aber bei argmax kann man das vernachlässigen

Klassifikation

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

$t^* = \underset{t}{\operatorname{argmax}}\ P(t|\bar{x}) = \underset{t}{\operatorname{argmax}} \left[P(\bar{x}|t)\,P(t)\right]$ — ants. parametr. $\mathcal{N}(\bar{x}|\theta_t)$ → nicht-parametr.

$= \underset{t}{\operatorname{argmax}} \left[\ln P(\bar{x}|t) + \ln P(t)\right]$

$= \underset{t}{\operatorname{argmax}}\ g_t(\bar{x})$

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

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

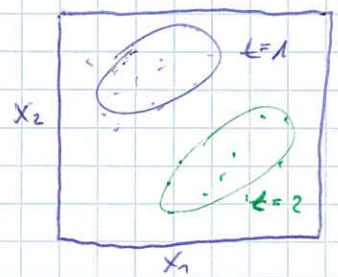
$p(\bar{x}|t) = \mathcal{N}(\bar{x}|\theta_t)$

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

geg: X

Klasse $t=1$: $X^1 = \{x^n | t^n = 1\}$ ML-Schätzer $\Rightarrow$ $\theta_1 = \{\mu_1, \Sigma_1\}$

$t=2$: $X^2 = \{x^n | t^n = 2\}$ $\Rightarrow$ $\theta_2 = \{\mu_2, \Sigma_2\}$



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

Gaußvert.

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

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

$-\frac{d}{2}\ln(2\pi)$: const

3 Fälle:

1) $\Sigma_t = \sigma^2 I$ (I: Einheitsmatrix)

$(\bar{x}-\bar{\mu}_t)^T(\bar{x}-\bar{\mu}_t)$: Euklidische Dist.

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

↳ vernachlässigbar, wenn $P(t)$ uniform.

⇒ lineare Entscheidungsgrenze

"Entscheidungsgrenze"

$$g_t(\bar{x}) = g_i(\bar{x}) \qquad i \neq t$$

$$\Rightarrow P(t \mid \bar{x}) = P(i \mid \bar{x})$$

2) $\Sigma_t = \Sigma$ → voll besetzt → irgend eine Ellipse

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

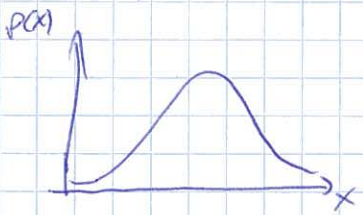
Mahalanobis Dist.

⇒ lineare Entscheidungsgrenze

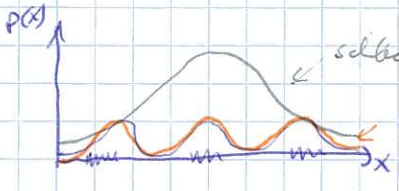
3) Σ_t beliebig $\quad g_t(\bar{x}) = -\frac{1}{2}(\bar{x}-\bar{\mu}_t)^T \Sigma_t^{-1} (\bar{x}-\bar{\mu}_t) - \frac{1}{2}\ln(|\Sigma_t|) + \ln(P(t))$

⇒ hyperquadratische Entscheidungsgr.

Bis jetzt sind wir davon ausgegangen, dass die Daten Gauß verteilt sind



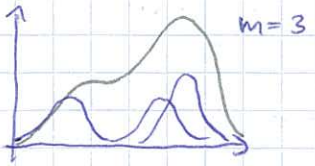
Doch was wenn Daten so verteilt sind:



besser → Gauß'sche Mischverteilung.

Gaußsche Mischverteilung

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



M ... # ... Komponenten

Bed:

$$\sum_{m=1}^{M} \alpha_m = 1 = p(n) \qquad 0 \le \alpha_m \le 1$$

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

$$\Theta = \{\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M\}$$

$\hat{\Theta} = ?$

Schätzproblem:

geg: $X = \{\bar{x}_1, \dots, \bar{x}_N\} \quad \bar{x} \in \mathbb{R}^d$

ges: Θ

$\log := \ln$

ML-Schätzer

$$\Theta = \arg\max_{\Theta} \{\log p(X \mid \Theta)\}$$

$$\frac{\partial \log P(X \mid \Theta)}{\partial \Theta} \stackrel{!}{=} 0$$

$$\log P(X \mid \Theta) \stackrel{i.i.d}{=} \log \prod_{n=1}^{N} \underbrace{p(\bar{x}_n \mid \Theta)}_{\text{Gaußsche Mischvert.}} = \sum_{n=1}^{N} \log \left[\sum_{m=1}^{M} \alpha_m \mathcal{N}(\bar{x}_n \mid \bar{\mu}_m \Sigma_m) \right]$$

