线性回归推导、误差分析、极大似然估计 + 正态分布推最小二乘

线性回归算法推导

前言

上一个文章我们说了，使用$J(\theta)$推导出正规方程，但是我们还没有了解这个$J(\theta)$是怎么来的，我们来研究一下它

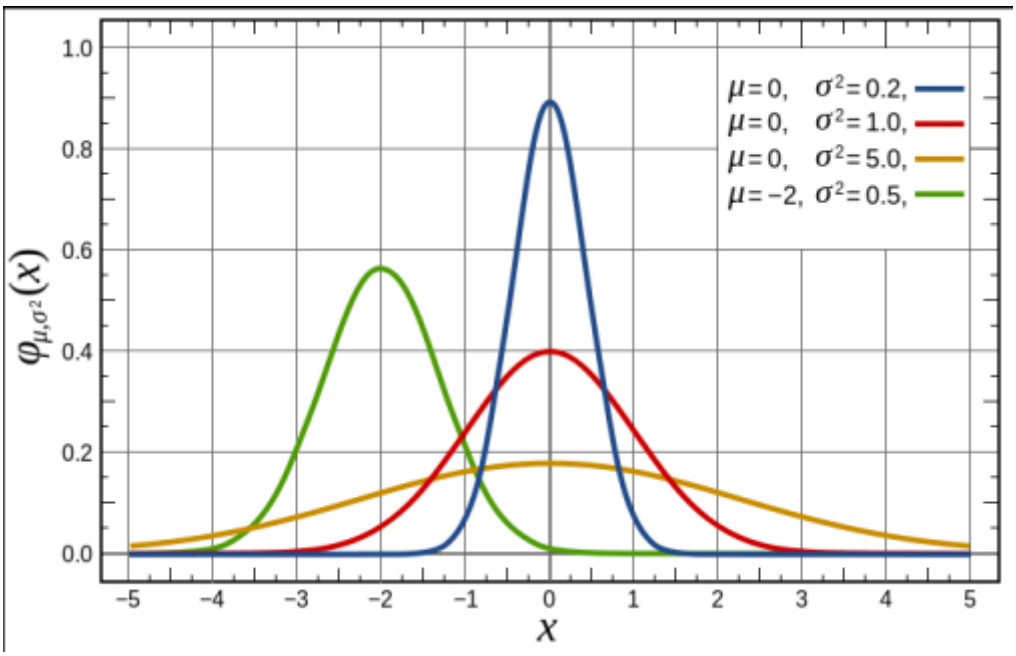
深入理解回归

回归简单来说就是“回归平均值”（regression to the mean）。但是这里的 mean 并不是把历史数据直接当成未来的预测值，而是会把期望值当作预测值。追溯根源回归这个词就是一个叫高尔顿的人发明的，他通过大量地观察数据发现：父亲比较高、儿子比较高；父亲比较矮、儿子也就会比较矮。

在这些相关的数据中，就存在着一定的规律规则，这就是**正态分布（高斯分布）**，高斯深入研究了正态分布，最终推导出了线性回归的原理：最小二乘法

正态分布公式：

https://bkimg.cdn.bcebos.com/formula/6c9d4bf224281678113352b45366d4cd.svg



误差分析

误差ε_i等于第i个样本实际的值y_i减去预测的值$\hat{y}_i$，公式可以表达如下：

$$\varepsilon_i = |y_i - \hat{y}_i| \tag{1}$$

$$\tag{2}$$

$$\varepsilon_i = |y_i - W^T x_i| \tag{3}$$

假定所有的样本误差都是**独立的**，有上下的震荡，震荡认为是随机变量，最够多的随机变量叠加后形成的分布，他服从的就是正态分布，因为他是正常情况下的分布，也就是高斯分布。**均值**是某一个值，**方差**是某一个值。方差我们暂时不管，均值我们总有办法让他去等于零的，因为我们合理是有截距b，所以误差我们就可以认为是独立分布的，$1 \leq i \leq n$ ，服从均值为 0，方差为某定值的高斯分布。机器学习中我们假设误差符合均值为0，方差就是正态分布！（“均值” 是指样本误差分布的平均值，也就是所有样本误差值的算术平均数。）

