Exponential Family, Maximum Likelihood, EM Algorithmus und Gaussian Mixture Models

Korbinian Schwinger

23. November 2003

Inhaltsverzeichnis

1 Exponential Family

1.1 Definition

Zur Exponential Family gehören alle Dichtefunktionen, deren zugehörige Verteilungsfunktion sich folgendermassen schreiben lassen :

$$f(\mathcal{X} \mid \theta) = h(x)c(\theta)\exp(b(\theta) + \sum_{i=1}^{N} w_i(\theta)t_i(x)))$$

Hierbei ist zu beachten, daß θ alle Parameter darstellt, die für die entsprechende Dichtefunktion charakteristisch sind, d.h. für die Normalverteilung im eindimensionalen Raum bestünde θ aus dem Erwartungswert $E[\mathcal{X}] = \mu$ und der Varianz σ^2, man kann sich θ also als einen Vektor aller Parameter vorstellen. Die Funktionen h, b, w_i und t_i sind reellwertig. $h(x), t_i(x)$ hängen ausschließlich von x, $c(\theta)$ und $w_i(\theta)$ ausschließlich von θ ab.
Diese Schreibweise wird verwendet, da sich aus ihr die minimale Sufficient Statistic, auf die ich hier nicht weiter eingehen werde, leicht errechnen lässt :

$$S(X_1, \cdots, X_n) = (\sum_i t_1(x_i), \cdots, \sum_i t_k(x_i))$$

1.2 Beispiel 1 - Normalverteilung in 1 Dimension

In einer Dimension wird eine Normalverteilung durch 2 Parameter, der Varianz σ^2 und dem Erwartungswert μ festgelegt. Den Erwartungswert kann man als beste Schätzung dafür ansehen, welchen Ausgang ein Zufallsexperiment haben wird, wenn man es genau einmal durchführt. Die Varianz ist dagegen ein Maß für die Streuung, d.h. wie weit die möglichen Ergebnisse vom Erwartungswert abweichen können. Die Varianz gibt aber nicht die durchschnittliche Abweichung wieder, sondern das Quadrat davon. Der Grund hierfür ist, daß die Summe aller Ausgänge eines normalverteilten Zufallsexperiment jeweils gewichtet mit der dazugehörigen Wahrscheinlichkeit 0 wäre, d.h. um ein Maß für die mittlere Abweichung zu bekommen müsste man jeweils den Betrag der gewichteten Zufallswerte bilden. Die Betragsfunktion ist allerdings unhandlich, weswegen man statt dessen das Quadrat verwendet - damit erhält man die Varianz. Die Dichtefunktion ordnet jedem möglichen Ausgang seine Wahrscheinlichkeit zu und die Verteilung, wie wahrscheinlich es ist, daß ein Zufallswert, der kleiner oder gleich dem gezogenen auftritt.