Mittelwert:

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

$$\stackrel{\#}{=} \sum_{n=1}^{N} \underbrace{\frac{\alpha_m \mathcal{N}(\bar{x}_n \mid \bar{\mu}_m \Sigma_m)}{\sum_{m'=1}^{M} \alpha_{m'} \mathcal{N}(\bar{x}_n \mid \bar{\mu}_{m'} \Sigma_{m'})}}_{=: r_m^n} \cdot \frac{\partial \log [\alpha_m \mathcal{N}(\bar{x}_n \mid \bar{\mu}_m \Sigma_m)]}{\partial \bar{\mu}_m} = (*)$$

$$\frac{\partial \log \mathcal{N}(\bar{x}_n \mid \bar{\mu}_m \; \Sigma_m)}{\partial \bar{\mu}_m} =$$

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

$$\frac{\partial (\bar{a} - \bar{x})^T C (\bar{a} - \bar{x})}{\partial \bar{x}} = (C + C^T)(\bar{a} - \bar{x})$$

$$= -\frac{1}{2}\left(\Sigma_m^{-1} + (\Sigma_m^{-1})^T\right)(\bar{x}_n - \bar{\mu}_m)$$

(symetrisch)

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

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

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

$$-\sum_{n=1}^{N} r_m^n \cdot \Sigma_m^{-1}(\bar{x}_n - \bar{\mu}_m) = 0 \quad | \cdot \Sigma_m$$

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

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

$=: N_m$ ... Effektive # von Datenpunkten, die von der m-ten Komponente modelliert werden

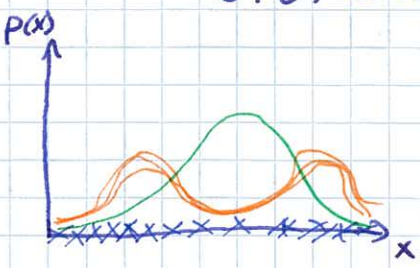
$r_m^n = P(m|\bar{x}^n, \theta)$ $\Leftarrow$ posterior prob.

Henne-Ei-Problem:

für r_m^n braucht ich θ.
Aber θ will ich schätzen.
$\Rightarrow$ iterativ!

Bsp: $\bar{x} \in \mathbb{R}^2$

CI SPSC 5.6.2012



Gauß'sche Mischverteilung

$$p(\bar{x}_n \mid \Theta) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\bar{x}_n \mid \bar{\mu}_m \Sigma_m)$$

M ... # Komponenten

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

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

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

↳ ML - Schätzer

ML - Schätzer

$X = \{\bar{x}_1, \dots, \bar{x}_N\}$ N ... # Samples

standard ML-Schätzer (Prüfungsfrage)

$\Theta_{ML} = ?$

$\Theta_{ML} = \underset{\Theta}{\operatorname{argmax}} [\log P(X \mid \Theta)]$

Lösung: $\frac{\partial \log P(X \mid \Theta)}{\partial \Theta} \stackrel{!}{=} 0$

$$\log P(X \mid \Theta) \stackrel{i.i.d}{=} \log \prod_{n=1}^{N} p(\bar{x}_n \mid \Theta)$$

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

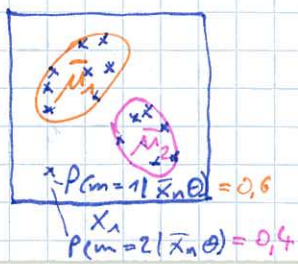
$$\frac{\partial \log P(X \mid \Theta)}{\partial \bar{\mu}_m} \stackrel{!}{=} 0 \qquad \bar{\mu}_m = \frac{\sum_{n=1}^{N} r_m^n \bar{x}_n}{N_m}$$

