EM 算法学习笔记

欧高炎

1 预备知识

拉格朗日乘法

设给定二元函数 $z=f(x,y)$和附加条件$\varphi(x,y)=0$，为寻找 $z=f(x,y)$在附加条件下的极值点，先做拉格朗日函数 $L(x,y)=f(x,y)+\lambda\varphi(x,y)$，其中$\lambda$为参数。求 $L(x,y)$对 x 和 y 的一阶偏导数，令它们等于零，并与附加条件联立，即

$L'_x(x,y)=f'_x(x,y)+\lambda\varphi'_x(x,y)=0$，

$L'_y(x,y)=f'_y(x,y)+\lambda\varphi'_y(x,y)=0$，

$\varphi(x,y)=0$

由上述方程组解出 x，y 及λ，如此求得的(x,y)，就是函数 $z=f(x,y)$在附加条件$\varphi(x,y)=0$下的可能极值点。

贝叶斯公式

P(X,Y)表示X,Y的联合概率，有如下公式P(X,Y)=P(Y|X)P(X)，由于P(X,Y)=P(Y,X)，于是我们得到P(Y|X)P(X)=P(X|Y)P(Y)，将左边P(X)移到右边得到

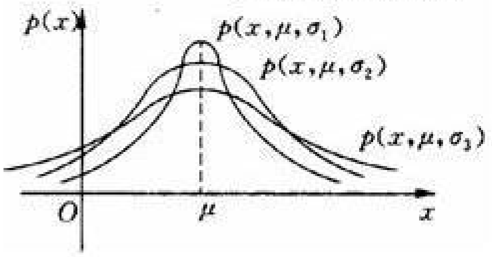
$$p(Y|X)=\frac{p(X|Y)p(Y)}{p(X)}$$

这就是贝叶斯公式，其中 P(Y|X)称为后验分布，P(X)称为先验分布，P(X|Y)称为似然函数。

高斯分布

高斯分布又成为正态分布。对于随机变量 X，其概率密度函数如图所示。称其分布为高斯分布或正态分布，记为 N（μ，σ2），其中为分布的参数，分别为高斯分布的期望和方差。

高斯分布概率示意图



$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(\sigma>0)$$

高斯分布公式

相对熵（Kullback-Leibler 发散度）

设 p(x)和 q(x)为随机变量 X 的两个不同的分布密度，则他们的相对熵定义为：

$$\mathrm{D}(p\,\|\,q)=\sum_x p(x)\log\frac{p(x)}{q(x)} \qquad (1)$$

相对熵的一些性质：D(p||q)不是对称的，也不满足三角不等式，但是可以用它来度量两个分布之间的差异度。D(p||q)的取值始终非负，当且仅当 p=q 时取值为零。

其他的一些准备知识包括求导（偏导、向量求导、矩阵求导）。这里就不一一列出了。

2.EM 算法的一般形式

EM 算法是极大似然估计的一种经典算法。主要用于解决数据量不足和似然函数中含有隐形变量的情形。设 X 为观测变量、Z 为隐形变量，θ 为参数。则：

$$p(X,Z|\theta)=p(X|\theta)p(Z|X,\theta)$$

$$\begin{aligned}=>\ln p(X|\theta)&=\ln p(X,Z|\theta)-\ln p(Z|X,\theta)\\&=\ln p(X,Z|\theta)-\ln q(Z)-[\ln p(Z|X,\theta)-\ln q(Z)]\\&=\ln\frac{p(X,Z|\theta)}{q(Z)}-\ln\frac{p(Z|X,\theta)}{q(Z)}\end{aligned}$$

等式两边同时乘以去$q(Z)$,并对z求和，有

$$\sum_z q(Z)\ln p(X|\theta)=\sum_z[q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}-q(Z)\ln\frac{p(Z|X,\theta)}{q(Z)}]$$

由于Z与平$p(\mathrm{X}|\theta)$独立，且$\sum_z q(Z)=1$，则

$$\begin{aligned}\ln p(X|\theta)&=\sum_z[q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}-q(Z)\ln\frac{p(Z|X,\theta)}{q(Z)}]\\&=\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}+\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta)} \qquad (2)\end{aligned}$$

