一文搞懂损失函数（中）

P9工作法
分布式架构、上百人团队管理、全球支付实践与AI技术

上文 一文搞懂损失函数（上）讨论了回归问题的损失函数，这篇继续讨论分类问题的损失函数。这里分为概率论和信息熵两部分，所以还需要再拆解为两篇。这篇为概率论部分，下一篇为交叉熵部分。

分类是把一堆数据分为几类，要么是one-hot（后面会单独有文章来分析什么是one-hot）的表达形式，要么就是概率（比如一张图片是猫的概率为90%，是狗的概率是10%）。当然one-hot的表达中，1也可以表达成概率为100%。所以统筹起来用概率来做分类问题是比较好的。

MSE法

上文已经提到MSE，它就是 $(y-\hat{y})^2$ ，但这个数字范围理论上是0到正无穷的，而概率是0到1之间的，要实现这个转换还需要一个函数来做变换，即将 $(y-\hat{y})^2$ 转换为0到1之间的数，而这个函数也是有的，那就是 $y=\frac{1}{1+e^{-z}}$ ，这就是sigmoid函数。

按照之前MSE的方法用链式法则进行求偏导，得到如下的式子。

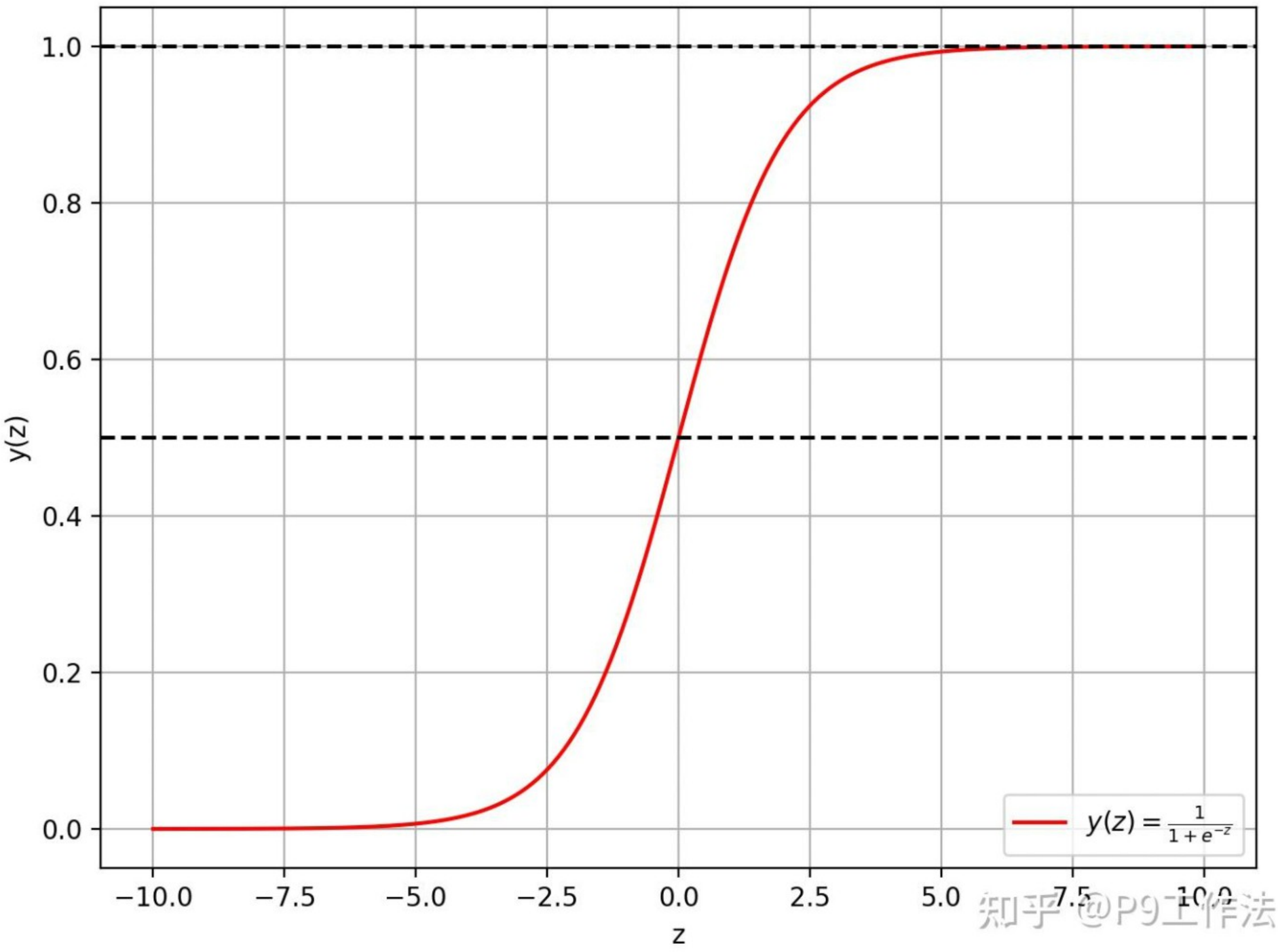
$$\mathrm{MSE}=\frac{1}{2}(y-\hat{y})^2$$