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

posterior

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

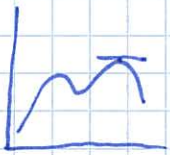
$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$:



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

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

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



Optimierung mit Lagrange Multiplikator

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

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

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

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

$$\sum_{m=1}^{M} \underbrace{\sum_{n=1}^{N} r_m^n}_{N_m} + \sum_{m=1}^{M} \lambda \alpha_m \stackrel{!}{=} 0 \qquad | \sum_{m=1}^{M}$$

$$\underbrace{\sum_{m=1}^{M} N_m}_{N} + \lambda \underbrace{\sum_{m=1}^{M} \alpha_m}_{=1} \stackrel{!}{=} 0 \quad \Rightarrow \boxed{\lambda = -N}$$

$$\underbrace{\sum_{n=1}^{N} r_m^n}_{N_m} - N \cdot \alpha_m = 0 \quad \Rightarrow \boxed{\alpha_m = \frac{N_m}{N}}$$

Algorithmus für Θ_{ML} von Gauß'schen Mischverteilungen

(EM - Algorithmus)

Expectation
Maximization

EM:

1) Init: $\Theta^{(0)} = \{ \bar{\mu}_m^{(0)}, \Sigma_m^{(0)}, \alpha_m^{(0)} \}_{m=1}^{M}$

$t = 0$

2) E-Step: $r_m^n = p(m \mid \bar{x}_n, \Theta^{(t)})$ $\quad m = 1..M$, $n = 1..N$

3) M-Step:

$$\mu_m^{(t+1)} = \frac{1}{N_m} \sum_{n=1}^{N} r_m^n \bar{x}_n$$

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

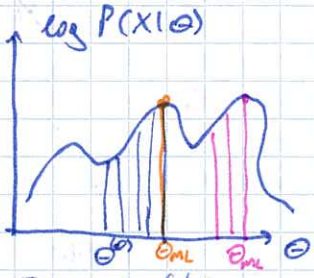
$$\alpha_m^{(t+1)} = \frac{N_m}{N}$$

$m = 1..M$

$t = t + 1$

4) Termination: if $\log P(X \mid \Theta) \rightarrow$ konvergiert $\Rightarrow \Theta_{ML} = \Theta^{(t)}$

otherwise $\rightarrow$ step 2



Eigenschaft:

- $\log P(X \mid \Theta)$ wird mit jedem Schritt monoton besser.
- EM konvergiert gegen lokales Optimum.
- Lösung von $\Theta^{(0)}$ abhängig

Init: $\alpha_m \rightarrow$ uniform verteilen

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

$\mu_m \rightarrow$ zufällig gewählte $\bar{x}_n$

↳ oder: k-means Algorithmus.

Anwendung:
- Sprache

EM für GMM $\rightarrow$ k-means

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

b) α_m ... vernachlässigen

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

E-step:

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

Bayes Klassifikator: CASE (a)

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

$$\Rightarrow g_m(\bar{x}_n) = -\|\bar{x}_n - \bar{\mu}_m\|^2$$

Euklidische Distanz

(c)

$$m^* = \arg\max_m (-\|\bar{x}_n - \bar{\mu}_m\|) = \arg\min_m \|\bar{x}_n - \bar{\mu}_m\|$$

k-means: (k-means, daher $m =: k$, $n =: K$)

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

$t = 0$

2) Step 1: Annahme (a) – (c)

$$Y_k = \{\bar{x}_n \mid k = \arg\min_{k'} \|\bar{x}_n - \bar{\mu}_{k'}\|^2\} \qquad n = 1..N, \; k = 1..K$$

3) Step 2:

$$\bar{\mu}_k^{(t+1)} = \frac{1}{|Y_k|} \sum_{\bar{x}_n \in Y_k} \bar{x}_n \qquad t = t+1 \qquad k = 1..K$$

$|Y_k|$ ... # samples in Y_k

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

otherwise Step 1

Eigenschaften:

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

Anwendung:

- Vektorquantisierung
- Init von μ_m bei GMM.

