

# 曲线拟合之概率回访

emacsun

## 目录

1 回忆	1
2 贝叶斯估计	1
3 概率模型	3

## 1 回忆

我们之前在 [曲线拟合的过程中](#) 采用的是最小化误差函数的方法来确定拟合的 $\mathbf{w}$ 系数。拟合的多项式为：

$$y(x, \mathbf{w}) = w_0 + w_1x + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1.1)$$

误差函数为：

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad (1.2)$$

在那里，我们为了解决过度拟合问题还采用了一种叫做正则化的方法。今天，我们从概率的角度来审视多项式曲线拟合问题。通过概率角度，我们可以更深入的理解误差函数和正则化。

## 2 贝叶斯估计

曲线拟合的目的是对于给定输入 $x$ 估计输出 $t$ 。当然，我们有训练数据：对于 $\mathbf{x} = (x_1, \dots, x_N)^T$ ，对应的值是 $\mathbf{t} = (t_1, \dots, t_N)^T$ 。对于任意的新的输入值 $x$ ，我们可以把对 $t$ 的估计写成一个条件概率估计。什么样的概率密度函数最合适



呢？**正态分布最合适**！即，对于给定的输入 $x$ ，我们假设 $t$ 具有正态分布，均值是 $y(x, \mathbf{w})$ ：

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (2.1)$$

其中 $\beta$ 是精度参数，等于(2.1)的方差的倒数，即 $\beta^{-1} = \sigma^2$ 。

式(2.1)的示意图如1所示。

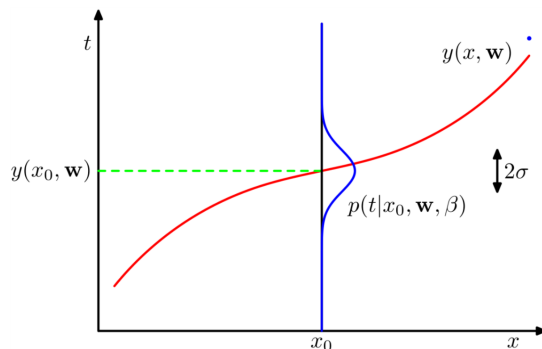


图 1: 式(2.1)的示意图

从图1中可以看出蓝色曲线就是假设的高斯分布。而精度值 $\beta$ 体现了分布的方差。

现在我们用不同于**以往**的方法来求 $\mathbf{w}, \beta$ 。如果所有的训练数据都是从(2.1)中独立获得的，也就是说假设 $\mathbf{t}$ 是独立同分布的。那么关于 $\mathbf{t}, \mathbf{x}$ 分布的似然函数是：

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \quad (2.2)$$

其中：

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad (2.3)$$

把(2.3)带入(2.2)，并对(2.2)左右两端取自然对数：

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \quad (2.4)$$

我们从(2.4)推出曲线拟合系数 $\mathbf{w}$ 的最大似然解。显然，我们可以忽略(2.4)的后两项，因为这两项与 $\mathbf{w}$ 没有关系。另外我们也发现 $\mathbf{w}$ 的最大似然解与等号右边第一项的系数没有关系，这个系数只是起到缩放作用，我们还可以把 $\beta/2$ 用 $1/2$ 代替。最大化似然函数等效于最小化负的似然函数。最后我们发现最大化(2.4)和最小化(1.2)是一回事儿。因此(1.2)所示的误差函数最小值的解是假定噪声为高斯噪声的最大似然解。



另外我们还可以使用最大似然准则求得精度值 $\beta$ 的最优解。把(2.4)当做 $\beta$ 的函数，我们有 $\beta$ 的最大似然解满足：

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 \quad (2.5)$$

所以我们可以先求得 $\mathbf{w}$ 的最大似然解 $\mathbf{w}_{ML}$ ，然后求得 $\frac{1}{\beta_{ML}}$ 。如此，我们便得到了所需高斯分布的两个重要参数，对于任意输入 $x$ ，我们可以使用这个模型来估计 $t$ 。

### 3 概率模型

现在我们有了 $\mathbf{w}_{ML}$ ， $\frac{1}{\beta_{ML}}$ ，我们就有了一个概率模型：

$$p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1}) \quad (3.1)$$

对于给定的 $x$ 我们用(1.1)来计算其均值 $y(x, \mathbf{w}_{ML})$ ，然后用(3.1)给出 $t$ 的估计。

现在让我们更深入的理解这个问题。首先，我们引入对(1.1)中系数 $\mathbf{w}$ 的一个先验估计：

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2} \mathbf{w}^T \mathbf{w}\right) \quad (3.2)$$

其中 $\alpha$ 是先验概率分布的精度。 $M+1$ 是 $M$ 阶多项式中的系数个数。 $\alpha$ 控制着模型的参数（式(1.1)的参数），我们称 $\alpha$ 为超参数。据贝叶斯理论 $\mathbf{w}$ 的后验分布与先验分布和似然函数成比例，即：

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha) \quad (3.3)$$

利用给定的训练数据，我们通过最大化后验概率来确定 $\mathbf{w}$ 。这个准则叫做最大后验概率准则(maximum posterior, MAP). 结合(3.3)(2.4)(3.2)，我们发现最大后验概率等效于最小化(3.4)：

$$\frac{\beta}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} \quad (3.4)$$

即，最大化后验概率等效于最小化带有正则参数 $\lambda = \alpha/\beta$ 的均方误差函数。