$$y=\frac{1}{1+e^{-z}}$$

$$z=Wx+b$$

$$\frac{\partial \mathrm{MSE}}{\partial w}=\frac{\partial \mathrm{MSE}}{\partial a}\cdot\frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial w}=(y-\hat{y})\cdot y'(z)\cdot x$$

从sigmoid的函数图像来看，就知道这个函数在0和1这两个地方的梯度是很趋近于0，这会导致后续通过梯度下降算法无法很好学习到好的参数。



当维度很多的时，MSE不是凸函数，可能出现局部最优解。所以综合这两个缺点，其实MSE是不适合用来做分类问题的损失函数。而更好的做法就是要用到概率相关方法，这就是要涉及到概率和信息论相关知识。

概率法

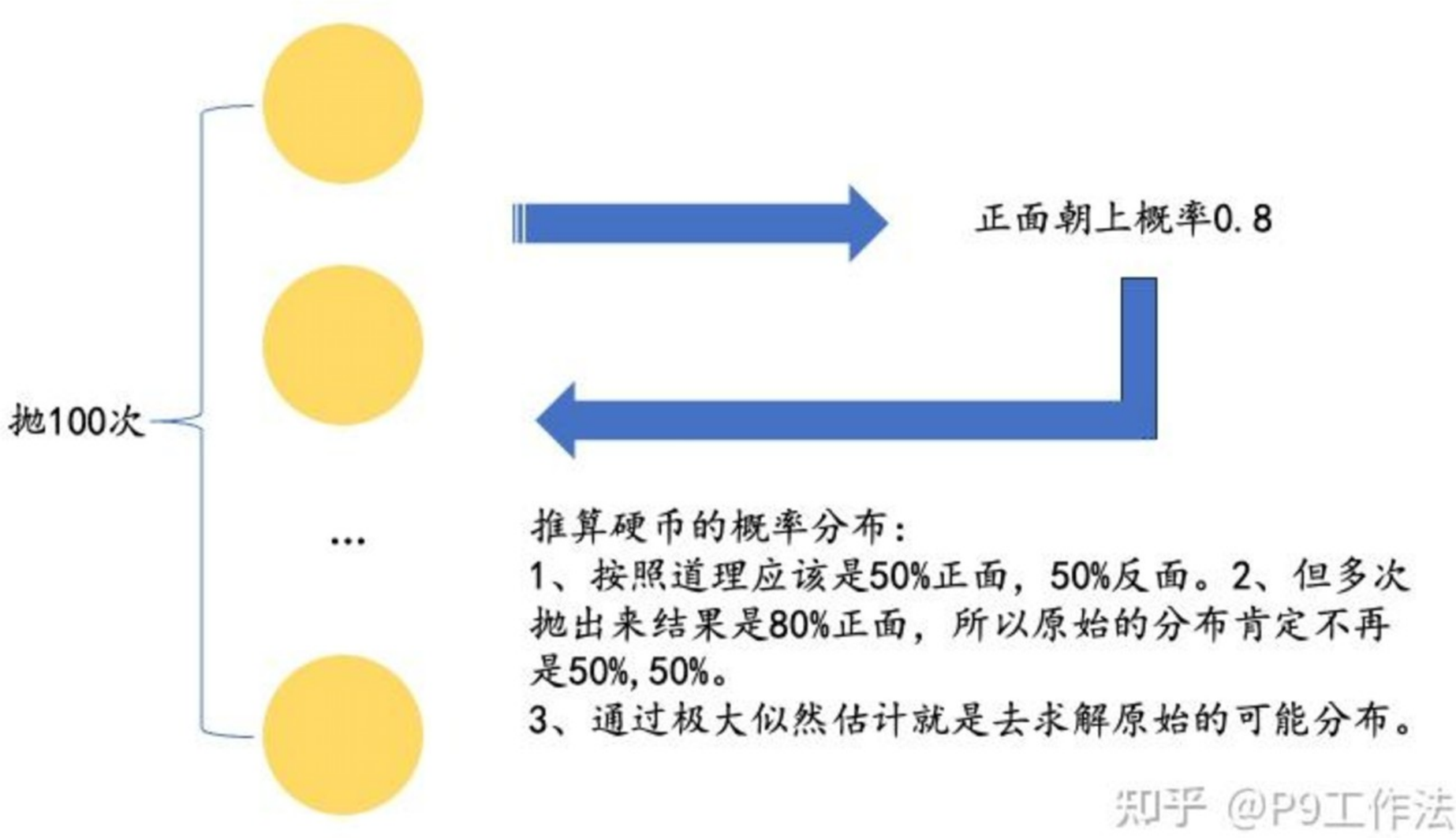
极大似然法

极大似然也叫极大似然估计，MLE（Maximum Likelihood Estimation）

概率与似然的区别

概率是一件事情发生的可能性，一旦事件发生概率就变成确定性的事件，比如抛硬币，出现正面的概率是50%，但只要抛出去就变成了确定性事件。

而似然是基于已经确定的结果来推测是某个原因导致的可能性。极大似然估计就是反推最具有可能或者说最大概率导致这些样本结果出现的模型参数。



比如抛1000次硬币，出现正面的概率是0.8，反面是0.2，那说明这个硬币应该是特殊制造（比如质地不均匀等），那么似然估计就是去计算这个质地是怎么分布的。

假设w是模型参数，y为事件发生的结果。那么概率就表达为P(y|w），表示在参数为w的条件下，y发生的概率。而似然表达为L(w|y），表示在已知道结果y的情况下去求得w。概率P是关于y的函数，似然L是关于w的函数。

极大似然的本质

极大似然其实是推断w为什么时，结果y最有可能发生。以抛硬币为例，假设出现正面的概率为w，出现反面的概率为1-w，这时候并不知道w的具体值。

但是可以通过抽样得到结果，比如结果为：“正面，反面，正面，反面，正面”。那么正面的概率为0.6，反面为0.4.

$$L(w \mid y) = w \times (1-w) \times w \times (1-w) \times w = w^3 \times (1-w)^2$$

取对数运算：

