四、6、期望最大化算法

2016年2月21日 星期日　　上午9:59

一、简介

期望最大化算法，EM，Expectation-Maximization。

是一种参数估计的方法，主要针对模型中有隐变量的情况，它是一种迭代式算法。EM算法就是含有隐变量的极大似然估计。

二、算法思想

（E步）根据前一步估计出的参数θ^t，和已经观察的变量Y推断出隐变量Z^t，在目前情况下的最优值；（M步）把隐变量Z^t看做已知，根据已知隐变量和已观测变量Y（训练数据），对参数θ^t+1做极大似然估计。

观察变量（observable variable）、隐变量（latent variable）

这里待估计的参数隐变量，可以是缺失的训练数据（最初的设计确实是这样的），也可以是模型中的隐变量/隐参数/未知参数。如，PLSA中的“主题参数”，k-means中的“各类均值”。

三、算法具体过程

1、E步

首先，观测数据的似然函数为

$$P(Y \mid \theta) = \sum_{Z} P(Z \mid \theta) P(Y \mid Z, \theta)$$

对观测数据+隐藏数据的概率分布是

$$P(Y, Z \mid \theta)$$

我们通过对Z计算期望，最大化观测数据的对数“边际似然”（merginal likelihood）

ln（P（Y | θ））= ln（Σ_z P（Y, Z | θ））

我们根据当前参数θ^t推测隐变量的分布Z^t，得到P（Z|Y，θ^t）

在根据Z^t，计算LL（θ|Y，Z），得到关于Z的期望

$$Q(\Theta \mid \Theta^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^t}\, LL(\Theta \mid \mathbf{X}, \mathbf{Z})$$

这个是EM算法中很关键的Q函数。

Q函数就是完全数据的对数似然log（P（Y,Z|θ））），关于

给定观测数据Y，和当前时刻参数θ下，对未观测数据Z的条件概率分布P（Z|Y，θ）的期望

2、M步

根据Z的期望，寻找参数最大化期望似然。

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)})$$

四、CS299对EM算法的讲解（更加本质）

1、琴生不等式 Jensen's inequality

A 维基百科

对于凸函数，有以下推论：过一个凸函数上任意两点所作割线一定在这两点间的函数图象的上方

$$tf(x_1) + (1-t)f(x_2) \geq f\left(tx_1 + (1-t)x_2\right), 0 \leq t \leq 1.$$

这个公式和机器学习-周志华中对凸函数的定义类似，

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f(x_1) + f(x_2)}{2}$$

Jensen's inequality 是一种更一般化的表示。

Jensen's inequality 有不同的表示形式，CS229中介绍的是概率论的版本

B CS229

如果f是任一凸函数，必有

$$\mathrm{E}[f(X)] \geq f(\mathrm{E}X)$$

具体推导也没有细说，只是“凸函数 -> 对于任意x，二阶

导>=0 -> hessian矩阵H半正定(H>=0)”

2、EM算法

首先，我们需要表示出优化目标。

这里优化目标是，找出一组参数θ，让最大似然观测数据x（这里所谓的x，是广义的数据，包括有监督训练的数据x和标签y，也包含无监督训练的x）

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x;\theta)$$

$$= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta)$$

其次，表示出E步的目标。

E步，我们是根据现有参数θ^T 和 已观测的数据x，推测隐变量。由于隐变量是未知的，我们假设隐变量z的某种分布是Q_i(z) 。注意，这里的Q函数和统计机器学习中的不一样！！统计机器学习中是期望，这里是z的某种分布。

把Q_i(z)带入优化目标，并且通过化简，利用Jensen's inequality，找到优化目标在这一步迭代中的下界。

可以看做是：

E步：基于现有参数和已观测数据，找到让优化目标的下界的表示形式；

M步：在下界已知的基础上，通过参数优化调整，最大化这个下界。

（我们的目标是最大化，我们希望让优化目标的下界尽量大，使得最坏情况也不是很坏）

化简过程如下：

$$\begin{aligned}\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) \\ &= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \\ &\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\end{aligned}$$

$$f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right]$$

我们可以看出来，优化目标l（θ）的下界，最终是f（P（x，z；θ）/Q）的期望。我们在E步“要找到优化目标l（θ）下界”的目标基本达成，下界就是这个期望。

第三，找到Q函数

下界是期望，不过期望中的Q是z的某种分布，具体是哪种分布？

这里求解的过程中，由于我们是要对θ求解，所以假设期望相对于z是一个常量，所以有

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

进而推导出

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta)$$

Q可以看做是，z基于参数θ在观测数据x之后的后验概率。

第四、优化参数θ，找到让期望最大的一组参数θ

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

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

3、坐标下降角度的理解

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

优化目标可以写为，整个EM的过程可以参数是对J的坐标下降。

E步是把θ看做常量，优化Q（或者说隐变量z）；M步是隐变量z看做常量，优化θ。

五、应用

1、k-means

简介

k-means（k-均值）聚类，KM是一种无监督的学习方法。

E（根据现有参数，估计隐变量）：

通过现有的类划分方式（已知参数）和各样本点位置（已观测数据），计算各个类中心位置（隐变量），并且找出到各个节点最近的类中心。

M（根据隐变量，对参数进一步估计）：

通过各节点到各个类中心的距离，重新对类进行划分（进一步估计参数），每个节点分配到距离类中心最近的类中，相当于最小化各节点到各类中心距离。这里待估计参数是类的划分方式。

2、高斯混合模型

简介

聚类对每个样本hard的分到某个类中，混合模型可以理解为聚类的一种soft的形式，混合模型也是有k个分模型，相当于k

y)

个类。k个分模型对每个样本点都有一个概率密度值，也就是说每个样本都有一定的可能属于每个类。

高斯混合模型是最常用的混合模型，假设各样本点在k个分模型上服从高斯分布。

定义

定义 9.2（高斯混合模型） 高斯混合模型是指具有如下形式的概率分布模型：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k) \tag{9.24}$$

其中，α_k 是系数，$\alpha_k\geqslant 0$，$\sum_{k=1}^{K}\alpha_k=1$；$\phi(y|\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{9.25}$$

称为第 k 个分模型.

EM算法

和k-means的思路很相似

类的划分方式是待估计的隐变量

在k-means中，这个划分的方式是hard的：某个样本属于哪一类

在混合模型中，这个方式是soft的，是一个概率值，并且服从多项分布：某个样本数据某一类的概率是多少

高斯模型的参数均值、方差是待估计参数θ

E（根据现有参数，估计隐变量）：

根据目前k个分模型的均值、方差（现有参数θ^T）和样本点分布（已观测数据），估计“各样本数据各类的概率”（隐参数）

M（根据隐变量，对参数进一步估计）：

根据样本点分布（已观测数据）、“各样本数据各类的概率”（隐参数）计算当前条件下，各类均值和方差的最大值。（分别对均值、方差求导，令导数=0）

这部分在统计机器学习和CS229中都有讲解。

3、 PLSA

简单列出公式

E step: p(z|w,d)=p(z|d)*p(w|z)/sum_z{p(z|d)*p(w|z)}

M step:

p(z|d)=sum_w{p(z|w,d)}/sum_wz{p(z|w,d)}==sum_w{p(z|w,d)}/w_count

详见 http://blog.jqian.net/post/plsa.html

六、推广与拓展

广义期望最大化算法（GEM，generalized expectation-maximization），比普通的EM条件宽松一些，只要求M步求解θ^t+1时，找到一个相比θ^t有改善的θ^t+1就行，不要求是极大似然的。（推导详见统计学习方法P168）

（统计学习方法，也有详细的描述）