最大似然估计（计算过程中的一些方法）

最大似然估计（maximum likeihood estimation，MLE）一种重要而普遍的求估计的方法。最大似然估计明确地使用概率模型，其目标是寻找能够以高概率产生观察数据的系统发生树。最大似然估计是一类完全基于统计的系统发生树重建方法的代表。

举个例子吧：

假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。我们想知道罐中白球和黑球的比例，但我们不能把罐中的球全部拿出来数。现在我们可以每次任意从已经摇匀的罐中拿一个球出来，记录球的颜色，然后把拿出来的球再放回罐中。这个过程可以重复，我们可以用记录的球的颜色来估计罐中黑白球的比例。假如在前面的一百次重复记录中，有七十次是白球，请问罐中白球所占的比例最有可能是多少?

70%?

详细推导（高中数学）：

1. 最大似然估计，计算
2. 白球概率是p，黑球是1-p(罐子中非黑即白)
3. 罐子中取一个请问是白球的概率是多少? p
4. 罐子中取两个球，两个球都是白色，概率是多少? p^2
5. 罐子中取10个球，9个是白色，一个是黑色，概率是多少呢?

$C_{10}^{9}(1-p) = C_{10}^{1}p^{9}(1-p)$

6. 那么一百次重复记录中，有七十次是白球，请问罐中白球所占的比例最有可能是多少?

$C_{100}^{30}p^{70}(1-p)^{30}$

那么什么是**最大似然估计**，就是什么时候 p 最大，求导！

去掉 C_{100}^{30} 常数，求导：

$$P' = 70 * p^{69} * (1-p)^{30} + p^{70} * 30 * (1-p)^{29} * (-1)$$

令 P' 等于 0：

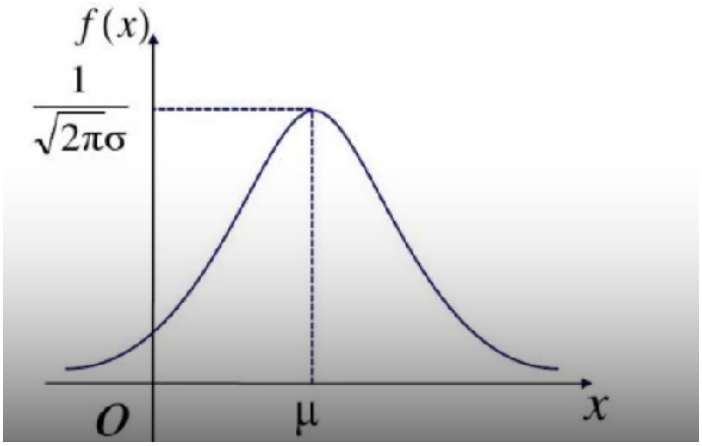
求出 p=70% √

高斯分布—概率密度函数

常见的连续概率分布是正态分布，也叫高斯分布，而这正是我们所需要的，其密度函数如下：

μ是平均值，σ是标准差，知道二者就可以计算它的密度

https://bkimg.cdn.bcebos.com/formula/6c9d4bf224281678113352b45366d4cd.svg



公式如下 f(x)=p：

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

随着参数 μ和σ 的变化、概率分布也产生变化。下面重要的步骤来了，我们把一组数据误差出现的**总似然**，也就是一组数据之所以对应误差出现的在**整体可能性**表达出来了，因为数据的误差我们假设服从高斯分布，并通过**截距**项来平移整体分布的位置从而使得 $\mu = 0$，所以样本的误差我们可以表达概率密度函数的值如下：

$$f(\varepsilon|\mu=0,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon-0)^2}{2\sigma^2}} \tag{4}$$

简化如下： (5)

$$f(\varepsilon|0,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}} \tag{6}$$

误差总似然

类似于前边所介绍的黑白球问题的**累乘**

$$P = \prod_{i=0}^{n} f(\varepsilon_i|0,\sigma^2) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} \tag{7}$$

根据前面的公式$\varepsilon_i = |y_i - W^T x_i|$误差公式可以推导出如下公式：

$$P = \prod_{i=0}^{n} f(\varepsilon_i|0,\sigma^2) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - W^T x_i)^2}{2\sigma^2}} \tag{8}$$