$$\ln L(w \mid y) = 3 \ln w + 2 \ln(1-w)$$

$$\frac{d \ln L(w|y)}{dw} = \frac{3}{w} - \frac{2}{1-w}$$

令其等于0，即

$$\frac{3}{w} - \frac{2}{1-w} = 0$$

求得：

$$w = \frac{3}{5}$$

也就是当 $w = \frac{3}{5}$ 时，正面概率为0.6的可能性最大。

这样的求解过程得出来的w为0.6时，最优可能出现上面 “正面，反面，正面，反面，正面” 这样的概率结果。

我们再把这个含义延展一下，从头梳理下思路。

- 假设有一个抛硬币这件事本来就有一个概率分布，可能有如下几个分布可能：

1. 分布1：得出正面的概率为0.8，反面为0.2。
2. 分布2：得出正面的概率为0.9，反面为0.1。
3. 分布3：得出怎么的概率为0.6，反面为0.4。

如果是在上面的三个概率分布下，哪一个最有可能抛出来 “正面，反面，正面，反面，正面” 这样一个结果呢。我们可以计算下：

1. 对于分布1：概率为0.8*0.2*0.8*0.2*0.8 = 0.02048
2. 对于分布2：概率为0.9*0.1*0.9*0.1*0.9 = 0.00729
3. 对于分布3：概率为0.6*0.4*0.6*0.4*0.6 = 0.03456

从结果来看显然是分布3是最有可能出现 “正面，反面，正面，反面，正面” 这样一个结果的。这其实也印证了上面的求解过程。
这个很好理解，因为不管是哪一个分布，其实都可以扔出来 “正面，反面，正面，反面，正面” 这样一个结果（想想，对于分布1来说，正面概率为0.8，反面概率为0.2，但要扔出来正、反、正、反、正还是可能的，只是需要些运气）。
当然这里我们得出来抛硬币的这件的概率分布是0.6，和现实中认为的0.5有差距呢，这个怎么解释。两种解释：

1. 第一种就是因为我们的观察结果太少，只有5次，所以算出来为0.6。我们通过扩大观察样本，就能够得出来0.5；
2. 第二种就是这个硬币构造有问题，比如质地不均匀，就会产生这样的概率分布。

回到似然估计的本质，这是不是意味着，我们拿着现实既定的结果去推定了理想中的概率。对于神经网络而言，是不是可以认为通过训练数据去推定了神经网络中的概率分布，使得其最有可能出现现实世界的概率。
通过这个可以解释清楚为什么需要更多的数据去训练模型，因为数据越多推定出来的概率分布就越接近理想中的概率分布。