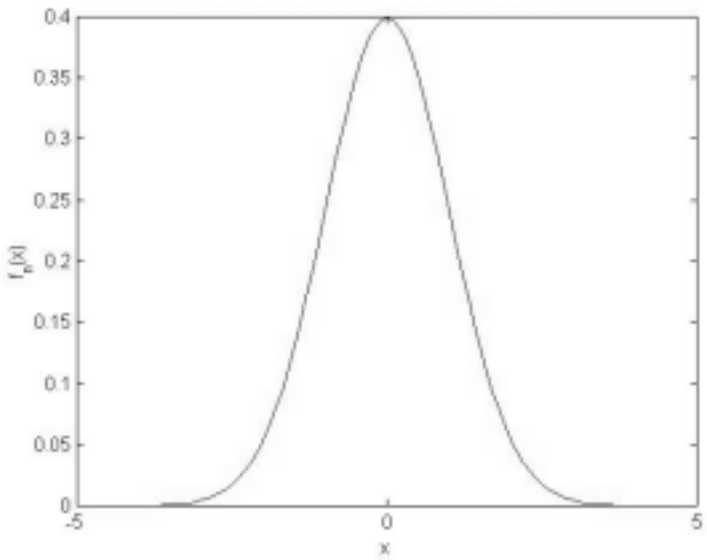

Abbildung 1: Dichtefunktion der 1-dimensionalen Standardnormalverteilung

$$Dichtefunktion : f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

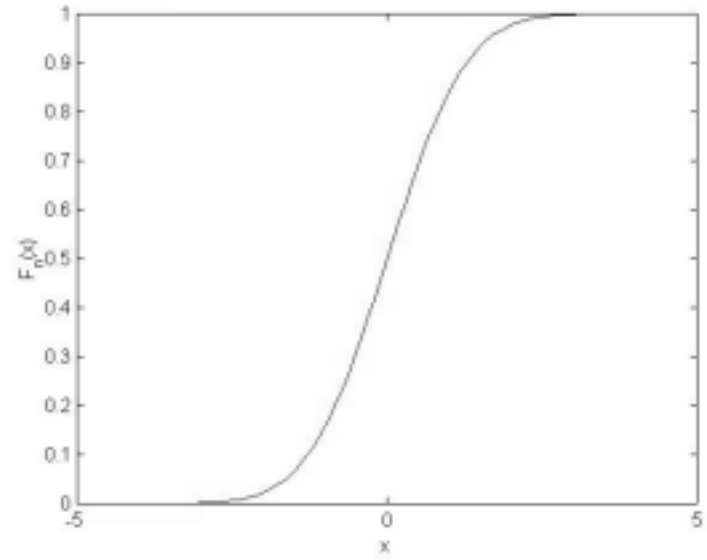

Abbildung 2: Verteilungsfunktion der 1-dimensionalen Standardnormalverteilung

$$Verteilungsfunktion : F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Ausgedrückt in der obigen Definition der Exponential Family könnte man die Verteilungsfunktion so umschreiben, daß $h(x) = 1$, $c(\theta) = \frac{1}{\sqrt{2\pi}}$, $w_i(\theta) = -\frac{1}{2}$, $t_i(x) = x^2$ und alle anderen Funktionen gleich 0. Da es sich hier um eine kontinuierliche Verteilung handelt, musste das Summenzeichen durch ein Integral ersetzt werden.

1.3 Beispiel 2 - Normalverteilung in 2 Dimensionen

In 2 Dimensionen verändern sich die Parameter der Normalverteilung: Der Erwartungswert wird zu einem 2 dimensionalen Vektor und die Varianz wird zur Kovarianzmatrix, die alle Kovarianzen enthält. Die Kovarianz ist ein Maß dafür, ob 2 Zufallsvariablen zusammenhängen, besteht kein Zusamenhang, so ist die Kovarianz 0.

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

$$\sigma_{ij} = E[(x_i - \mu_i)\,(x_j - \mu_j)]$$

In der Diagonalen der Kovarianzmatrix stehen damit die Varianzen der Zufallsvariablen. In einer abelschen Gruppe, d.h. wenn a*b=b*a gilt, ist die Kovarianzmatrix symmetrisch, da sich die beiden Faktoren der Kovarianz einfach vertauschen lassen.
Eine Dichtefunktion könnte folgendermassen aussehen :

$$\mu = \begin{pmatrix} 10 \\ 1 \end{pmatrix}$$

$$\sum = \begin{pmatrix} 9 & 1 \\ 1 & 2 \end{pmatrix}$$

1.4 Beispiel 3 - Normalverteilung in N Dimensionen

In N Dimensionen sehen die Parameter folgendermaßen aus :

