1、模型评估

之前学到了很多模型，线性回归、logistic 回归、神经网络，但有时候预测效果不是那么的理想，最直观的表现就是代价值较高，因此本文就如何选择算法以及算法中参数值的设置作简要讨论。

2、训练集与测试集

一直以来，所建立的模型，通常是将所有数据直接用来做训练集，现在，将从这随机分布的数据中分出 70%作为训练集，另 30%作为测试集，模型的建立根据最小化那70%数据的代价值作为依据。然后用令 30%测试集，计算代价值以评估模型。但不幸的是，这不是一个好的方法，我么以模型次数选择为例。

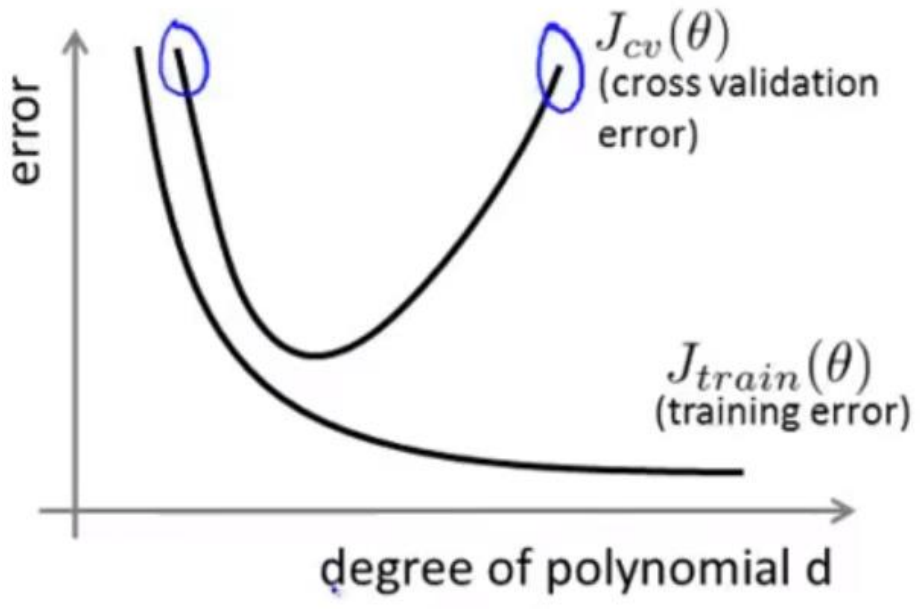
1. $h_\theta(x) = \theta_0 + \theta_1 x$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

$\vdots$

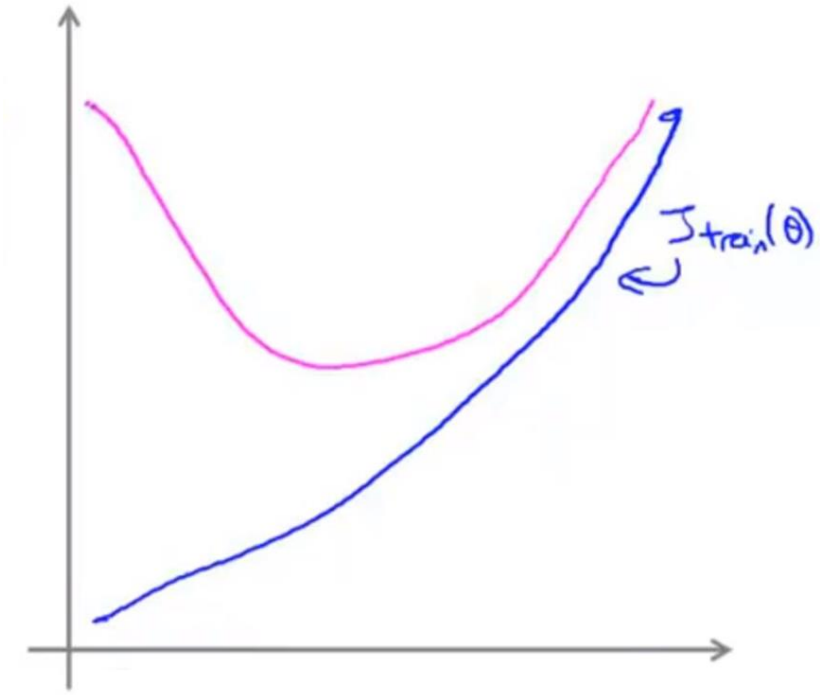
10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$

一开始，我们并不知道该选择次数项为多少的模型来拟合该数据，依据之前的方法，首先用训练集拟合数据，并最小化代价函数 $J(\theta)_1$，然后使用测试集计算新的代价值$J(\theta)_2$作评估，最终取$J(\theta)_2$最小的那组作为模型来预测。但这是不公平的，利用测试集最小化代价值本质上仍然建立在原有训练集上，无法正真的评估出它的泛化能力。

所以，接下来又引入了三组结构，分为训练集、验证集和测试集，训练集和测试集分别计算 $J(\theta)_1$和 $J(\theta)_2$，最后用测试集的$J(\theta)_3$作为模型泛化能力的评估标准。



上图很好的说明了模型选择中次数的影响，选择低次数，会出现欠拟合的情况，这时不论训练集还是验证集代价值都很高，我们称之为高偏差现象；而当次数较高时，训练集代价值较低，而验证集代价值很高，即出现了过拟合问题，我们称之为高方差现象。



同样，也可以得到训练误差和验证误差关于λ的变化情况，左边的λ值最小，对应高方差，右边最大，对应高偏差。