# 机器学习学习笔记

李宏毅教授机器学习(视频地址：<https://www.bilibili.com/video/BV1JE411g7XF?p=3&vd_source=0a27b8c0284e7cd3b72bc0c89478f2d7>)

第一课：Regression：Case Study

Regression(回归)，是让我们找到一个函数，使得给定一个input，它可以尽可能地通过这个函数转化成我们想要的output。  
例如我们想要预测一个宝可梦的CP值，也就是我们需要找到一个函数，使得给定一个宝可梦，它能尽可能准确地预测出该宝可梦进化后的CP值。  
首先我们要建立一个model，例如  
y = b + w1 \* xcp，  
y = b + w1 \* xcp + w2 \* (xcp)²,  
……  
我们通过training data构建了这个model，接着我们将testing data放入我们建立的model中去计算我们预测的cp值与实际的cp值之间的偏差，以此用来代表我们的model的好坏程度。  
我们发现，在一定程度上我们的model表达式越复杂，它的预测效果就越好，但是当表达式过于复杂的时候，我们的预测效果反而会变得很糟糕，这种现象我们称之为**Overfitting**，所以我们的model并不是越复杂就越好的，构建一个合适的model就成了一个非常关键的问题。  
而model的好坏往往通过它的error程度来反映，error的来源有两个：  
1.bias  
2.variance  
我们可以bias看作样本均值与总体均值的偏差，而variance看作样本的方差，这两方面就可以体现出我们model的好坏情况了。  
当样本总量N不断增大时，样本均值m就会越来越接近总体均值μ，所以样本均值的期望E（m）=μ。  
当样本总量不断增大时，样本方差s²也会越来越接近总体方差σ²，但是样本方差的期望并不等于σ²，E（s²）=(N-1)/N \*σ²。  
一般来说当我们的model比较简单的时候，我们的variance也会比较小；而当model比较复杂的时候，我们的variance也会比较大。  
我们的model比较坏有两种现象：

**1.underfitting**

表现为bias比较大，这个时候就要求我们重新设计我们的model。

**2.overfitting**

表现为variance较大。对于overfitting我们有两种方法来修正：  
一、是增加data量，也就是增大N，这个做法不会影响到我们原有的bias，如果无法拥有这么多data量的话，我们可以自己制造假的data，这个假的data是根据我们自己的理解合理创造出来的，比如只有向右行驶的火车的图片，我们可以将图片镜像翻转，就得到向左行驶的火车的图片了。  
二、regularization，通过增加一些参数，使得所得的曲线更加平滑一些，它不需要我们寻找新的data，但这个做法却可能会影响我们原有的bias，因为我们需要更平滑的曲线，所以我们就有可能会舍弃掉原来那些没那么平滑的曲线，而这个舍弃的操作就会影响到我们原有的bias。