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix}$$

$$\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{pmatrix}$$

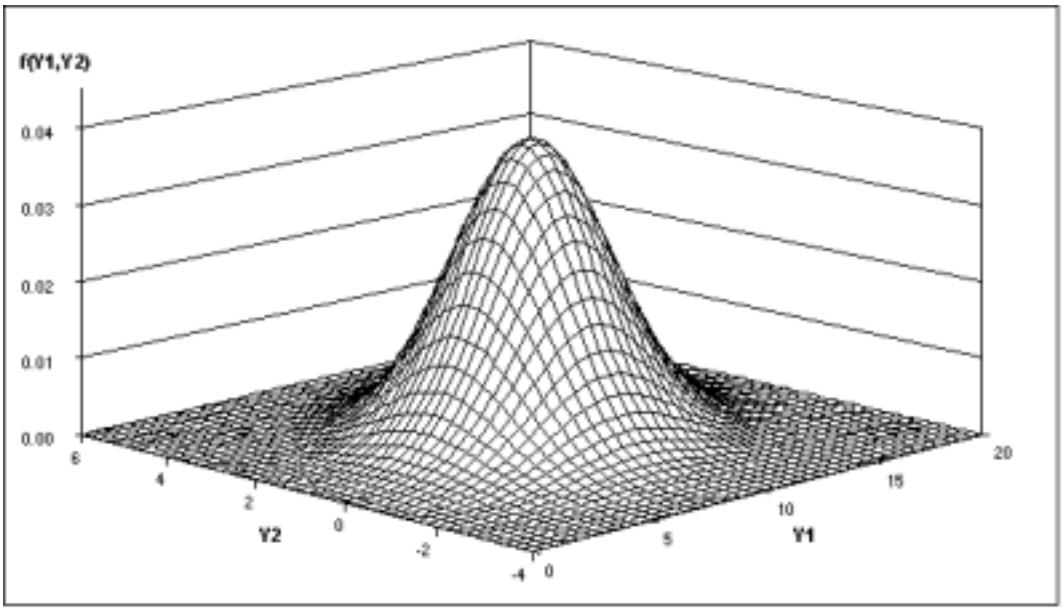

Abbildung 3: Verteilungsfunktion der 2-dimensionalen Normalverteilung

$$\sigma_{ij} = E[(x_i - \mu_i)\,(x_j - \mu_j)]$$

2 Maximum Likelihood

2.1 Prinzip

Als Ausgangspunkt hat man N Messwerte, welche von einer Dichtefunktion bekannten Typs erzeugt wurden. Die Parameter, die der Dichtefunktion zu grunde liegen, sind dagegen unbekannt. Im Prinzip kommen für die gemessenen Werte mehrere Dichtefunktionen in Frage, die sie erzeugt haben könnten, bei einer Normalverteilung könnte das Problem folgendermaßen sein :

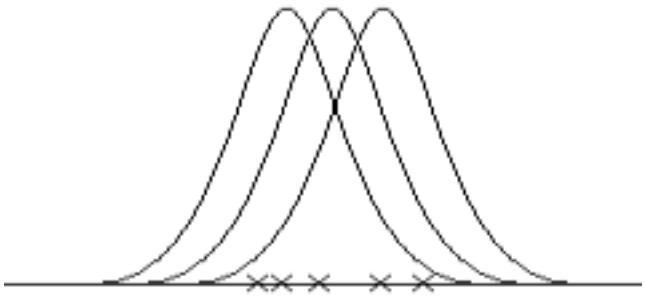

Abbildung 4: alle diese Normalverteilungen könnten die Punkte erzeugt haben

Nun würde man gerne herausfinden, welche der Funktionen wohl die erzeugende war – allein mit den Messwerten lässt sich das aber nicht mehr herausfinden, dashalb versucht man diejenige Funktion zu ermitteln, die mit der größten Wahrscheinlichkeit die Punkte erzeugt hat. Die Dichtefunktion ermittelt zu jedem Punkt die Wahrscheinlichkeit, mit der ein Zufallsexperiment auf diese Weise endet, d.h. man kann zu jedem gemessenen Wert ermitteln, mit welcher Wahrscheinlichkeit er beim Zufallsexperiment erzeugt wurde. Multipliziert man alle diese Wahrscheinlichkeiten, erhält man ein Maß dafür, wie wahrscheinlich alle Messwerte von einer bestimmten Dichtefunktion erzeugt wurden - mit diesem Gedanken erhält man eine neue Dichtefunktion, die Likelihood–Funktion :

$$p(\mathcal{X} \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta) = \mathcal{L}(\theta \mid \mathcal{X})$$

Diese Funktion übernimmt die Parameter einer Dichtefunktion und liefert die Wahrscheinlichkeit zurück, mit der die gemessenen Werte durch die Dichtefunktion mit den eingesetzten Parametern erzeugt wurde. Hiermit hat man das Problem auf ein analytisches zurückgeführt, nämlich dem Finden eines Maximums einer Funktion. Das Maximum dieser Dichtefunktion ist also derjenige Vektor von Parametern, den die Dichtefunktion am wahrscheinlichsten hatte :

$$\theta^{max} = argmax_{\theta} \mathcal{L}(\theta \mid \mathcal{X})$$

Auch wenn dieses Problem schon sehr viel besser vorstellbar ist, kann es dennoch ziemlich schwierig sein, es zu lösen, weswegen man auch oft den Logarithmus der Likelihoodfunktion maximiert, da dieser das selbe Steigungsverhalten hat, wie die Likelihoodfunktion selbst, d.h. an der Stelle, wo die Likelihoodfunktion ihr Maximum hat, hat auch ihr Logarithmus ein Maximum.

2.2 Beispiel

Angenommen man hat eine Normalverteilung mit den Parametern Varianz und Erwartungswert. Die allgemeine Dichtefunktion der Normalverteilung ist :

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

Multipliziert man nun über alle N Messwerte, kann man den konstanten Bruch N-mal aus dem Produkt ziehen und bekommt damit die Likelihood Funktion :

$$\mathcal{L}(\theta \mid \mathcal{X}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \prod_{i=1}^{N} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Jetzt bildet man den Logarithmus dieser Funktion, um die Rechnung zu vereinfachen :

$$\ln(\mathcal{L}(\theta \mid \mathcal{X})) = -N(\ln(\sqrt{2\pi} + \ln\sigma) + \sum_{i=1}^{N}\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Über Ableitungen dieser Funktion nach Erwartungswert und Varianz können die Parameter ermittelt werden, die die Funktion maximieren, denn beim Maximum einer Funktion ist die Ableitung, also deren Steigung immer gleich 0 :

$$\frac{\delta \ln \mathcal{L}}{\delta\mu} = \sum_{i=1}^{N}\frac{x_i-\mu}{\sigma^2} = 0$$

$$\frac{\delta \ln \mathcal{L}}{\delta\sigma} = -\frac{N}{\sigma} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^3} = 0$$

