插值与拟合

1. 插值

插值是通过给定的一组点去还原原来的函数曲线。

* 1. Lagrange插值法

对于给定的n+1组离散的点：，拉格朗日插值法的想法是找到一种开关，每当插值出的多项式曲线通过所给定的某一个离散点时，就使得其他项失效，仅留，且，因此可以定义为：

从该拉格朗日基数可以看出，对于点，其值为1，而当x为其余所给离散点的x值时，该基数为0。

拉格朗日插值法所插值的函数可以被写为：

1. 拟合

拟合是指根据一组数据，这组数据因为是通过采样等方式得到的，因而有概率存在误差，因此只尝试找到一个函数或曲线，使其最好地描述这组数据的趋势或规律，是一种尝试去逼近原来的函数的方法。

* 1. 最小二乘法

最小二乘法的思想是通过缩小我们所假设的曲线与采样的函数值之间的误差到最小，从而得到拟合曲线的方法，其数学表达为：

一般是通过求各个参数的偏导数并令其为0，接触法方程而得出所求函数结果。

* 1. Weierstrass定理

除了通过使用最小二乘法后采用法方程求解或者梯度下降法求解外，还可以通过使用“多项式拟合”的方式求解曲线。之所以能够使用多项式方法，主要是因为有Weierstrass定理做支撑，即任意连续函数，都存在多项式能够表达这个曲线：

* 1. RBF函数

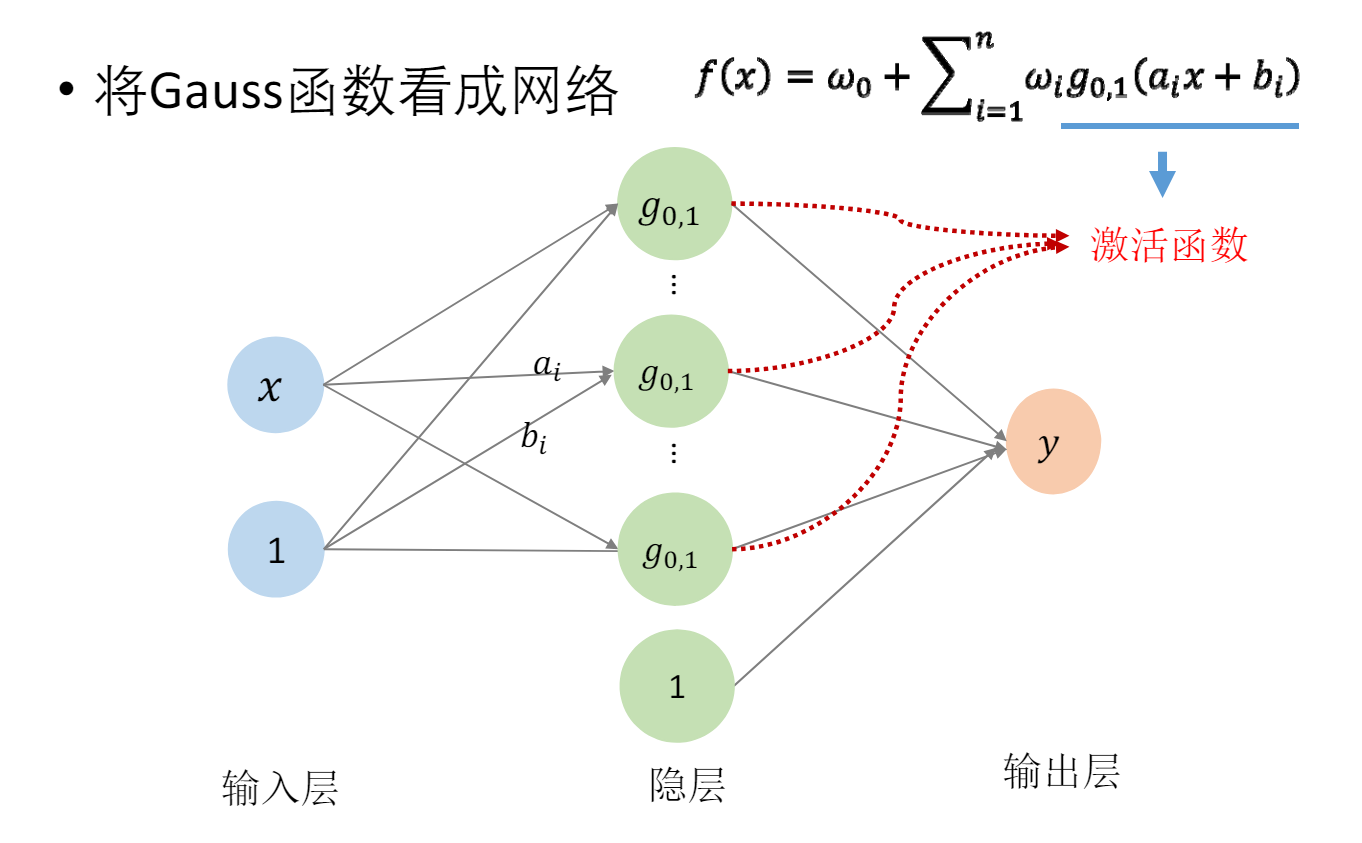
除了使用多项式作为基函数（如Bernstein多项式）去拟合函数，还可以使用Gauss函数作为基函数去拟合，这就是RBF函数：

