Homework 01 Polynomial Curve Fitting

王晓捷 (11521053)

May 13, 2016

1 引言

拟合是指针对一些已有的数据，根据其规律将其方程化，从而在将来有未知数据输入时，我们也能够得到其相应的预测值。方程化的形式就是预测值与其对应的真实值之间的误差越小越好。如何衡量误差？如何将数据进行方程化？

2 方法概述

将数据进行方程化的方法有很多，本次作业主要使用多项式拟合方法。多项式拟合的任务是假设给定数据由M次多项式函数生成，选择最有可能产生这些数据的M次多项式函数，即在M次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数。通常情况下，设M次多项式为

$$f_M(x,w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{2.1}$$

式中x是单变量输入，$w_0, w_1, ..., w_M$是$M+1$个参数。

解决这一问题的方法可以是这样的。首先确定模型的复杂度，也就是确定多项式的次数；然后在给定模型复杂度的情况下，按照经验风险最小化的策略，求解参数。具体地，求以下经验风险最小化[1]：

$$\min_w L(w) = 1/2 \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 \tag{2.2}$$

对式(2.2)求最小值，最直接的方法便是对每一个参数求偏导，偏导为0的点即为极值点。从另外一种角度出发，可以将其用矩阵的形式表示，对其求导数，运用到线性代数的知识去求解，如迹、矩阵求导等知识，最终可以得到

$$W = (X^T X)^{-1} X^T Y \tag{2.3}$$

根据上式得到的结果，我们可以很容易的使用Python通过进行一些简单的矩阵运算从而获得拟合模型。

但假设我们假设的模型复杂度比真模型的复杂程度要高很多，为了得到较好的拟合效果，一种典型的方法是在经验风险上加一个正则化项，这一项一般是模型复杂度的单调递增函数，模型越复杂，该项的值就越大。比如，正则化项可以是模型参数向量的范数[2]。带有正则化项的式子一般具有如下形式:

$$\min_w L(w) = 1/2 \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda J(f) \tag{2.4}$$

对式(2.4)同样可以使用矩阵计算的方式来进行求解，结果为

$$W = (X^T X + \lambda I)^{-1} X^T Y \tag{2.5}$$

3 实验结果

真实函数为sin函数,$x \in [0,1]$。下图例中蓝色圆点代表具有一定噪声的真实数据，蓝色曲线代表sin函数对应的曲线，绿色曲线代表多项式拟合的结果。

3.1 10 samples, fit degree = 3 or 9

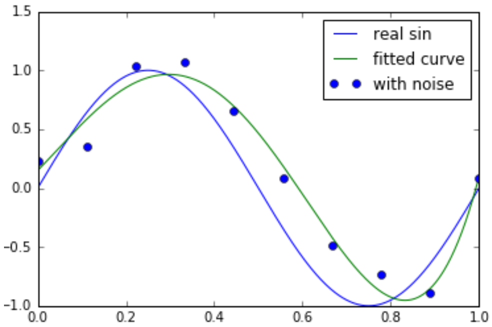

图 1: 10 samples,$M = 3$

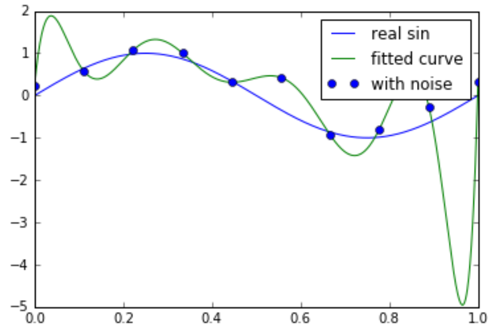

图 2: 10 samples,$M = 9$

3.2 15 or 100 samples, fit degree = 9

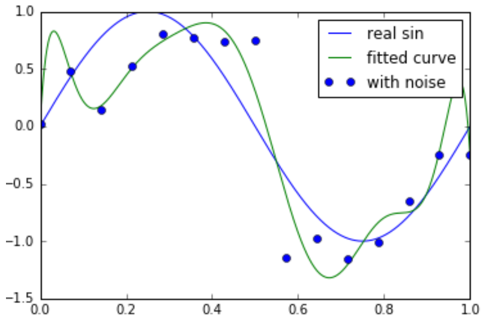

图 3: 15 samples,$M = 9$

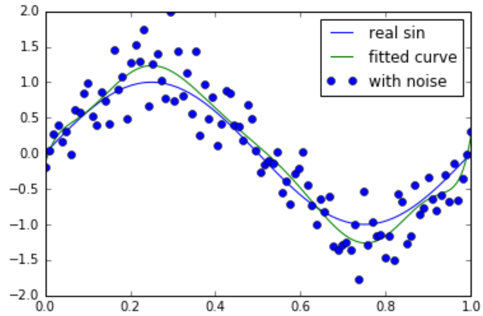

图 4: 100 samples,$M = 9$

3.3 10 samples, fit degree=9, with regularization term

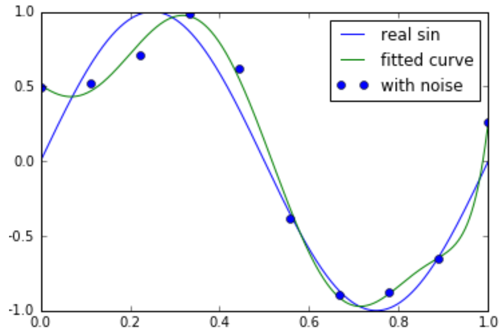

图 5: 100 samples,$M = 9$,with regularization

4 小结与讨论

对第3部分中的实验结果进行一定的观察与分析，可以得到以下结论。

图1和图2 对应的数据样本个数均为10 个，在样本相同的情况下，讨论不同模型复杂度对拟合效果的影响。其中，图2 的复杂程度比图1高很多，明显图1 拟合的效果要比图2 好很多，对未知数据具有更好的预测能力。但是图2 对已知数据具有很好的预测能力，但对未知数据预测能力很差。图3和图4 则是在相同模型复杂程度的基础上，讨论不同样本数量对拟合效果的影响。明显，样本数为100的拟合效果更好。图5 与图2的样本数、模型复杂度都相同，唯一不同的是增加了一个正则化项，可以明显看出图5的拟合效果比图2要好很多。

由以上实验可以看出模型选择的重要性。如果一味追求提高对已有的训练数据的预测能力，所选模型的复杂度则往往会比真实模型更高，例如图2所示，这也就是常说的过拟合现象，即在学习时选择的模型所包含的参数过多，以至于出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象。所以，模型其实不是越复杂越好，相反，我们应该在保证能够达到期望的前提下，使模型的复杂度越低越好。

从实验结果中，也可以清楚的看到正则化项对较高复杂度的模型的惩罚作用，从而帮助我们选择出经验风险和模型复杂度同时较小的模型。

参考文献

[1] Andrew Ng. Cs229 lecture notes. CS229 Lecture notes, 1(1):1–3, 2000.

[2] 李航. 统计学习方法. 清华大学出版社, 2012.