极大似然函数

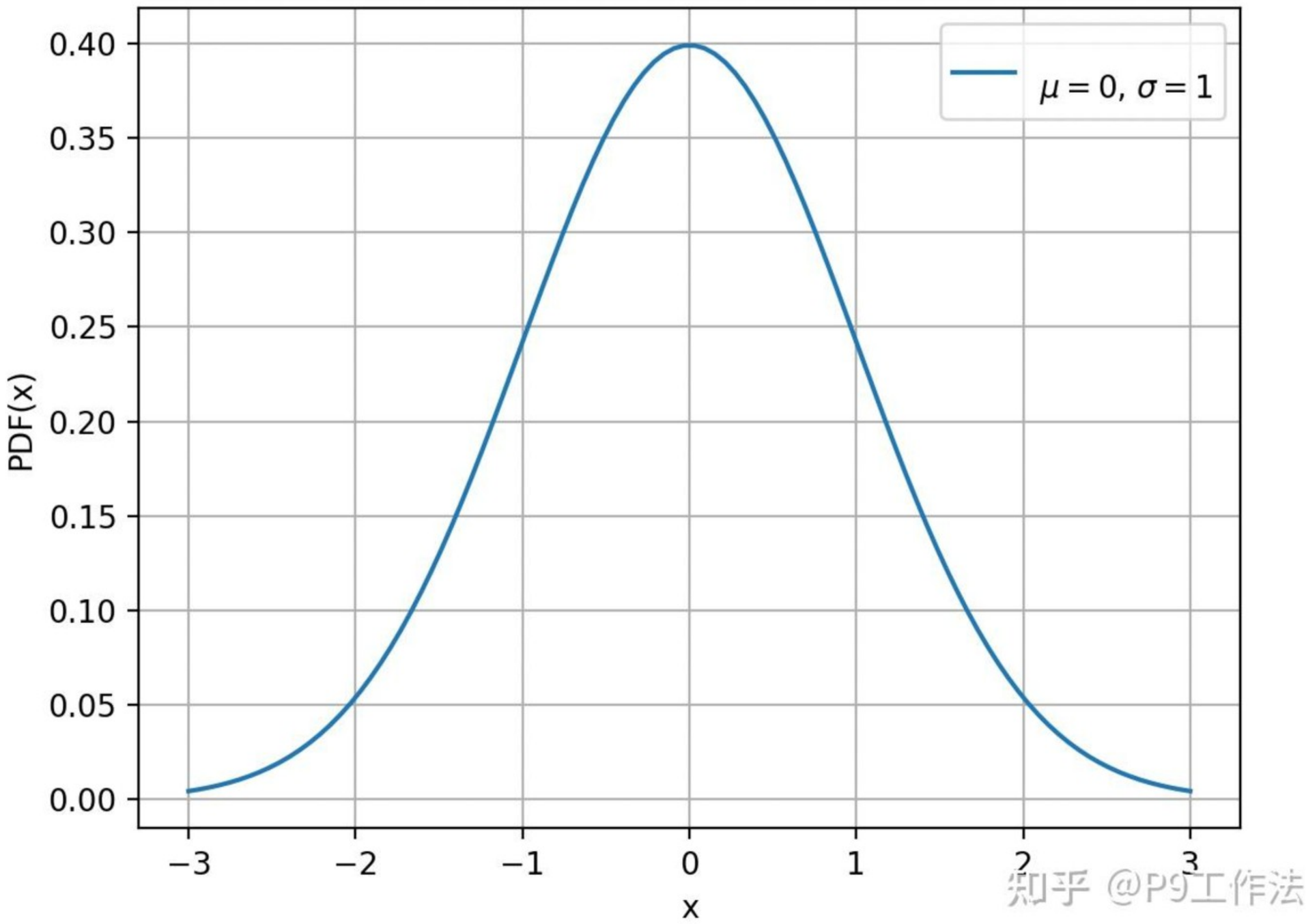
极大似然的损失函数表达为：

$$L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(x_i)$$

这个是连乘符号，为了方便计算，取对数得到如下函数：

$$\log L(\theta \mid x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log P_\theta(x_i)$$

我们假设这些预测值与真实值之间的误差是服从高斯分布的，



而根据中心极限定理（大量独立随机变量的和（或平均值）趋向于正态分布（高斯分布）的现象，不论这些随机变量本身遵循什么样的概率分布），认为服从高斯分布是没有任何问题的。则，上面的似然函数可以写成

$$P(y_i \mid x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i;\theta))^2}{2\sigma^2}\right)$$

取对数为：

$$\log P(y_i \mid x_i, \theta) = -\frac{(y_i - f(x_i;\theta))^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})$$

对所有数据点求和为：

$$\log L(\theta \mid x_1, x_2, \ldots, x_n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 + \text{constant}$$

忽略常数项，极大似然估计的目标可以简化为最小化平方误差之和

$$\max_\theta \log L(\theta \mid x_1, x_2, \ldots, x_n) \equiv \min_\theta \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2$$

所以说，最小二乘法其实高斯分布下似然函数的特殊情况。