Diese Gleichungen werden von folgenden Werte erfüllt:

$$\mu = \bar{x}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

3 EM Algorithmus

3.1 Prinzip

Das dem EM-Algorithmus zugrundeliegende Problem ist dem der Likelihoodfunktion sehr ähnlich. Wieder hat man einige Messwerte, die von einer Dichtefunktion bekannten Typs erzeugt wurden, aber diesmal ist bekannt, daß einige Messwerte bzw. Teile davon falsch oder schlicht nicht vorhanden sind. Dies könnte daraus resultieren, daß der Beobachtungsprozess einigen Beschränkungen unterliegt, allerdings ist es ebenso möglich, daß mit dieser Annahme die späteren Rechnungen vereinfacht werden, wie man später im Teil Gaussian Mixture Models sehen wird. Zuerst nimmt man an, daß die korrekten Messwerte von einer bestimmten Verteilung erzeugt wurden. Außerdem nimmt man an, daß die fehlenden Werte ebenfalls einer Verteilung unterliegen. Man bezeichnet die gemessenen, korrekten Werte mit $\mathcal{X}$, die fehlenden bzw. falschen mit $\mathcal{Y}$ und alle Messwerte zusammen mit $\mathcal{Z}$. Mit diesen Annahmen kann man eine neue Dichtefunktion definieren:

$$p(z \mid \theta) = p(x, y \mid \theta) = p(y \mid x, \theta)p(x \mid \theta)$$

Mit dieser neuen Dichtefunktion kann man eine neue Likelihoodfunktion definieren, die Likelihoodfunktion der kompletten Daten :

$$\mathcal{L}(\theta \mid \mathcal{Z}) = \mathcal{L}(\theta \mid \mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y} \mid \theta)$$

Der erste Schritt, den der EM Algorithmus ausführt, der E-Schritt, wertet den Erwartungswert des Logarithmus der Likelihoodfunktion aus :

$$\mathcal{Q}(\theta, \theta^{(i-1)}) = E[log(p(\mathcal{X}, \mathcal{Y} \mid \theta) \mid \mathcal{X}, \theta^{(i-1)}))]$$

Bei dieser Formel bezeichnet $\theta^{(i-1)}$ keinen Parameter, der in die Funktion eingesetzt wird, sondern eine Konstante, die bisher beste Näherung an die Parameter der Dichtefunktion. θ selbst dagegen ist der Parameter der Funktion, der verwendet wird, um das Maximum der Funktion zu finden.
Der Erwartungswert auf der rechten Seite kann umgeschrieben werden als :

$$E[log(p(\mathcal{X}, \mathcal{Y} \mid \theta) \mid \mathcal{X}, \theta^{(i-1)}))] = \sum_{y \in Y} log(p(\mathcal{X}, y \mid \theta) f(y \mid \mathcal{X}, \theta^{(i-1)}))dy$$

Hierbei bezeichnet Y die Menge all derjenigen Werte, die y annehmen kann und $f(y \mid \mathcal{X}, \theta^{(i-1)})$ die Verteilung der unbekannten Daten, die sowohl von der bisher besten Schätzung der Parameter, $\theta^{(i-1)}$, als auch von den Messwerten abhängt.
Der zweite Schritt des EM-Algorithmus, der M-Schritt sucht nun das Maximum dieser Funktion :

$$\theta^i = argmax_\theta \mathcal{Q}(\theta \mid \theta^{(i-1)})$$

$\theta^{(i)}$ ist damit die nächstbessere Näherung für die Dichtefunktion, diese kann verwendet werden, um in einer weiteren Iteration eine noch bessere Näherung für die Dichtefunktion zu ermitteln. Diese Iteration kann beliebig oft wiederholt werden, je nachdem, wie genau das Ergebnis sein soll - jede weitere Iteration verbessert das Ergebnis. Die erste Schätzung der Parameter muß man selbst übernehmen - je nachdem welchen Typ die Dichtefunktion hat, man kann z.B. bei der Normalverteilung annehmen, der Erwartungswert μ sei gleich 0 und die Varianz σ^2 gleich 1.