其中，当的时候则是标准正态分布，实际上可以这样定义，把所有的都变为类似于标准正态分布的形式：

这样定义基函数，就可以将e的次数也转化为一个线性的方程：

并使原先每个前的系数吸收掉，则重新写出的公式为：

这就是RBF函数，之所以要将改写为以标准正态本部为底的形式，主要是为了适应人工智能领域的训练，也就是说，可以将标准正太分布函数视作一个激活函数，如图所示。

图2.1 RBF神经网络

1. 过拟合

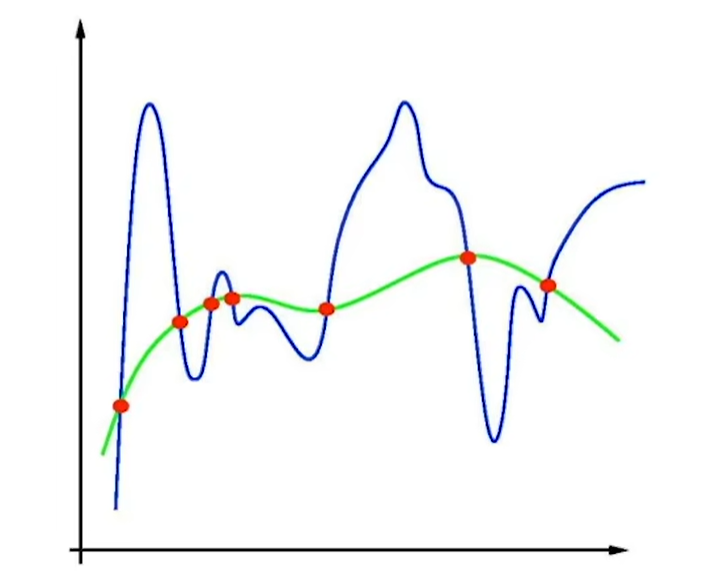
虽然通过拟合的方法可能可以得到与采样点误差很小的函数，但该函数的其他部分与实际函数图形相差甚远（图3.1），这种现象就是过拟合

图3.1 过拟合

* 1. 过拟合的修正方法
     1. 岭回归

通过最小二成可以得到回归系数向量，将其模乘以一个惩罚系数作为惩罚项，通过考虑这个惩罚项来进行正则化的方法被称为岭回归：

* + 1. 稀疏正则化

稀疏正则化考虑到了有些时候所得函数的基函数可能有很多线性相关的项，因此为了减少函数的过完备性，必须减少系数不为0的基函数的个数：