号:

我爱机器学习

机器学习干货站

首页 > 机器学习 > THE EM(EXPECTATION–MAXIMIZATION) ALGORITHM 详解

公告栏

期待您的加入

欢迎加入我爱机器学习
QQ14群：336582044

The EM(Expectation–Maximization) 解

smallroof 2012年10月10日 0

最新文章列表

NIPS 2016 — Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

EM，Expectation–Maximization）算法是在概率（

数最大似然估计的算法，其中概率模型依赖于无法观测的隐

最大期望经常用在机器学习和计算机视觉的数据集聚（Data Clustering）领域。

通常来说，聚类是一种无指导的机器学习问题，如此问题描述：给你一堆数据点，让你将它们最靠谱地分成一堆一堆的。聚类算法很多，不同的算法适应于不同的问题，这

Concepts and Techniques 》中的图：

核心，生成了大致两堆点。我们的聚类算法就是需要根据给出来的

正态分布的核心在什么位置，以及分布的参数是多少。这很明显又

，但这次不同的是，答案是连续的且有无穷多种可能性，更糟的是，

哪些点属于同一个正态分布圈的时候才能够对这个分布的参数作出靠

现在两堆点混在一块我们又不知道哪些点属于第一个正态分布，哪些属于第

反过来，只有当我们对分布的参数作出了靠谱的预测时候，才能知道到底哪些点

分布，哪些点属于第二个分布。这就成了一个先有鸡还是先有蛋的问题了。

一方要先打破僵局，说，不管了，我先随便整一个值出

来，看你怎么变，然后我再根据你的变化调整我的变化，然后如此迭代着不断互相推

是 EM 算法。

EM 的意思是 "Expectation-Maximazation"，在这个聚类问题里面，我们是先随便猜一

：如核心在什么地方，方差是多少。然后计算出每个数据点更

还是第二个正态分布圈，这个是属于 Expectation 一步。有了每个数据

点的归属，我们就可以根据属于第一个分布的数据点来重新评估第一个分布的参数（从

nazation 。如此往复，直到参数基本不再发生变化为止。

方法在第二步，根据数据点求分布的参数上面。

在此表示感谢，对其进行总结可得到以下结论。

法，需要多次循环迭代

要目标

公告栏

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

号:

M 步骤

2016深

十大指数级

12月12日

课程中我们学过，统计推断的基本问题可以分为两大类，一类

险验问题。我们主要学习了两类估计问题：点估计和区间估

学习了两类问题：矩估计和最大似然估计。

学习零基础

式为已知，但它的

2016年12月11日

博客，我自己又添加了一些理解此文所需要的预备知识，希望

友，或者对此有困惑的朋友可以留言和我讨论，我们共同学习

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习

QQ14群：336582044

选择月份 ▾

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

SEARCH

号:

016深度
大指数级
12月12日

习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

2016年12月8日

欢迎加入我爱机器学习
QQ14群：336582044

选择月份 ▾

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

$, \dots, x_n)$。如果其所有的二阶偏导数都存在，则 f 函数的 Hessian

$D_i f(x)$，也即如下的形式：↵

$$\begin{bmatrix} \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \cdots & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ & \vdots & \ddots & \vdots \\ \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$ ↵

些概念。设 f 是定义域为实数的函数，如果对于所有的实数 x，

是凸函数。当 x 是向量时，如果其 hessian 矩阵 H 是半正定的（$H \ge 0$），

或者 $H > 0$，那么称 f 是严格凸函数。↵

变量，那么↵

凸函数，那么 $E[f(X)] = f(EX)$ 当且仅当 $p(x = E[X]) = 1$，也就是

说

号:

:016深度

十大指数级

:12月12日

学习零基础

2016年12月11日

公告栏

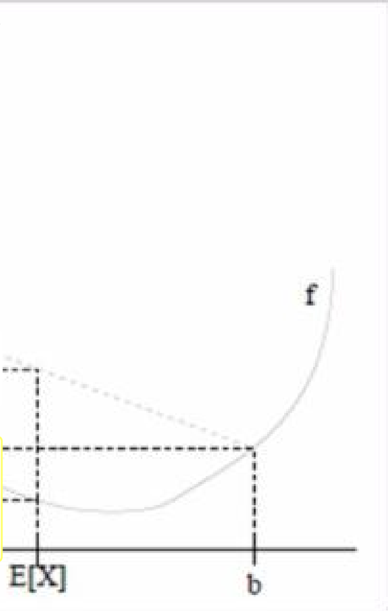
Yoshua

期待您的加入

2016年12月8日

欢迎加入我爱机器学习QQ14群：336582044

选择月份 ▾



最新文章列表

图中，实线f是凸函数，X是随机变量，有0.5的概率是a，有0.5的概率是b。（就像

a和b的中值了，图中可以看到$E[f(X)] \ge f(EX)$成立。

仅当f是（严格）凸函数。

凹函数时，不等号方向反向，也就是$E[f(X)] \le f(EX)$。

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

SEARCH

号:

2016深度…十大指数级

12月12日

学习零基础

2016年12月11日

公告栏

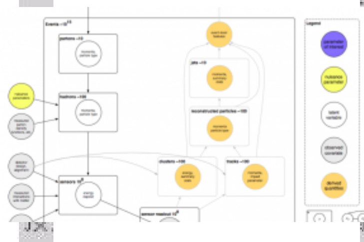
Yoshua

期待您的加入

欢迎加入我爱机器学习QQ14群：336582044

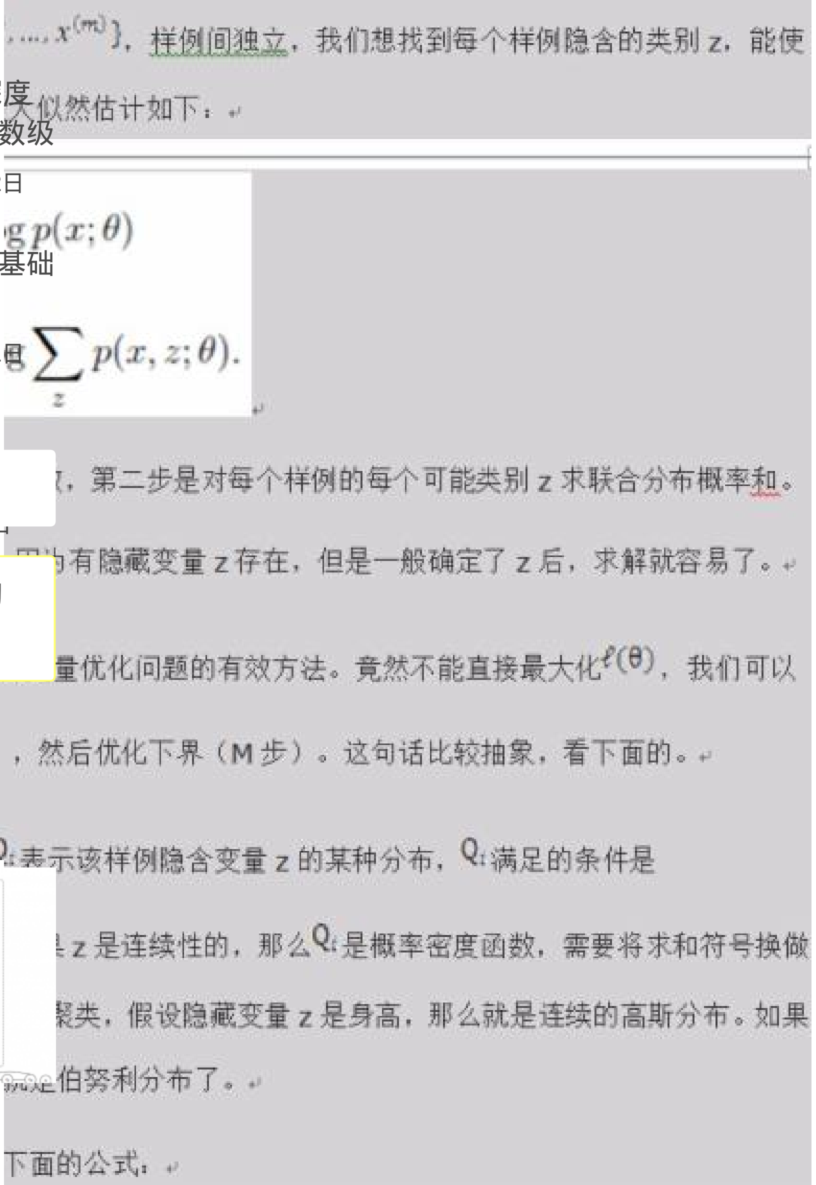
选择月份

最新文章列表



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0



…, $x^{(m)}$}，样例间独立，我们想找到每个样例隐含的类别 z，能使

…似然估计如下：

$$\log p(x;\theta)$$

$$\sum_z p(x,z;\theta).$$

…，第二步是对每个样例的每个可能类别 z 求联合分布概率和。

…为有隐藏变量 z 存在，但是一般确定了 z 后，求解就容易了。

…量优化问题的有效方法。竟然不能直接最大化$\ell(\theta)$，我们可以

…），然后优化下界（M 步）。这句话比较抽象，看下面的。

对于每一个样例 i，让Q_i表示该样例隐含变量 z 的某种分布，Q_i满足的条件是

…z 是连续性的，那么Q_i是概率密度函数，需要将求和符号换做

…聚类，假设隐藏变量 z 是身高，那么就是连续的高斯分布。如果

…就是伯努利分布了。

…下面的公式：

号:

016深度
十大指数级
12月12日

学习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习

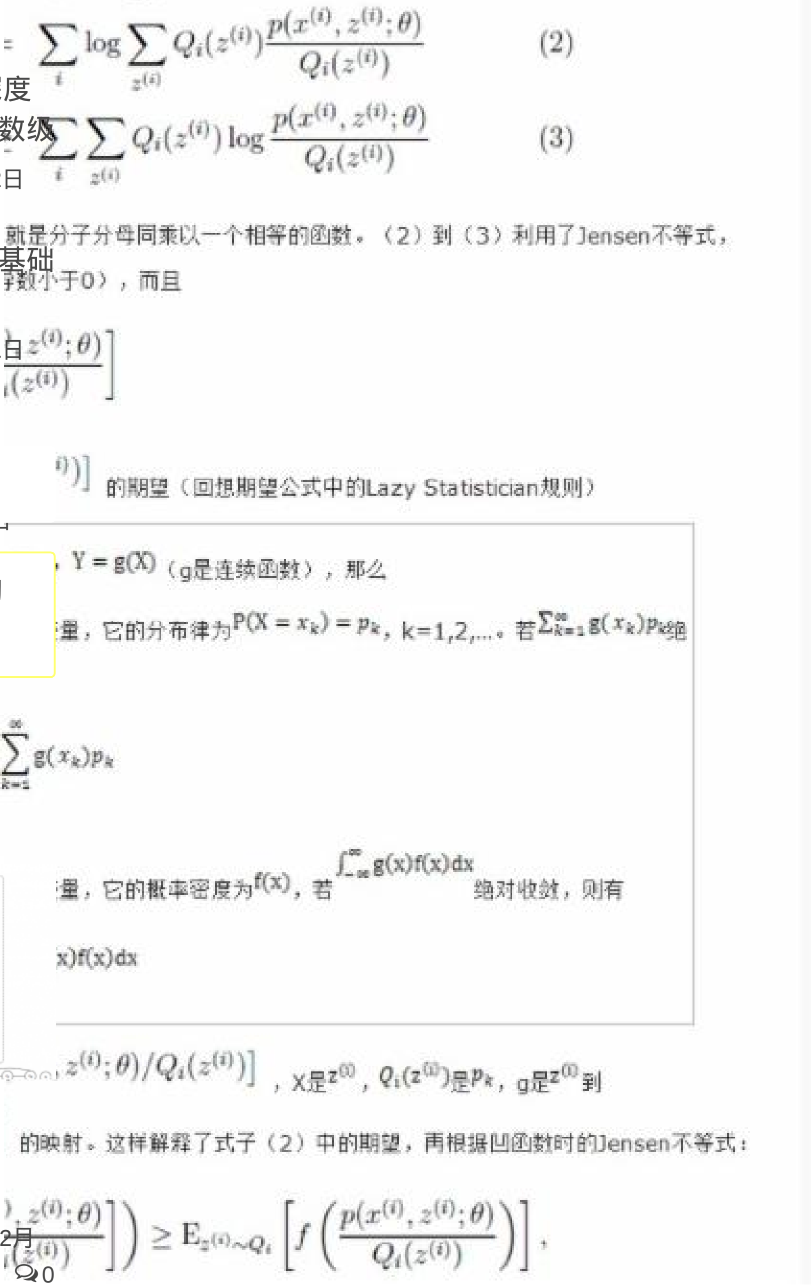
QQ14群：336582044

选择月份 ▼

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0



$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \qquad (2)$$

$$\sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \qquad (3)$$

就是分子分母同乘以一个相等的函数。（2）到（3）利用了Jensen不等式，

数小于0），而且

$$z^{(i)}; \theta) \Big/ (z^{(i)}) \Big]$$

$$)] $$ 的期望（回想期望公式中的Lazy Statistician规则）

$Y = g(X)$（g是连续函数），那么

量，它的分布律为$P(X = x_k) = p_k$，k=1,2,…。若$\sum_{k=1}^{\infty} g(x_k)p_k$绝

$$\sum_{k=1}^{\infty} g(x_k)p_k$$

量，它的概率密度为$f(x)$，若$\int_{-\infty}^{\infty} g(x)f(x)dx$绝对收敛，则有

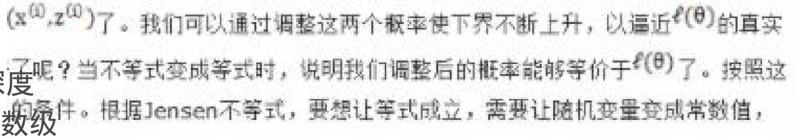
$x)f(x)dx$

$z^{(i)}; \theta)/Q_i(z^{(i)})]$，X是$z^{(i)}$，$Q_i(z^{(i)})$是$p_k$，g是$z^{(i)}$到

的映射。这样解释了式子（2）中的期望，再根据凹函数时的Jensen不等式：

$$z^{(i)}; \theta) \Big/ (z^{(i)}) \Big]\Big) \geq E_{z^{(i)} \sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right)\right],$$

号:


$(x^{(i)},z^{(i)})$了。我们可以通过调整这两个概率使下界不断上升，以逼近$\ell(\theta)$的真实
了呢？当不等式变成等式时，说明我们调整后的概率能够等价于$\ell(\theta)$了。按照这
的条件。根据Jensen不等式，要想让等式成立，需要让随机变量变成常数值，

:016深度
十大指数级
:12月12日

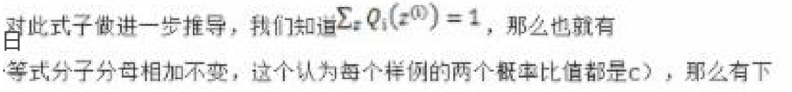
学习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入