3.2 Beispiel

Gegeben seien 4 Punkte im 2 dimensionalen Raum, wobei die erste Koordinate des 4.Punktes fehlt. Diese Punkte sollen durch eine 2 dimensionale Normalverteilung entstanden sein - ihre Parameter sind aber unbekannt:

$$\{\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}?\\4\end{pmatrix}\}$$

$$\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}$$

Da angenommen wurde, daß hier eine Normalverteilung vorliegt, kann man die ersten Parameter Schätzen :

$$\theta^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Jetzt kann man die nächstbessere Näherung, θ^1, berechnen :

$$\mathcal{Q}(\theta;\theta^0) = E_{x_{41}}[ln(p(x_g, x_b; \theta \mid \theta^0; \mathcal{D}_g))]$$

$$= \int_{-\infty}^{\infty} \left(\sum_{k=1}^{3} ln(p(x_k \mid \theta)) + ln(p(x_4 \mid \theta)) \right) p(x_{41} \mid \theta^0; x_{42} = 4) dx_{41}$$

$$= \sum_{k=1}^{3} [ln(p(x_k \mid \theta))] + \int_{-\infty}^{\infty} ln(p(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \theta)) \underbrace{\frac{p(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \theta^0)}{\left(\int_{-\infty}^{\infty} p(\begin{pmatrix} x'_{41} \\ 4 \end{pmatrix} \mid \theta^0) dx'_{41} \right)}}_{=K} dx_{41}$$

Hierbei ist x_{41} die erste, unbekannte Koordinate des vierten Punktes.
K ist eine Konstante, die nicht von x_{41} abhängt und kann damit aus dem Integral gezogen werden :

$$\mathcal{Q}(\theta;\theta^0) = \sum_{k=1}^{3} [ln(p(x_k \mid \theta))] + \frac{1}{K} \int_{-\infty}^{\infty} ln(p(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \theta)) \frac{1}{2\pi |\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}|} \exp(-\frac{1}{2}(x_{41}^2 + 4^2)) dx_{41}$$

$$= \sum_{k=1}^{3} [ln(p(x_k \mid \theta))] - \frac{1 + \mu_1^2}{2\sigma_1^2} - \frac{(4 - \mu_2)^2}{2\sigma_2^2} - ln(2\pi\sigma_1\sigma_2)$$

Über die Ableitung nach $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ erhält man folgende Schätzung für die Parameter der Funktion :

$$\theta^{(1)} = \begin{pmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{pmatrix}$$

Bereits nach der 3. Iteration hat man eine Näherung, die sich kaum noch verbessert. Graphisch dargestellt würde das Iterationsverfahren folgendermassen aussehen, die Ellipsen zeigen den Ort der Geschätzten Normalverteilung:

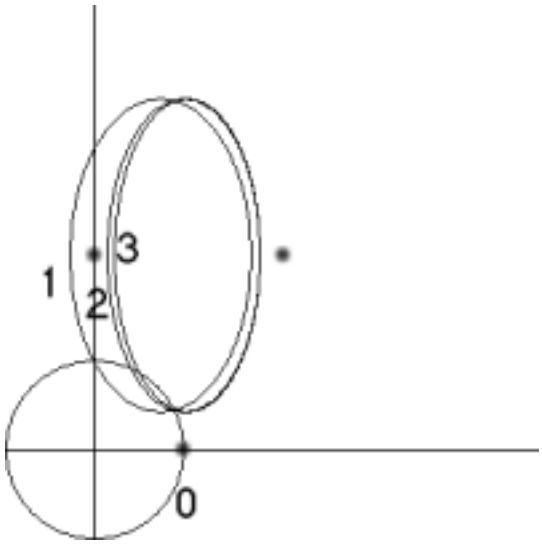

Abbildung 5: Jede Iteration liefer eine bessere Näherung der Parameter der Dichtefunktion

4 Gaussian Mixture Models

4.1 Prinzip

Im folgenden Abschnitt sind alle Rechnungen bis zu einem bestimmten Punkt allgemein und auf beliebige Verteilungen anwendbar, an entsprechender Stelle ist der Wechsel zum Spezialfall Normalverteilung vermerkt.
Man hat einige Messwerte, die in N verschiedene Gruppen aufgeteilt werden können, in denen jeweils alle Punkte enthalten sind, die von der selben Dichtefunktion erzeugt wurden. Alle Punkte insgesamt sind also von einer Dichtefunktion erzeugt, die als Linearkombination der anderen Dichtefunktionen angesehen werden kann :

$$p(x \mid \theta) = \sum_{i=1}^{N} \alpha_i p_i(x \mid \theta_i)$$

Hierbei bezeichnen die θ_i die Parameter der Dichtefunktionen, die die einzelnen Gruppen erzeugt haben und die α_i die Gewichtung der dazugehörigen Dichtefunktion und es gilt :

$$\sum_{i=1}^{N} \alpha_i = 1$$

$$\theta = (\alpha_1, \alpha_2, \cdots, \alpha_N, \theta_1, \theta_2, \cdots, \theta_N)$$

