EM Algorithmus

Machine Learning I

Prof. K-R Müller
(Dr. S. Lemm)

TU Berlin

WS 2009/10

Outline

EM-Algorithmus: warm up

- **EM** Algorithmus = **E**xpectation **M**aximization Algorithmus
- EM ist **kein** Modell, wie z.B. Gaussian mixture model, K-means
- EM ist eine Methode zur Bestimmung der Modellparameter
- EM ist **keine** Kostenfunktion
- EM ein Verfahren zur Optimierung der Likelihood insbesondere für unvollständig beobachtbare Daten (d.h. missing values, latente Variablen, ...)

Allgemeine Problembeschreibung

Gegeben:

- Beobachtungen $x_1, x_2, \ldots, x_n$ gezogen aus Menge $\mathcal{X}$
- parametrisierte Verteilung $p(x|\theta)$ mit Parameter $\theta \in \Omega$, so dass

$$\int_{x \in \mathcal{X}} p(x|\theta) = 1, \quad p(x|\theta) \leq 0, \forall \theta \in \Omega$$

- Annahme: $x_1, x_2, \ldots, x_n$ unabhängig identisch verteilt (u.i.v.) mit $x_i \sim p(x|\theta^*)$, für ein $\theta^* \in \Omega$
- Gesucht: der Parameter $\hat{\theta} \in \Omega$, bzw $P(x|\hat{\theta})$ welche die Daten am wahrscheinlichsten erzeugt hat.

Likelihood

- Die Likelihood $L(\theta)$ ist die Wkt., dass $x_1, x_2, \ldots, x_n$ von der Verteilung $p(x|\theta)$ erzeugt wurden.

$$L(\theta) = p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

- Log-likelihood ist entsprechend

$$\ell(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

- Der ML-Schätzer ist

$$\hat{\theta}_{ML} = arg\max_{\theta} \ell(\theta) = arg\max_{\theta} L(\theta)$$

1. Beispiel: Münzwurf

- Beobachtungsraum: $\mathcal{X} = \{H, T\}$,
- $x_1, x_2, \ldots, x_n$ Sequenz $\langle HHTTHTHHHT \rangle$
- für $\theta \in [0, 1]$ ist

$$p(H|\theta) = \theta \text{ und } p(T|\theta) = 1 - \theta$$

dann ist die Likelihood

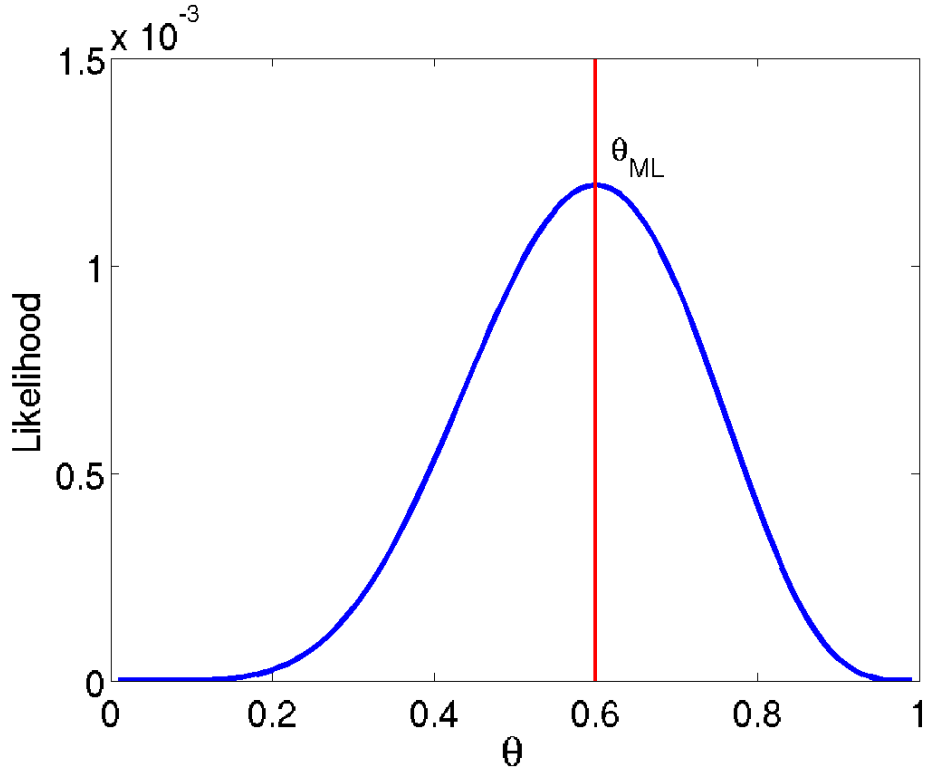
$$L(\theta) = \theta^{\#H}(1-\theta)^{\#T}$$

bzw. Log-likelihood

$$\ell(\theta) = \#H \log(\theta) + \#T \log(1-\theta)$$

1. Beispiel: Münzwurf

Für $\langle HHTTHTHHHT \rangle$



$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \qquad \Longrightarrow \qquad \theta_{ML} = \frac{\#H}{n} = 0.6$$