对此式子做进一步推导，我们知道$\sum_z Q_i(z^{(i)})=1$，那么也就有
等式分子分母相加不变，这个认为每个样例的两个概率比值都是c），那么有下

$;\theta)$

$,z;\theta)$

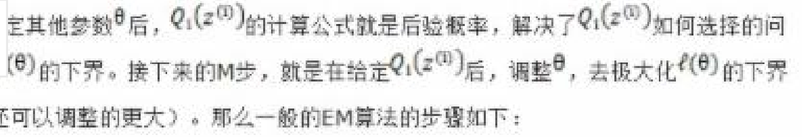
$;\theta)$

$\theta)$

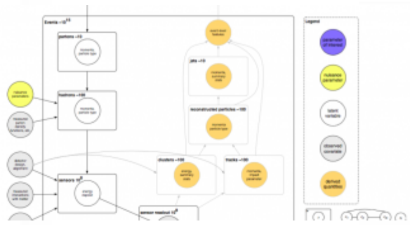
$;\theta)$

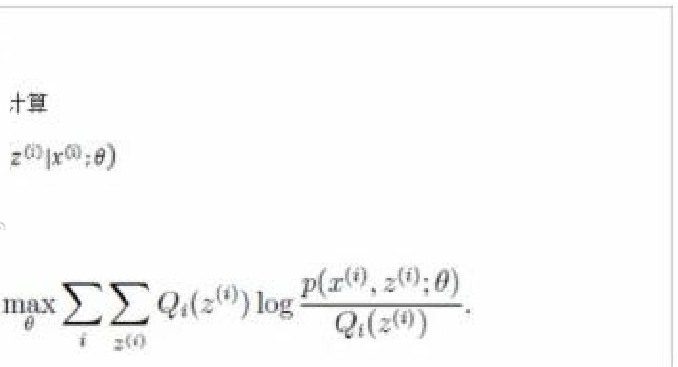
欢迎加入我爱机器学习 QQ14群：336582044

选择月份 ▼


定其他参数θ后，$Q_i(z^{(i)})$的计算公式就是后验概率，解决了$Q_i(z^{(i)})$如何选择的问
(θ)的下界。接下来的M步，就是在给定$Q_i(z^{(i)})$后，调整θ，去极大化$\ell(\theta)$的下界
（在固定$Q_i(z^{(i)})$后，下界还可以调整的更大）。那么一般的EM算法的步骤如下：

最新文章列表




计算

$z^{(i)}|x^{(i)};\theta)$

$$\arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

号:

:016深度

十大指数级

:12月12日

学习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习

QQ14群：336582044

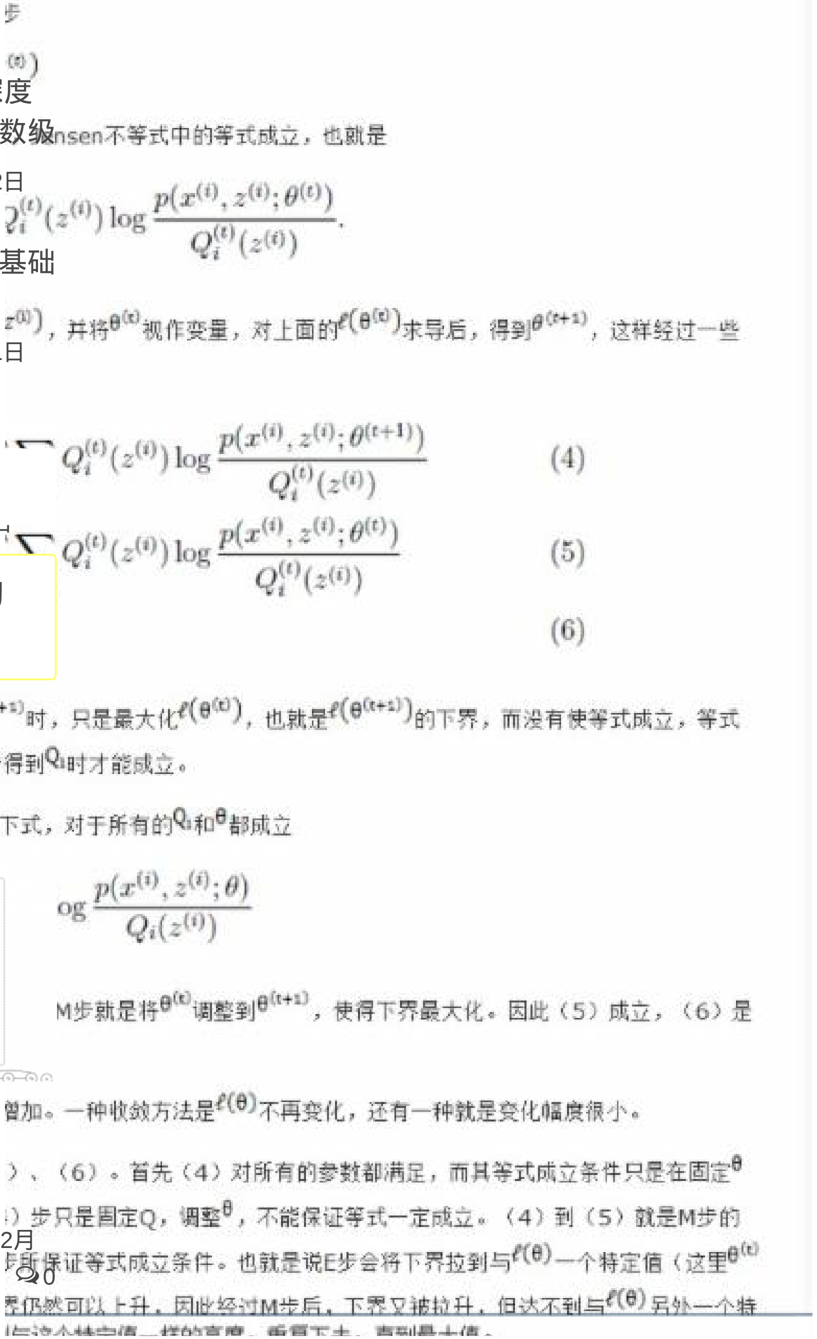
选择月份 ▼

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

0



步

$^{(t)})$

nsen不等式中的等式成立，也就是

$$Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})}.$$

$z^{(i)})$，并将$\theta^{(t)}$视作变量，对上面的$\ell(\theta^{(t)})$求导后，得到$\theta^{(t+1)}$，这样经过一些

$$\sum Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \quad (4)$$

$$\sum Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \quad (5)$$

$$(6)$$

$^{t+1)}$时，只是最大化$\ell(\theta^{(t)})$，也就是$\ell(\theta^{(t+1)})$的下界，而没有使等式成立，等式步得到Q_i时才能成立。

况且根据我们前面得到的下式，对于所有的Q_i和θ都成立

$$og \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}$$

M步就是将$\theta^{(t)}$调整到$\theta^{(t+1)}$，使得下界最大化。因此（5）成立，（6）是

增加。一种收敛方法是$\ell(\theta)$不再变化，还有一种就是变化幅度很小。

）、（6）。首先（4）对所有的参数都满足，而其等式成立条件只是在固定θ）步只是固定Q，调整θ，不能保证等式一定成立。（4）到（5）就是M步的所保证等式成立条件。也就是说E步会将下界拉到与$\ell(\theta)$一个特定值（这里$\theta^{(t)}$界仍然可以上升，因此经过M步后，下界又被拉升，但达不到与$\ell(\theta)$另外一个特与这个特定值一样的高度，重复下去，直到最大值。

号:

016深度

大指数级

12月12日

习零基础

2016年12月11日

Yoshua

公告栏

期待您的加入

欢迎加入我爱机器学习

QQ14群：336582044

选择月份 ▼

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

0

$$Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})},$$

J(Q,θ)，EM可以看作是J的坐标上升法，E步固定θ，优化Q，M步固定Q优

和推导过程，再次审视一下混合高斯模型。之前提到的混合高斯模型的参数φ，μ

得出的，有些没有说明来由。为了简单，这里在M步只给出φ和μ的推导方法。

$$(z^{(i)}=j|x^{(i)};\phi,\mu,\Sigma).$$

类别$z^{(i)}$为j的概率可以通过后验概率计算得到。

$^{(i)})$后最大化最大似然估计，也就是

$$\log\frac{p(x^{(i)},z^{(i)};\phi,\mu,\Sigma)}{Q_i(z^{(i)})}$$

$$=\sum^{m}\sum^{k}Q_i(z^{(i)}=j)\log\frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{Q_i(z^{(i)}=j)}$$

$$\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot\phi_j}{w_j^{(i)}}$$

子，未知参数ϕ_j、μ_j和Σ_j。

$$\log\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot\phi_j}{w_j^{(i)}}$$

$$\sum w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)$$

号:

$\Sigma_l^{-1}x^{(i)} - \Sigma_l^{-1}\mu_l)$

016深度

大指数级

12月12日

习零基础

2016年12月11日

$w_j^{(i)}$，再次使用 $\sum_{j=1}^k \phi_j = 1$，得到

$$w_j^{(i)} = \sum_{i=1}^m 1 = m.$$

公告栏

Yoshua

期待您的加入

的更新公式：

欢迎加入我爱机器学习QQ14群：336582044

选择月份 ▾

微复杂一些，毕竟是矩阵。结果在之前的混合高斯模型中已经给出。

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

号:

2016深度
十大指数级
12月12日

学习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

2016年12月8日

欢迎加入我爱机器学习

QQ14群：336582044

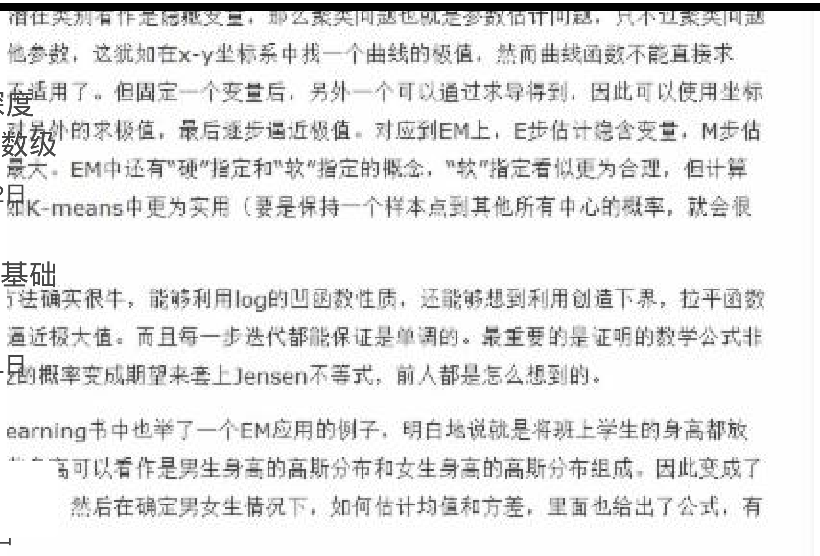
选择月份 ▾

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

0



潜在类别看作是隐含变量，那么聚类问题也就是参数估计问题，只不过聚类问题
他参数，这犹如在x-y坐标系中找一个曲线的极值，然而曲线函数不能直接求
不适用了。但固定一个变量后，另外一个可以通过求导得到，因此可以使用坐标
对另外的求极值，最后逐步逼近极值。对应到EM上，E步估计隐含变量，M步估
最大。EM中还有"硬"指定和"软"指定的概念，"软"指定看似更为合理，但计算
如K-means中更为实用（要是保持一个样本点到其他所有中心的概率，就会很

方法确实很牛，能够利用log的凹函数性质，还能够想到利用创造下界，拉平函数
逼近极大值。而且每一步迭代都能保证是单调的。最重要的是证明的数学公式非
的概率变成期望来套上Jensen不等式，前人都是怎么想到的。

earning书中也举了一个EM应用的例子，明白地说就是将班上学生的身高都放
高可以看作是男生身高的高斯分布和女生身高的高斯分布组成。因此变成了
然后在确定男女生情况下，如何估计均值和方差，里面也给出了公式，有

Q14群：336582044

号:

016深度
大指数级
12月12日
机器学习公众号
习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习
QQ14群：336582044

知器解决异

下一篇文章
支持向量机(SVM)：简介

选择月份 ▾

最新文章列表

或许您还喜欢这些文章

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0

Ma
Me

机器学习

DMX – SQL SERVER 数据挖掘聚类

大数据

社会网络分析：方法与实践

SEARCH

号:

016深度
大指数级
12月12日

高斯混合模型

趣味数据挖掘(7)——团拜会与鸡尾酒会上的聚类

习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习QQ14群：336582044

选择月份 ▼

最新文章列表

NIPS 2016—Daily Highlights

姓:

我爱机器学习(52ml.net) 2016年12月13日

0

号:

016深度

大指数级

12月12日

习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

欢迎加入我爱机器学习 QQ14群：336582044

选择月份 ▾

最新文章列表

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日 0