Der Logarithmus der Likelihoodfunktion der unvollständigen Daten, $\mathcal{X}$ wird damit zu :

$$\log(\mathcal{L}(\theta \mid \mathcal{X})) = \log(\prod_{i=1}^{N} p(x_i \mid \theta)) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} a_j p_j(x_i \mid \theta_j)\right)$$

Diese Formel ist unhandlich, da sie den Logarithmus einer Summe enthält. Nun nimmt man an, daß $\mathcal{X}$ nicht vollständig ist, und weitere, unbeobachtete Größen existieren, die jeweils für die einzelnen Messwerte angeben, von welcher Dichtefunktion sie erzeugt wurden, d.h. eine Menge $\mathcal{Y}$, bei der gilt $y_i = k$, wenn der i-te Messwert von der k-ten Verteilung erzeugt wurde. Nimmt man nun an, daß $\mathcal{Y}$ bekannt ist, kann man fortfahren :

$$\log(\mathcal{L}(\theta \mid \mathcal{X}, \mathcal{Y})) = \log(P(\mathcal{X}, \mathcal{Y} \mid \theta)) = \sum_{i=1}^{N} \log(P(x_i \mid y_i)P(y)) = \sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(x_i \mid \theta_{y_i}))$$

Jetzt stellt sich das Problem, daß die Werte in $\mathcal{Y}$ eben nicht bekannt sind, nimmt man allerdings an, daß $\mathcal{Y}$ ein Zufallsvektor ist, kann man fortfahren.
Als nächstes muß ein Ausdruck für die Mischverteilung gefunden werden, also nimmt man an, daß $\theta^g = (\alpha_1^g, \cdots, \alpha_M^g, \theta_1^g, \cdots, \theta_M^g)$ die Parameter der Veteilung $\mathcal{L}(\theta^g) \mid \mathcal{X}, \mathcal{Y})$ sind. Mit θ^g lässt sich $p_j(x_i \mid \theta_j^g)$ berechnen. Mit der Regel von Bayes kann man weiterrechnen :

$$p(y_i \mid x_i, \theta^g) = \frac{\alpha_{y_i}^g p_{y_i}(x_i \mid \theta_{y_i}^g)}{p(x_i \mid \theta^g)} = \frac{\alpha_{y_i}^g p_{y_i}(x_i \mid \theta_{y_i}^g)}{\sum_{k=1}^{M} \alpha_k^g p_k(x_i \mid \theta_k^g)}$$

$$p(y \mid \mathcal{X}, \theta^g) = \prod_{i=1}^{N} p(y_i \mid x_i, \theta^g)$$

Dabei ist $y = (y_1, \cdots, y_N)$ ein Element der unbekannten Daten, $\mathcal{Y}$. Dies ist eine Dichtefunktion, genau wie sie gesucht wurde,und man kann damit den Erwartungswert des Logarithmus der Likelihoodfunktion umschreiben :

$$\mathcal{Q}(\theta, \theta^g) = \sum_{y \in Y} \log(\mathcal{L}(\theta \mid \mathcal{X}, y)) p(y \mid \mathcal{X}, \theta^g)$$

$$= \sum_{y \in Y} \sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(x_i \mid \theta_{y_i})) \prod_{j=1}^{N} p(y_j \mid x_j, \theta^g)$$

$$= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(x_i \mid \theta_{y_i})) \prod_{j=1}^{N} p(y_j \mid x_j, \theta^g)$$

$$= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{l,y_i} \log(a_l p_l(x_i \mid \theta_l)) \prod_{j=1}^{N} p(y_j \mid x_j, \theta^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(a_l p_l(x_i \mid \theta_l) \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{l,y_i} \prod_{j=1}^{N} (y_i \mid x_j, \theta^g)$$

In dieser Form ist $\mathcal{Q}(\theta \mid \theta^g)$ sehr unhandlich, die folgende Nebenrechnung verdeutlicht, wie man den Ausdruck vereinfachen kann :

$$\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p(y_j \mid x_j,\theta^g)$$

$$\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N}p(y_j \mid x_j,\theta^g)\right)p(l \mid x_j,\theta^g)$$

$$=\prod_{j=1,j\neq i}^{N}\left(\sum_{y_j=1}^{M}p(y_j \mid x_j,\theta^g)\right)p(l \mid x_i,\theta^g)=p(l \mid x_i,\theta^g)$$