Markov Modell (MM)

Bsp.: Gramatikmodell für natürliche Sprache

$q_n \in W^*$ $\quad |W| = 1000$

n... Zeitpunkt, Position in Sequenz

q... Zustand / State

W... Wörterbuch = {"ich", "du", "ist", ...}

Satz: "Ich gehe einkaufen" = $q_1 .. q_n$

$P(q_1 .. q_n) = ?$

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

Parameter: $|W| \quad |W|^2 \quad \ldots \quad |W|^n$

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

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

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

Bigram-Modell

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

Bsp: $P(...) = P(\text{"ich"}) \cdot P(\text{"gehe"} | \text{"ich"})\, P(\text{"einkaufen"} | \text{"gehe"})$

↳ Markov Modell (MM) 1ster Ordnung

Trigram-Modell:

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

↳ MM – 2ter Ordnung

Bsp: Wettersequenzen

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

MM – 1ster Ordnung

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

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

Hidden Markov Modell (HMM) (siehe Slides auf Homepage)

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

- $S = \{s_1, \dots, s_n\}$ ... Menge der Zustände / $N_S = |S|$
- $P(q_1 = s_i) = \pi_i$ ... prior W! $\pi = \{\pi_1, \dots, \pi_{N_S}\}$
- $P(q_n = s_j \mid q_{n-1} = s_i) = a_{ij} = a_{q_{n-1} q_n}$ ... Übergangs-W!

mm / mn

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

HMM

- Beobachtungs W! $b_{i,x_n} = P(\bar{x}_n \mid q_n = s_i)$

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

$\bar{x}_n \in \mathbb{R}^d$ (kontinuierlich)
⇒ bei Spracherkennung: Gauß / Gauß'sche Mischv.

$B = [b_{i,x_n}]_{i=1 \dots N_S}$

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

Annahme

Normalisierungsbedingung:

$$\sum_{i=1}^{N_S} \pi_i = 1$$

$$\sum_{j=1}^{N_S} a_{ij} = 1 \qquad i = 1 \dots N_S$$

$$\int_{-\infty}^{\infty} b_{i,x_n} \, dx_n = 1 \qquad i = 1 \dots N_S$$

Zustandssequenz / Pfad

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

Beobachtungssequenz

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

3 Probleme:

1) Evaluierungsproblem / Klassifikation

gegeben: X, Θ → Forward / Backward Algorithmus

gesucht: $P(X \mid \Theta)$ ... Likelihood / Produktions W!

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

$$t^* = \underset{t}{\operatorname{argmax}} \, P(t \mid X) = \underset{t}{\operatorname{argmax}} \, \frac{P(X \mid t)}{P(X \mid \Theta_t)} \cdot P(t)$$

2) Dekodierungsproblem → Viterbi - Algorithmus

geg.: X, Θ

ges.: $Q^* = \underset{Q}{\operatorname{argmax}}\ P(Q \mid X \Theta)$

↳ Pfad der X bei Θ am besten erklärt

3) Schätzproblem

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

ges.: $\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\ P(X^{1:R} \mid \Theta) = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{R} \underbrace{P(X^i \mid \Theta)}_{\text{siehe 1)}}$

↳ EM - Algorithmus / Baum - Welch - Algorithmus

Zu 1) $P(X \mid \Theta) = ?$

P für state sequenz $Q \Rightarrow MM$

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

P für X bei geg. Θ, Q

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

$$P(X Q \mid \Theta) = P(X \mid Q \Theta)\, P(Q \mid \Theta)$$

$$P(X \mid Q) = \sum_{Q \in \mathcal{Q}} P(X Q \mid \Theta)$$

$\mathcal{Q} \dots$ Menge von Pfaden

$$|\mathcal{Q}| = (N_s)^n$$

Rechenoperationen: $\mathcal{O}(2 \cdot n\, (N_s)^n)$

Effizientere Methode: (Forward / Backward)