我们的目标是使似然函数$\ln p(X|\theta)$最大化。结合公式（1）中相对熵的定义和性质，上式（2）右边的第二项可以看成 q(Z)和 p(Z|X,θ)的相对熵。由于相对熵始终非负，所以$\ln p(X|\theta)$的下界为公式（2）右边的第一项$\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$。E M算法的基本思想是不断提高$\ln p(X|\theta)$的下界$\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta)}$，直到收敛。

我们做以下标记：

$$L(q,\theta)=\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)} \qquad (3)$$

$$D(p\,\|\,q)=\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta)} \qquad (4)$$

则有 $\ln p(X|\theta) = L(q,\theta) + D(p\,\|\,q)$ （5）

下面来以此来介绍E M算法的两个步骤。

E 步骤：固定一个θ，记为$\theta^{(n)}$，求一个分布 q(Z),使得$L(q,\theta)$最大化。

分析：由于 Z 和$\ln p(X|\theta)$无关，即公式（5）的左边相对于 Z 来说是是一个常量。所以使$L(q,\theta)$最大化，等价于使$D(p\|q)$最小化。由相对熵的定义，我们知道$D(p\|q)$最小为 0，当且仅当 p=q，即我们所要求的分布为$q(Z)=p(Z|X,\theta^{(n)})$。

M 步骤：固定 q(Z),求一个θ,记为$\theta^{(n+1)}$，使得似然函数的下界$L(q,\theta)$最大化。

分析：由 E 步骤已知$\theta^{(n)}$值，则

$$L(q,\theta)=\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$$

$$L(q,\theta)=\sum_z p(Z|X,\theta^{(n)})\ln p(X,Z|\theta)-\sum_z p(Z|X,\theta^{(n)})\ln p(X,Z|\theta^{(n)})+L(q,\theta^{(n)})$$

上式中后两项相对于θ为常量，即求$L(q,\theta)$最大值转化为求

$\sum_z p(Z|X,\theta^{(n)})\ln p(X,Z|\theta)$的最大值。假设此时的解为$\theta^{(n+1)}$。

EM 算法不断重复 E 步骤和 M 步骤，直到$\ln p(X|\theta)$收敛。

附录 1 EM 算法的另一种分析方法

由公式（2）

$$L(\theta)=\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}+\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta)} \qquad (2)$$

则有：

$$\begin{aligned}
&L(\theta)-L(\theta^{(n)})= \\
&\sum_z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}-\sum_z q(Z)\ln\frac{p(X,Z|\theta^{(n)})}{q(Z)} \\
&\quad+\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta)}-\sum_z q(Z)\ln\frac{q(Z)}{p(Z|X,\theta^{(n)})} \\
&=\sum_z q(Z)\ln p(X,Z|\theta)-\sum_z q(Z)\ln p(X,Z|\theta^{(n)})+\sum_z q(Z)\ln\frac{p(Z|X,\theta^{(n)})}{p(Z|X,\theta)} \qquad (6)
\end{aligned}$$

在 E 步骤中，已经计算得，$q(z)=p(Z|X,\theta^{(n)})$，代入公式（6），有：

$$
\begin{aligned}
&L(\theta)-L(\theta^{(n)}) \\
&=\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta)-\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta^{(n)}) \\
&\qquad+\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln\frac{p(Z\mid X,\theta^{(n)})}{p(Z\mid X,\theta)} \qquad (7)
\end{aligned}
$$

式（7）右边最后一项可以看作 $p(Z\mid X,\theta^{(n)})$ 和 $p(Z\mid X,\theta)$ 的相对熵，由相对熵的非负性质，有：

$$
\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln\frac{p(Z\mid X,\theta^{(n)})}{p(Z\mid X,\theta)}\geq 0
$$

所以式（7）可以转化为：

$$
L(\theta)\geq\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta)-\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta^{(n)})+L(\theta^{(n)}) \qquad (8)
$$

观察式（8），右边为 $L(\theta)$ 的一个下界，同时我们可以看到，右边三项相对于 θ 均可以视为常量，所以 $L(\theta)$ 的一个下界为 $\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta)$，这个下界即是所谓的Q 函数。记为：

$$
Q(\theta;\theta^{(n)})=\sum\nolimits_z p(Z\mid X,\theta^{(n)})\ln p(X,Z\mid\theta) \qquad (9)
$$