Weil $\sum_{i=1}^{M}p(i \mid x_j,\theta^g)=1$ kann man $\mathcal{Q}(\theta \mid \theta^g)$ umschreiben :

$$\mathcal{Q}(\theta \mid \theta^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}log(\alpha_l p_l(x_i \mid \theta_l))p(l \mid x_i,\theta^g)$$

$$\sum_{l=1}^{M}\sum_{i=1}^{N}log(\alpha_l)p(l \mid x_i,\theta^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}log(p_l(x_i \mid \theta_l))p(l \mid x_i,\theta^g)$$

Mit dem Wissen, daß die Summe aller α_i gleich 1 ist, bekommt man die folgende Gleichung, in der λ den Lagrange Multiplikator bezeichnet :

$$\frac{\delta}{\delta\alpha_l}[\sum_{l=1}^{M}\sum_{i=1}^{N}log(\alpha_l)p(l \mid x_i,\theta^g)+\lambda\left(\sum_{l}(\alpha_l-1)\right)]=0$$

und damit :

$$\sum_{i=1}^{N}\frac{1}{\alpha_l}p(l \mid x_i,\theta^g)+\lambda=0$$

Damit bekommt man, daß $\lambda=-N$ was bedeutet:

$$\alpha_l=\frac{1}{N}\sum_{i=1}^{N}p(l \mid x_i,\theta^g)$$

In dem von uns angenommenen Fall, daß eine Normalverteilung zu grunde liegt bekommt man:

$$p_l(x \mid \mu_l,\Sigma_l)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_l|^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l))$$

Bildet man von dieser Funktion den Logarithmus und ignoriert dabei alle konstanten Terme, da diese bei der Ableitung sowieso verschwinden würden, bekommt man folgenden Term:

$$\sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(x_i \mid \mu_l, \Sigma_l))p(l \mid x_i, \theta^g)$$

$$= \sum_{l=1}^{M}\sum_{i=1}^{N}\left(-\frac{1}{2}\log(|\Sigma_l|) - \frac{1}{2}(x_i-\mu_l)^T\Sigma_l^{-1}(x_i-\mu_l)\right)p(l \mid x_i, \theta^g) \qquad (*)$$

Leitet man diese Gleichung nach μ_l ab und setzt sie gleich 0, bekommt man :

$$\sum_{i=1}^{N}\Sigma_l^{-1}(x_i-\mu_l)p(l \mid x_i, \theta^g) = 0$$

Aus dieser Gleichung erhält man :

$$\mu_l = \frac{\sum_{i=1}^{N} x_i p(l \mid x_i, \theta^g)}{\sum_{i=1}^{N} p(l \mid x_i, \theta^g)}$$

Um Σ_l zu finden, schreibt man Term $(*)$ folgendermaßen, wobei tr(X) die Spur (engl. trace) der Matrix X bezeichnet und $N_{l,i} = (x_i-\mu_l)(x_i-\mu_l)^T$:

$$\sum_{l=1}^{M}[\frac{1}{2}\log(|\Sigma_l^{-1}|)\sum_{i=1}^{N}p(l \mid x_i, \theta^g) - \frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i, \theta^g)tr(\Sigma_l^{-1}(x_i-\mu_l)(x_i-\mu_l)^T)]$$

$$= \sum_{l=1}^{M}[\frac{1}{2}\log(|\Sigma_l^{-1}|)\sum_{i=1}^{N}p(l \mid x_i, \theta^g) - \frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i, \theta^g)tr(\Sigma_l^{-1}N_{l,i})]$$

Leitet man diesen Term nach Σ_l^{-1} ab, erhält man :

$$\frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i, \theta^g)(2\Sigma_l - diag(\Sigma_l)) - \frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i, \theta^g)(2N_{l,i} - diag(N_{l,i}))$$

$$= \frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i, \theta^g)(2M_{l,i} - diag(M_{l,i}))$$

$$= 2S - diagS$$

Hierbei ist $M_{l,i} = \Sigma_l - N_{l,i}$ und $S = \frac{1}{2}\sum_{i=1}^{N}p(l \mid x_i\theta^g)M_{l,i}$. Indem man die Ableitung von diesem Term gleich 0 setzt erhält man :

$$\sum_{i=1}^{N}p(l \mid x_i, \theta^g)(\Sigma_l - N_{l,i}) = 0$$

$$\Sigma_l = \frac{\sum_{i=1}^{N}P(l \mid x_i, \theta^g)N_{l,i}}{\sum_{i=1}^{N}p(l \mid x_i, \theta^g)} = \frac{\sum_{i=1}^{N}p(l \mid x_i, \theta^g)(x_i-\mu_l)(x_i-\mu_l)^T}{\sum_{i=1}^{N}p(l \mid x_i, \theta^g)}$$

Zusammenfassend sind die Näherungen also folgendermaßen berechnet worden :

$$\alpha_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l \mid x_i, \theta^g)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l \mid x_i, \theta^g)}{\sum_{i=1}^{N} p(l \mid x_i, \theta^g)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid x_i, \theta^g)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l \mid x_i, \theta^g)}$$

In diesem Beispiel wurden der E- und M-Schritt zusammengezogen, was in der Praxis wichtig für die Laufzeit der Programme, die auf diesem Algorithmus basieren, ist.