↑ Ausnützen des Distributivgesetzes

$$\underbrace{a(b+c)}_{} = \underbrace{a \cdot b + a \cdot c}_{\text{Naive}}$$

$\mathcal{O}(2 \cdot N_s^2 \cdot n)$

zu 2) geg: Θ, X

ges: Q^*

$$Q^* = \underset{Q}{\operatorname{argmax}}\ P(Q \mid X\Theta) = \underset{Q}{\operatorname{argmax}}\ \frac{P(XQ\mid\Theta)}{P(X\mid\Theta)} =$$

$$= \underset{Q\in\mathcal{Q}}{\operatorname{argmax}}\ P(XQ\mid\Theta)$$

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

Viterbi: → maximal erzielbare W! für Beobachtungen $x_1 \dots x_n$ entlang eines einzigen Pfades $q_1 \dots q_{n-1}$, der zum Zeitpunkt n im State $q_n = s_i$ mündet.

$$\delta_n(i) = \max_{q_1 \dots q_{n-1}} P(q_1 \dots q_{n-1}, q_n = s_i, x_i \dots x_n \mid \Theta)$$

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

rekursiv ↓

$\delta_2(i)$

$\vdots$

$\delta_n(i)$

$$\psi_n(i) \overset{!}{=} \underset{q_1 \dots q_{n-1}}{\operatorname{argmax}} P(\_\_\_ , \_\_\_)$$

↳ Welches $[\delta_{n-1}(i)\, a_{ij}]$ hat zum Maximum geführt

Rekursionsformel:

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

Algorithmus:

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

$\psi_1(i) = 0$

Rekursion: $\delta_n(j) = \max_{1\le i\le N_s} [\delta_{n-1}(i)\, a_{ij}]\, b_{j,x_n}$ $\qquad j = 1 \dots N_s$

$\psi_n(j) = \underset{1\le i\le N_s}{\operatorname{argmax}} [\delta_{n-1}(i)\, a_{ij}]$ $\qquad n = 2 \dots N$ $\quad N \dots$ Sequenzlänge

Termination: $P^*(X\mid\Theta) = \max_{1\le i\le N_s} \delta_N(i)$

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

Backtracking: Extraktion von Q^* aus $\psi_n(i)$

↳ $q_n^* = \psi_{n+1}(q_{n+1}^*) \qquad n = N-1 \dots 1$

Beispiel zu Viterbi ist auf den Slides (Homeapp) zu finde

19.6.2012

CI SPSC

Graphical Models

→ Tutorial + Slides auf der Homepage!

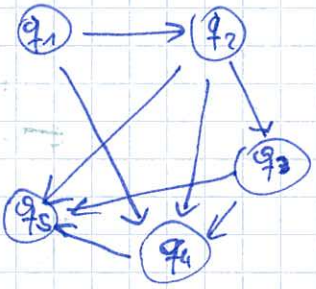
Beim Tutorial sind Section 1, 2 (2.1 & 2.2), Section 3 (3.1 - 3.4), und Section 5 + 5.1 relevant

Definitionen zu Graphen.

directed GM

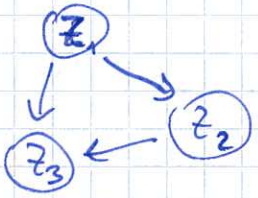
HMM kann als Bayes Netzwerk geschrieben werden.

Bsp: Bayes Netzwerke

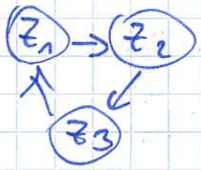


→ nicht voll verbunden: statistische Unabhängigkeit zw. einzelnen Samples.

Directed Acyclic Graph (DAG):



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



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

↳ geht nicht.

stat. independence: $x_i \perp\!\!\!\perp x_j$

$$P(x_i, x_j) = P(x_i)\, P(x_j) \Rightarrow P(x_i|x_j) = P(x_i)$$
$$P(x_j|x_i) = P(x_j)$$

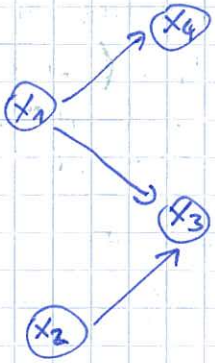
cond. independence: $x_i \perp\!\!\!\perp x_j | x_z$

$$P(x_i, x_j | x_z) = P(x_i|x_z) \cdot P(x_j|x_z) \Rightarrow P(x_i|x_j x_z) = P(x_i|x_z)$$
$$P(x_j|x_i x_z) = P(x_j|x_z)$$

Canonical examples → siehe Folien.