2. Beispiel: Gauss-verteilung

- Beobachtungsraum: $\mathcal{X} = \mathbb{R}^d$,
- $x_1, x_2, \ldots, x_n$ u.i.v.
- für $\theta = \{\mu, \Sigma\}$

$$p(x|\theta) = \left|(2\pi)^d \Sigma\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

dann ist die Log-likelihood

$$\ell(\theta) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) + const$$

2. Beispiel: Gaussverteilung

Die ML-schätzer ergeben sich wie folgt

$$\frac{\partial \ell}{\partial \mu} = 0 \implies \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (sample mean)}$$

$$\frac{\partial \ell}{\partial \Sigma} = 0 \implies \hat{\Sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top \text{ (sample covariance)}$$

Unvollständige Daten ???

- zensiert,fehlende/fehlerhafte Einträge, zusammengefasste Daten
- Latente Variablen: Mixture models, Hidden-Markov Model
- Beispiel (latente Variable): 3 Münzen ($\Theta = (\lambda, p_1, p_2)$)

  abhängig vom Wurf der ersten Münze wird 4 mal mit der zweiten (H) oder der dritten (T) Münze geworfen.

  - Vollst. Daten: $\langle H, TTTT\rangle\langle T, HHHH\rangle\langle H, TTTH\rangle\langle T, HHHH\rangle$
  - Unvollst. Daten: $\langle TTTT\rangle\langle HHHH\rangle\langle TTTH\rangle\langle HHHH\rangle$

Likelihood für unvollständigen Daten

Allgemeines Modell: Seien $x \in \mathcal{X}$ und $z \in \mathcal{Z}$ zwei Variablen mit gemeinsamer Verteilung $p(x, z|\theta)$, für $\theta \in \Omega$

- complete Log-likelihood: (x_i, z_i) vollst. beobachtbar

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i, z_i|\theta)$$

- incomplete Log-likelihood: z_i nicht beobachtbaren Variablen

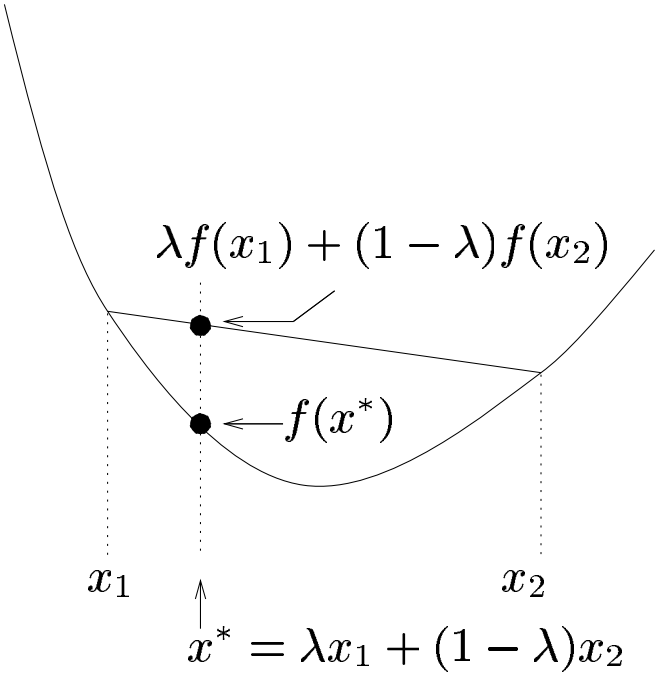
$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta) = \sum_{i=1}^{n} \log \int_{z \in \mathcal{Z}} p(x_i, z_i|\theta)\, \mathrm{dz}$$