4.2 Beispiel

Dieses Beispiel zeigt, wie Gaussian Mixture Models zum Clustering verwendet werden können. Zuerst wird eine Schätzung für die Parameter der einzelnen Dichtefunktionen ermittelt. Als nächstes kann dann ermittelt werden, zu welcher Verteilung die Punkte am wahrscheinlichsten gehören. Die folgenden Bilder sind einem Java-Applet entnommen, welches unter [5] zu finden ist. Die Kreise können als Höhenlinien der 2-dimensionalen Dichtefunktionen interpretiert werden.

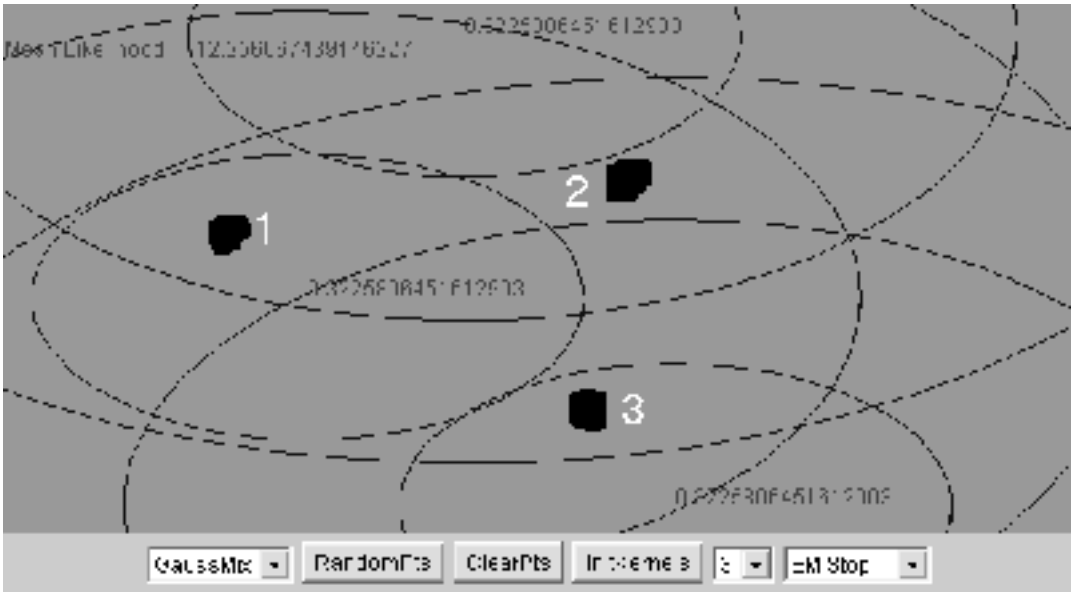

Abbildung 6: 3 Gruppen von Punkten wurden von 3 Dichtefunktionen erzeugt, deren Parameter im ersten Schritt geschätzt werden

Abbildung 7: Nach 7 Iterationen sind die Parameter der Dichtefunktionen bereits konvergiert und die Punkte können Gruppen zugeteilt werden.

5 Zusammenfassung

Die vorgestellten mathematischen Methoden eignen sich sehr gut, um Messpunkte in unterschiedliche Gruppen aufzuteilen, soweit die einzelnen Gruppen durch einfache Ellipsen voneinander trennbar sind. Man erhält in diesem Fall relativ schnell und, je nach Anzahl der Iterationen, sehr gute Ergebnisse. Sind allerdings die Gruppen ineinander verschlungen d.h. können sie nur durch komplexere geometrische Gebilde voneinander getrennt werden, werden die Ergebnisse nicht mehr brauchbar sein, und man muß andere Methoden verwenden, die andere Ansätze verfolgen.

Literaturverzeichnis

[1] *Dr. A. J. Canty*
Statistics 743 - Foundations of Statistics, Vorlesung 6
http://www.math.mcmaster.ca/canty/teaching/stat743/lectures6.pdf,
Seite 73

[2] *J. A. Bilmes*
A gentle tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Seite 1-7

[3] *R. O. Duda, P. E. Hart, D. G. Stork*
Pattern Classification, Kapitel 3.2

[4] *T. Schickinger, A. Steger*
Diskrete Strunkturen, Band 2, Seite 141

[5] *S. Akaho*
EM algorithm for Mixture models (test version)
http://www.neurosci.aist.go.jp/ ˜akaho/MixtureEM.html