ad converging connection: $z_i \perp\!\!\!\perp z_j$ aber $z_i \not\perp\!\!\!\perp z_j | z_k$

Bsp:



Mit Hilfe der canonical examples ablesen:

$\rightarrow x_1 \perp\!\!\!\perp x_2$

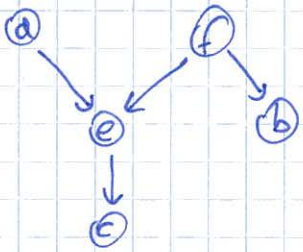
$\rightarrow x_4 \perp\!\!\!\perp x_3 \mid x_1$

$\rightarrow x_2 \perp\!\!\!\perp x_4$

$\rightarrow x_2 \perp\!\!\!\perp x_4 \mid x_1$

⋮

Bsp:



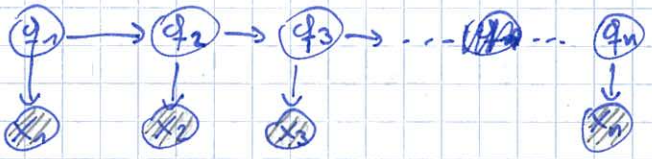
? $a \perp\!\!\!\perp b \mid f$ ✓ ← Behauptung stimmt

? $a \perp\!\!\!\perp b \mid e$ ↯ → $a \perp\!\!\!\perp b$

? $a \perp\!\!\!\perp b \mid c$ ↯ → stochastisch stimmt nicht ▽ 0

Bsp: HMM - revisited

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



x_n ... schraffiert heißt beobachtet.

$q_1 \perp\!\!\!\perp q_3 \mid q_2$ $\qquad$ $q_{n+1:N} \perp\!\!\!\perp q_{1:n-1} \mid q_n$

$x_2 \perp\!\!\!\perp x_1 \mid q_2$

Produktionsw!: $p(x) = \sum_{Q \in \mathcal{Q}} P(X\,Q)$ ← inferenz.

Bsp: GMM - revisited

$$P(x) = \sum_{m=1}^{M} \underbrace{P(m)}_{\alpha(m)} \cdot \underbrace{\mathcal{N}(x \mid \theta_m)}_{P(x \mid m)}$$

$$P(x) = \sum_{m=1}^{M} \underbrace{P(m)\,P(x \mid m)}_{P(x,\,m)}$$

Inferenz:

$P(m|X) = ? \triangleq$ E-Step im EM-Algo.

$$P(m|X) = \frac{P(m,X)}{P(X)} = \frac{P(m)\,P(X|m)}{\sum_{m'} P(m')\,P(X|m')}$$

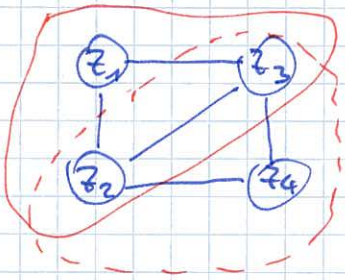
$P(m)$: $\alpha(m')$ — $P(X|m')$: $\mathcal{N}(x|\theta_{m'})$

HMM und GMM → Prüfungsrelevant!

undirected GM

Markov Netzwerk. (siehe Folien)

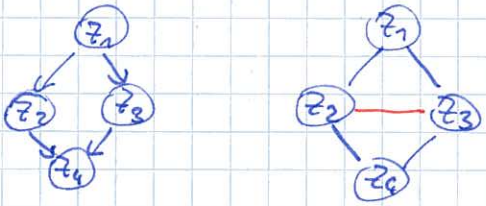
Bsp: Clique



Clique: $(z_1 - z_3)$, $(z_1 - z_2)$, $(z_3 - z_4)$, $(z_2 - z_4)$, $(z_1 - z_2 - z_3)$, $(z_2 - z_3 - z_4)$

max. Clique

Convert DGM to UGM (siehe Folien)



$$P(z_1 \dots z_4) = P(z_1)\,P(z_2|z_1)\,P(z_3|z_1)\,P(z_4|z_2 z_3)$$

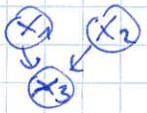
$$P(z_1 \dots z_4) = \frac{1}{W}\,\psi_{123}(z_1 z_2 z_3)\cdot\psi_{234}(z_2 z_3 z_4)$$

