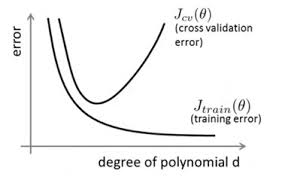
**关于过拟合**

在训练数据不够多时，或者overtraining时，常常会导致overfitting（过拟合）。其直观的表现如下图所示，随着训练过程的进行，模型复杂度增加，在training data上的error渐渐减小，但是在验证集上的error却反而渐渐增大——因为训练出来的网络过拟合了训练集，对训练集外的数据却不work。



在机器学习算法中，我们常常将原始数据集分为三部分：

1. training data

2. validation data

3. testing data。

**validation data是什么？**

它其实就是用来避免过拟合的，在训练过程中，我们通常用它来确定一些超参数（比如根据validation data上的accuracy来确定early stopping的epoch大小、根据validation data确定learning rate等等）。

**那为啥不直接在testing data上做这些呢？**

因为如果在testing data做这些，那么随着训练的进行，我们的网络实际上就是在一点一点地overfitting我们的testing data，导致最后得到的testing accuracy没有任何参考意义。因此，training data的作用是计算梯度更新权重，validation data如上所述，testing data则给出一个accuracy以判断网络的好坏。

**避免过拟合的方法**有很多：

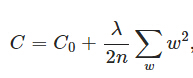
1. early stopping
2. 数据集扩增（Data augmentation）
3. 正则化（Regularization）包括L1、L2（L2 regularization也叫weight decay）
4. dropout。

**Early stopping**

所谓early stopping，即在每一个epoch结束时（一个epoch即对所有训练数据的一轮遍历）计算 validation data的accuracy，当accuracy不再提高时，就停止训练。这是很自然的做法，因为accuracy不再提高了，训练下去也没用。另外，这样做还能防止overfitting

**L2 regularization（weight decay）**

L2正则化就是在代价函数后面再加上一个正则化项



**L1 regularization**

**dropout**