Maximieren der Likelihood

- complete Log-likelihood
  - meist analytische Lösung (siehe Gauss, Münze)
  - selten Gradienten-Verfahren
- incomplete Log-likelihood
  - $\log(\sum)$ erschwert analytische Lösung
  - konjugierte Gradientenverfahren
  - alternative Idee: **EM**
    Iteriere
    - schätze latente Variablen z auf Basis der gegenwärtigen Parameter θ und der Daten
    - optimiere die ”complete Likelihood” bzgl der Parameter

Jensensche Ungleichung

Für f konvex, und λ_i positiv $\sum_{i=1}^{n} \lambda_i = 1$ gilt:



$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

bzw. für $(\Omega, \mathcal{F}, P)$

$$f\left(\int_{\Omega} g \, \mathrm{dP}\right) \leq \int_{\Omega} f \circ g \, \mathrm{dP}$$

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f\left(\lambda x_1 + (1-\lambda) x_2\right)$$

Untere Schranke der incomplete Likelihood

Sei $q(z)$ eine beliebige Wahrscheinlichkeitsdichte über den hidden Variablen z, dann ist die Log-likelihood beschränkt durch:

$$\begin{aligned} \ell(\theta) = \log \int q(z) \frac{p(x, z|\theta)}{q(z)} \, \mathrm{dz} &\geq \int q(z) \log \frac{p(x, z|\theta)}{q(z)} \, \mathrm{dz} \\ &=: \mathcal{F}(q, \theta) \end{aligned}$$

Direkte Anwendung der Jensenschen Ungleichung unter Verwendung, dass der Logarithmus konkav ist.

EM Algorithmus

$$\mathcal{F}(q,\theta) = \int q(z) \log p(x,z|\theta)\, \mathrm{dz} + \underbrace{\int q(z) \log q(z)\, \mathrm{dz}}_{=\mathcal{H}(q)}$$

EM verbessert $\ell(\theta)$ durch Maximieren von $\mathcal{F}(q,\theta)$ in zwei alternierenden Schritten:

1. **E-step:** maximiere $\mathcal{F}(q,\theta)$ bzgl $q(z)$ für θ fest:
$$q^{(k)}(z) = \arg\max_{q(z)} \mathcal{F}(q(z), \theta^{(k-1)}).$$

2. **M-step:** maximiere $\mathcal{F}(q,\theta)$ bzgl θ für $q(z)$ fest :
$$\begin{aligned} \theta^{(k)} &= \arg\max_{\theta} \mathcal{F}(q^{(k)}, \theta) \\ &= \arg\max_{\theta} \int q^{(k)}(z) \log p(x,z|\theta)\, \mathrm{dz}. \end{aligned}$$

Konvergenz

Betrachte die Differenz zwischen Log-likelihood und der unteren Schranke:

$$\begin{aligned}
\ell(\theta) - \mathcal{F}(q,\theta) &= \log p(x|\theta) - \int q(z) \log \frac{p(x,z|\theta)}{q(z)} \, \mathrm{dz} \\
&= \log p(x|\theta) - \int q(z) \log \frac{p(z|x,\theta)p(x|\theta)}{q(z)} \, \mathrm{dz} \\
&= -\int q(z) \log \frac{p(z|x,\theta)}{q(z)} \, \mathrm{dz} = \mathcal{KL}(q(z)\|p(z|x,\theta))
\end{aligned}$$

wobei für den Kullback-Leibler Abstand gilt

$$\mathcal{KL}(q(z)\|p(z|x,\theta)) = 0 \quad \Longleftrightarrow \quad q(z) = p(z|x,\theta)$$

$\Longrightarrow$ im **E step** wird $q(z) \longleftarrow p(z|x,\theta)$.