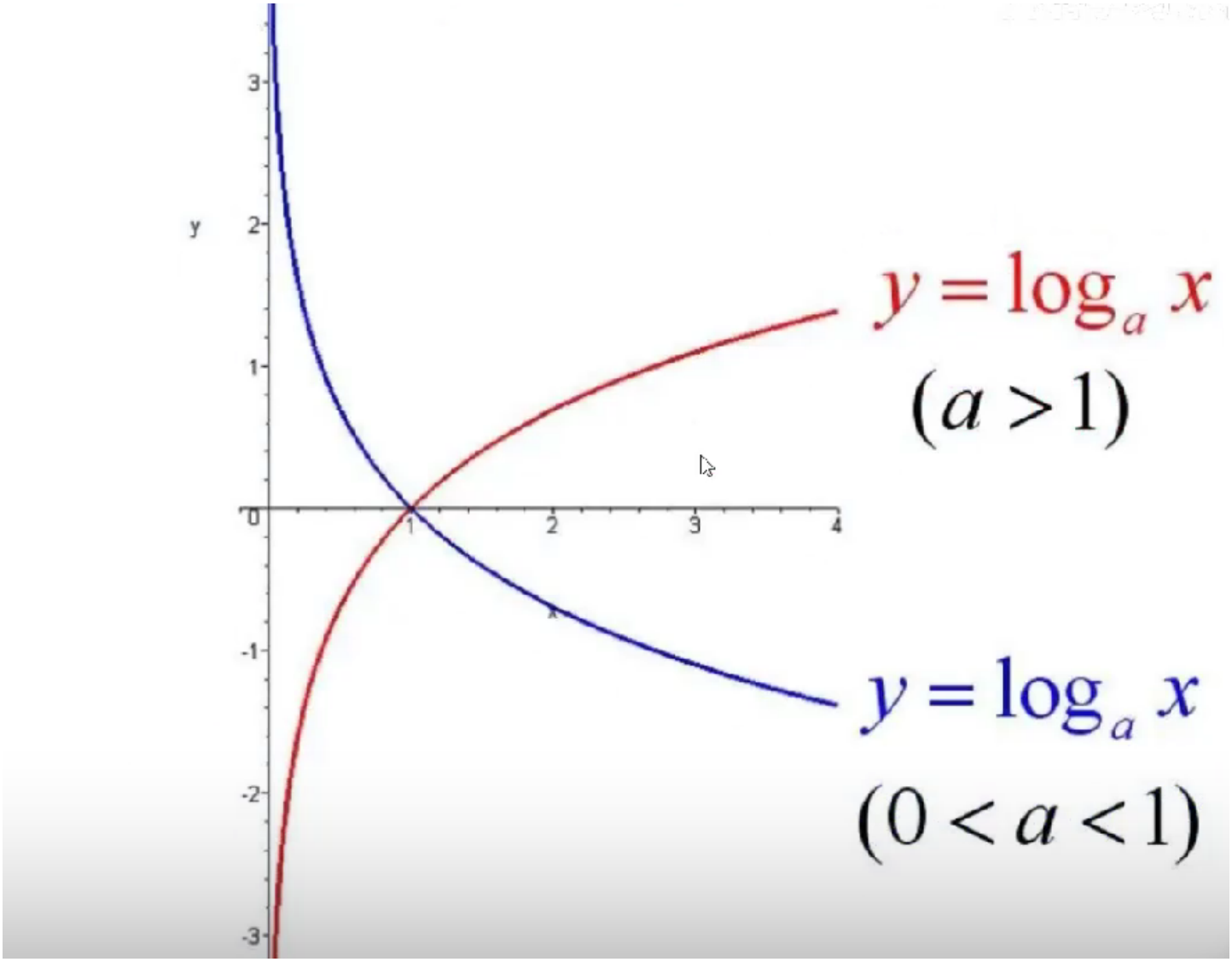
y_i表示真实值 (9)

$W^T x_i$代表预测值 (10)

公式中的**未知变量**就是 W^T ，即方程的系数，系数包含截距~如果，把上面当成一个方程，就是概率P关于W的方程。其余符合，都是常量。

$$P = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - W^T x_i)^2}{2\sigma^2}} \tag{11}$$