Factor Graph

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

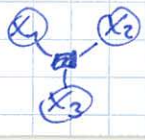
UGM: $P(z_1 \dots z_4) = \frac{1}{W}\,\psi_{12}(z_1 z_2)\,\psi_{23}(z_2 z_3)\,\psi_{34}(z_3 z_4)$

Bsp: $f_1(x_1)$ $f_2(x_2)$ $f_3(x_1 x_2 x_3)$



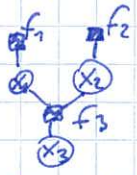
$$P(x_1 \dots x_3) = P(x_1)\,P(x_2)\,P(x_3|x_1 x_2)$$

FG:



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

FG:



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

zwei mögliche Faktor Graphen!

Lernen: => DGM / BN

↳ Was: > Parameter => $P(z_i \mid z_{\pi_i})$ => ML
Bayes Schätzer

> Graph-Struktur => Faktorisierung

↳ Greedy - Henristiken => Score

Inference (siehe Tutorial)

Lineare Transformationen - Dimensionsreduktion

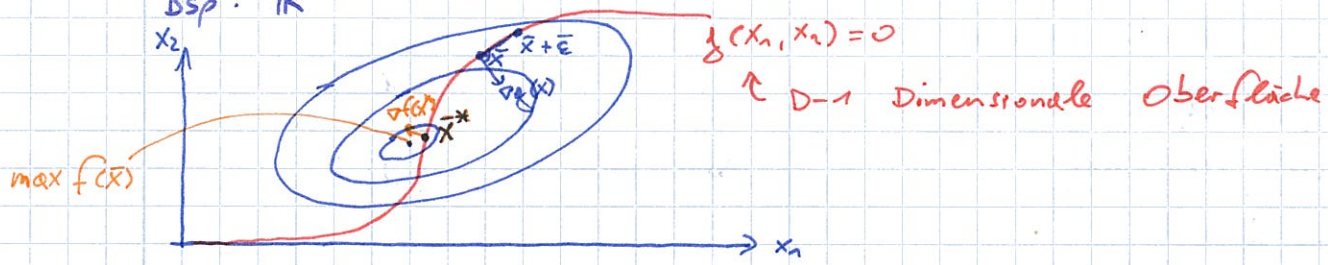
Exkurs: Lagrange - Multiplikatoren

→ maximieren einer Funktion, mit Bedingungen (Constraints)

Problem: $\max f(\bar{x})$ (Objective) $\quad \bar{x} = [x_1, \dots, x_D]^T$

s.t. $g(\bar{x}) = 0$ (Constraints) s.t. ... subject t

Bsp: $\mathbb{R}^2$



↑ D-1 Dimensionale Oberfläche

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

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

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

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

→ $\bar{\varepsilon}$ 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

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

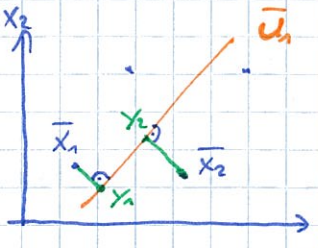
(λ: Lagrange-Multipl.)

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

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

Hintergrundinfo, kein Prüfungsstoff

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



→ Projektion auf:

$M = 1$- dim. Raum

$y_n = \bar{u}_1^T \bar{x}_n$

(1x1) (1xD) (Dx1)

→ $\bar{u}_1 = ?$

$M \leq D$ - dim. Raum

$\bar{y}_n = U \bar{x}_n = \begin{bmatrix} \bar{u}_1^T \\ \vdots \\ \bar{u}_M^T \end{bmatrix} \bar{x}_n$

(Mx1) (MxD)(Dx1)

statistische Eigenschaften der transf. Daten ($M = 1$)

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

Varianz:

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

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

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

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

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 dazu führen, dass $\|\bar{u}_1\| \to \infty$)

s.t. $\bar{u}_1^T \bar{u}_1 = 1$ → Nur Richtung von $\bar{u}_1$ wichtig

→ $\bar{u}_1^T \bar{u}_1 - 1 = 0$ → $\|\bar{u}_1\| = 1$ → $\bar{u}_1^T \bar{u}_1 = 1$