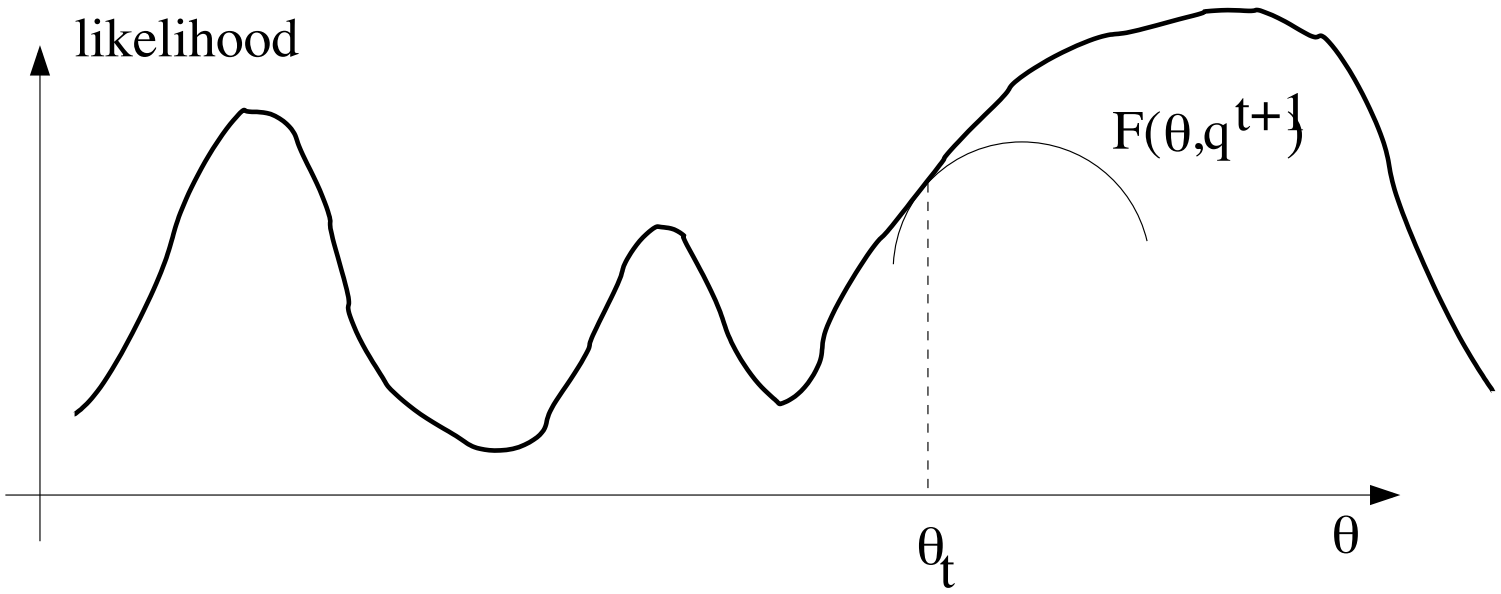
Konvergenz (2)

Somit verbessern der **E** und **M step** die Log-Likelihood, es gilt:

$$\ell(\theta^{(k-1)}) \overset{\text{E-Step}}{=} \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \overset{\text{M-Step}}{\leq} \mathcal{F}(q^{(k)}, \theta^{(k)}) \overset{\text{Jensen}}{\leq} \ell(\theta^{(k)})$$

$\Longrightarrow$ EM konvergiert zu (lokalem) Optimum von $\ell(\theta)$.

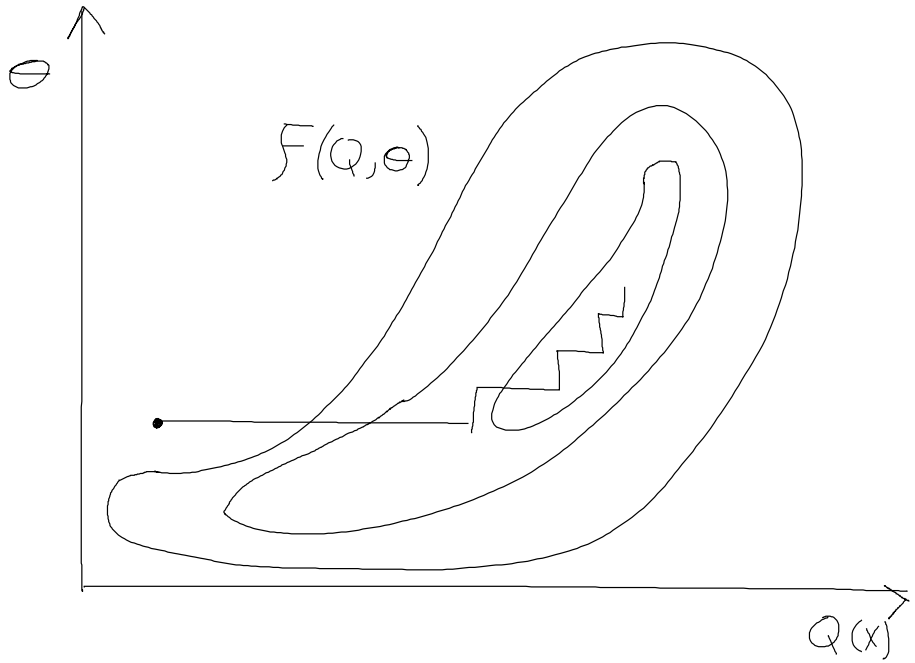
EM intuitiv (1)



copyright Sam Roweis

EM intuitiv (2)