现在问题，就变换成了**求最大似然问题**，不过累乘的最大似然计算十分麻烦，接下来我们通过**求对数**把**累乘问题**转化为累加问题：



最小二乘

$$P_W = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_i - W^T x_i)^2}{2\sigma^2}} \tag{12}$$

根据对数，单调性，对上面的公式求自然底数e的对数进行缩放，效果不变：

$$log_e(P_W) = log_e(\prod_{i=0}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - W^T x_i)^2}{2\sigma^2}}) \tag{13}$$

接下来 log 函数继续…数学上连乘是个十分的麻烦，即使交给计算机也会含用大量的计算，所以：

$$log_a(XY) = log_aX + log_aY \tag{14}$$
$$log_a\frac{X}{Y} = log_aX - log_aY \tag{15}$$
$$log_aX^n = n * log_aX \tag{16}$$
$$log_a(X_1X_2...X_n) = log_aX_1 + log_aX_2 + ... + log_aX_n \tag{17}$$
$$log_xx^n = n(n \in R) \tag{18}$$
$$log_a\frac{1}{X} = -log_aX \tag{19}$$
$$log_a\sqrt[x]{Ny} = \frac{y}{x}log_aN \tag{20}$$

所以，就求出：

$$\begin{aligned} log_e(P_W) &= log_e(\prod_{i=0}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-W^Tx_i)^2}{2\sigma^2}}) \\ &= \sum_{i=0}^{n}log_e(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-W^Tx_i)^2}{2\sigma^2}}) \end{aligned} \tag{21}$$

好我们继续：

$$\begin{aligned} &= \sum_{i=0}^{n}(log_e\frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i-W^Tx_i)^2}{2\sigma^2}) \\ &= \sum_{i=0}^{n}(log_e\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}(y_i-W^Tx_i)^2) \\ &\frac{1}{2}(y_i-W^Tx_i)^2\text{这不就是最小二乘法?} \end{aligned} \tag{22}$$

上面公式是最大似然求对数后的变形，其中π、σ都是常数，而$(y_i - W^Tx_i)^2$一定大于等于零，上面求最大值问题，就可转化为如下求最小值问题：

$L(W) = \frac{1}{2}\sum_{i=0}^{n}(y^{(i)} - W^Tx^{(i)})^2$，其中$L$代表$Loss$，损失函数，损失越小上面似然也就越大

有的书本上公式，也可以这么写，用$J(\theta)$表示一个意思，θ的角色就是W：

$$J(\theta) = \frac{1}{2}\sum_{i=0}^{n}(y^{(i)} - W^Tx^{(i)})^2 = \frac{1}{2}\sum_{i=0}^{n}(\theta^Tx^{(i)} - y^{(i)})^2$$

进一步提取：

$$J(\theta) = \frac{1}{2}\sum_{i=0}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$$

其中：

1. $\check{y} = h_\theta(X) = X\theta$ 表示全部数据，是矩阵，X表示多个数据，进行矩阵乘法时，放在前面
2. $\check{y}_i = h_\theta(x^{(i)}) = \theta^T x^{(i)}$ 表示第 i 个数据，是向量，所以进行乘法时，其最终一方需要转置

一位内最大似然公式中有个**负号**，所以最大似然变成了最小化后边负号部分。到此，我们就清楚了推导最小二乘（损失函数 $J(\theta)$ ），从公式上就可以看出名字的由来。

注：需要注意的是，最小二乘法作为损失函数，没有除以总样本数m，而均方误差(MSE)，除以总样本数m

由此得出谷神星的轨道！（

归纳总结升华

这种最小二乘法估计，其实我们就可以认为，假定了误差服从正态分布，认为样本误差的出现是随机的，独立的，使用最大似然估计思想，利用损失函数最小化MSE就可以求出最优解！所以反过来说，如果我们的数据误差不是互相独立的，或者不是随机出现的，那么就不适合假设为正态分布，就不能用正态分布的概率密度函数带入到总似然的函数中，故而就不能用MSE作为损失函数去求解最优解了！所以，最小二乘法不是万能的。

还有譬如结社误差从泊松分布，或者其他分布那就得用其他分布的概率密度去推导出损失函数了。

所以有时我们也可以把线性回归看成广义的线性回归。比如，逻辑回归，泊松回归都属于广义线性回归的一种，这里我们线性回归可以说是最小二乘线性回归。