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

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

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

oder: $\bar{u}_1^T S_x \bar{u}_1 = \tilde{\lambda}$ (→ größter EW $\tilde{\lambda}$)

→ Principal Component Analysis (PCA)

PCA

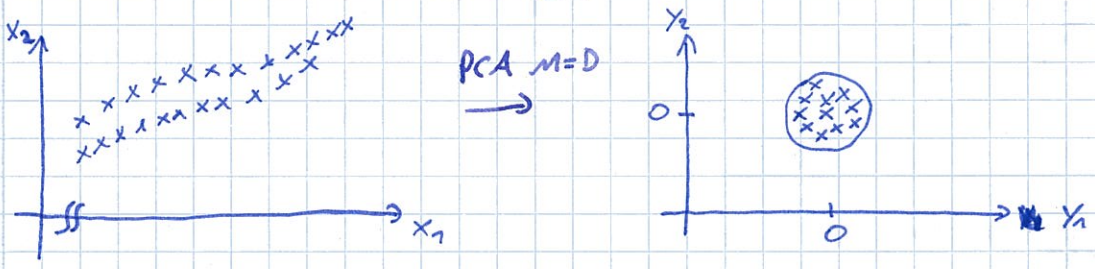
$y_n = \bar{u}_1^T \bar{x}_n$ (1. Hauptkomponente)

→ M Hauptkomponenten

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

U

PCA als Vorverarbeitung



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

$$S \cdot U = U \cdot L$$

PCA (modifiziert)

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

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

$$S_y = \frac{1}{N} \sum_{n=1}^{N} \bar{y}_n \bar{y}_n^T$$

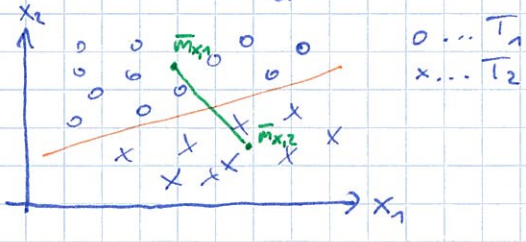
$$= \frac{1}{N} \sum_{n=1}^{N} L^{-\frac{1}{2}} U^T \underbrace{(\bar{x}_n - \bar{m}_x)(\bar{x}_n - \bar{m}_x)^T}_{\text{mit Summe} = S_x} U L^{-\frac{1}{2}}$$

$$= L^{-\frac{1}{2}} \underbrace{U^T S_x U}_{=L} L^{-\frac{1}{2}} = \underline{\underline{I}}$$

⇒ y_i, y_j paarweise dekorreliert (heißt i. A. nicht statistisch unabhängig!)

Linear Discriminant Analysis (LDA)

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



$\bar{m}_{x,1} = \frac{1}{N_1} \sum_{\bar{x} \in T_1} \bar{x}_n \; ; \quad \bar{m}_{x,2} = \dots$

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

Idee: $\bar{u} = \underset{\bar{u}}{\operatorname{argmax}} \; (m_{y,2} - m_{y,1})$

$= \underset{\bar{u}}{\operatorname{argmax}} \; \bar{u}^T (\bar{m}_{x,2} - \bar{m}_{x,1})$

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

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

$\vdots$

$\boxed{\bar{u} \propto \bar{m}_{x,2} - \bar{m}_{x,1}}$ → nicht optimal!

→ Varianzen einbeziehen!

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

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

→ Als $f(\bar{u}, \bar{x})$

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

$\rightarrow \sigma_{y,1}^2 + \sigma_{y,2}^2 = \bar{u}^T S_w \bar{u}$

maximieren: "Between-class cov"

$(m_{y,2} - m_{y,1})^2 = (u^T (\bar{m}_{x,2} - \bar{m}_{x,1}))^2 = \dots = \bar{u}^T S_B \bar{u}$

$\bar{u} = \underset{\bar{u}}{\operatorname{argmax}} \frac{\bar{u}^T S_B \bar{u}}{\bar{u}^T S_w \bar{u}}$

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

$\vdots$

$\boxed{\bar{u} \propto S_w^{-1} (\bar{m}_{x,2} - \bar{m}_{x,1})}$

↳ LDA

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