EM als koordinatenweise Maximierung von $\mathcal{F}$



copyright Zoubin Ghahramani

Univariate Gaussian mixture model

GMMs modellieren die Dichte eines Datenpunktes $x \in \mathbb{R}$ als gewichtete Summe von K Gaussverteilungen und lassen sich als generatives Modell interpretieren

$$\begin{aligned} p(x|\theta) &= \sum_{k=1}^{K} \alpha_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{(x-\mu_k)^2}{2\sigma_k^2} \right\} \\ &= \sum_{k=1}^{K} p(z=k|\theta) p(x|z=k,\theta) \end{aligned}$$

θ umfasst die Mittelwerte μ_k, Varianzen σ_k und die Mischungskoeffizienten $\alpha_k = p(z=k|\theta)$.

Die nicht beobachtbare Variable z_i gibt an welche Komponente einen Datenpunkt x_i generiert hat.

GMM-Parameterschätzung via EM

E-step: berechne die Posterior für z_i gegeben die momentanen Parameter

$$q(z_i) = p(z_i|x_i, \theta) = \frac{p(x_i|z_i, \theta)p(z_i|\theta)}{p(x_i|\theta)}$$

Zuordnung $x_i \to \mathcal{N}_k$

$$r_{i,k} := q(z_i = k) = \alpha_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}$$

Normalisiert auf $\sum_{k=1}^{K} r_{i,k} = 1$

GMM-Parameterschätzung via EM

M-step: maximiere folgende Summe (z diskret) bzgl θ

$$E(\theta) = \sum_{k,i} q(z_i = k) \log \left[p(z_i = k|\theta) p(x|z_i = k, \theta) \right]$$

$$= \sum_{i,k} r_{i,k} \left[\log \alpha_k - \log \sigma_k - \frac{(x_i - \mu_k)^2}{2\sigma_k^2} - \frac{1}{2} \log 2\pi \right]$$

$$\text{s.t.} \sum_{k=1}^{K} \alpha_k = 1.$$

mit Hilfe von Lagrange Multiplikatoren

$$E(\theta, \lambda) = E(\theta) + \lambda(1 - \sum_{k=1}^{K} \alpha_k)$$

GMM-Parameterschätzung via EM

$$\frac{\partial E}{\partial \mu_k} = \sum_i r_{i,k} \frac{x_i - \mu_k}{2\sigma_k^2} = 0 \qquad \Longrightarrow \mu_k = \frac{\sum_i r_{i,k} x_i}{\sum_i r_{i,k}},$$

$$\frac{\partial E}{\partial \sigma_k} = \sum_i r_{i,k} \left[\frac{(x_i - \mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k} \right] = 0 \Longrightarrow \sigma_k^2 = \frac{\sum_i r_{i,k} (x_i - \mu_k)^2}{\sum_i r_{i,k}}$$

$$\frac{\partial E}{\partial \alpha_k} = \sum_i r_{i,k} \frac{1}{\alpha_k} - \lambda = 0 \qquad \Longrightarrow \alpha_k = \frac{1}{\lambda} \sum_i r_{i,k}$$

$$\text{mit} \sum_{k=1}^{k} \alpha_k = 1 \qquad \Longrightarrow \lambda = n$$

$$\Longrightarrow \alpha_k = \frac{1}{n} \sum_i r_{i,k}$$

EM Algorithmus: wrap up

- Methode die Log-likelihood für Modelle mit latenten Variablen zu maximieren, wenn sich das (schwere) Problem in zwei (leichtere) Probleme aufspalten lässt:
  1. Schätze die fehlenden/nicht beobachtete Daten aus beobachteten Daten und gegenwärtigen Parametern
  2. Verwende die ”complete” Likelihood für ML-Schätzung der Parameter
- **Pro:**
  - keine Lernrate
  - schnelle Konvergenz für kleine Dimensionen
  - jede Iteration verbessert die Likelihood
- **Contra:**
  - verendet in lokalen Minima
  - sensitiv gegenüber Initialisierung
- Generalized EM: führt im M-step einen Gradientenschritt aus anstelle des ML-schätzers

Literatur

- Hastie, Tibshirani, and Friedman, The Elements of Statistical Learning, Chapter 8.5,
- A. P. Dempster, N. M. Laird and D. B. Rubin. (1977). ”Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society, B, vol. 39, no. 1, pp. 1-38.