# 初识机器学习

## 监督学习

监督学习（Supervised Learning）,输入样本已被标记。有两类问题，回归(Regression)和分类（Classification）。

Regression的输出是连续的，Classification输出离散。

**典型案例：**

Regression根据已知楼盘房价推测新楼盘房价。

Classification根据肿瘤样本的良性/恶性，推测新肿瘤是良性还是恶性。

## 无监督学习

无监督学习（Unsupervised Learning），输入样本没有被标记，需要根据样本间的相似性对样本集分类。

**典型案例：**

用户细分

混合音频的剥离

# 线性回归

监督学习中的回归问题，分为单变量线性回归（Linear regression with one variable）和多变量线性回归（(Linear Regression with Multiple Variables）。

一般公式为：

其中，，为变量数量，也叫特征数量。

矩阵公式：

## 代价函数

1. **函数原型**

文本

描述已自动生成

1. **python实现**

|  |
| --- |
| #代价函数  #x, y, theta为矩阵变量  def costFunctionJ(x, y, theta) :  #计算(x\*theta - y) ^ 2  inner = np.power(((x.dot(theta.T)) - y), 2)  #计算  累加和/2m  return np.sum(inner) / (2 \* len(x)) |

1. **原理**

假定训练实例原本为97x1的矩阵，为了计算矩阵相乘，前面插入一列1，为97x2矩阵。

矩阵为每个实例的输出，也是97x1矩阵。

m为实例数量，97。

矩阵为2x1矩阵，即，其中 = 1，每次运算有两个值。

回到代价函数公式，其中：

，

换成矩阵形式，即矩阵和矩阵相乘，，所得结果是97x1矩阵，包含了m个训练实例的每一个计算结果。

矩阵运算包含了每一个训练实例的计算结果与实际结构的差值。

python代码np.power((x.dot(theta.T), 2)，对差值矩阵的每个元素求平方。numpy使用dot进行矩阵乘法，使用multiply进行矩阵点乘。不要使用\*，容易乱。

最后的np.sum(inner) / (2 \* len(x))，也就是公式中功能。

## 批量梯度下降

1. **函数原型**

文本

描述已自动生成

1. **python实现**

|  |
| --- |
| def gradientDescent(x, y, theta, alpha, iters):      #theta矩阵清零      temp = np.matrix(np.zeros(theta.shape))      #revel()将多维数组降为一维      #返回theta元素个数      parameters = int(theta.ravel().shape[1])      #比如iters=5,则cost为长度为5的数组，值为0      cost = np.zeros(iters)        for i in range(iters):          error = (x.dot(theta.T)) - y          print(error)            for j in range(parameters):              term = np.multiply(error, x[:,j])              temp[0,j] = theta[0,j] - ((alpha / len(x)) \* np.sum(term))            theta = temp          cost[i] = costFunctionJ(x, y, theta)        return theta, cost |

1. **原理**

对于，梯度下降公式：

其中：

公式展开：

其中：

n为的特征数，本例中n=2。

求导后：

其中，代表矩阵中第行、第列，也就是第个训练实例的第个特征。

对于：

设矩阵R为计算结果，实际上是矩阵R和矩阵X的第一列相乘矩阵的累加。

回到python：

代码error = (x.dot(theta.T)) – y即为。

代码term = np.multiply(error, x[:,j])，即为，其中为矩阵的第列。

代码np.sum(term)即为。

代码temp[0,j] = theta[0,j] - ((alpha / len(x)) \* np.sum(term))即为对的迭代过程，其中在这里的取值为0，1。

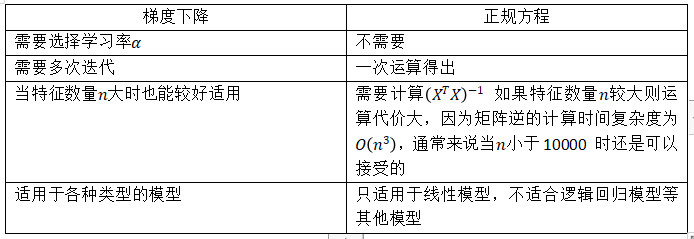
## 正规方程

正规方程（Normal Equation）是通过求解下面的方程来找出使得代价函数最小的参数的： 。

解得

TODO:推导过程涉及矩阵求导，目前不会，以后再推吧。

梯度下降与正规方程的比较：



总结，只要特征变量的数目并不大，标准方程是一个很好的计算参数的替代方法。具体地说，只要特征变量数量小于一万，我通常使用标准方程法，而不使用梯度下降法。

# 逻辑回归

Logistic Regression，监督学习中的分类问题，主要处理二分类中的线性可分问题。

对于线性可分，期望的假设函数处于[0, 1]范围，以便于判定：

当时，预测 = 1。

当时，预测 = 0。

因此结合了线性方程和Logistics/Sigmoid方程。

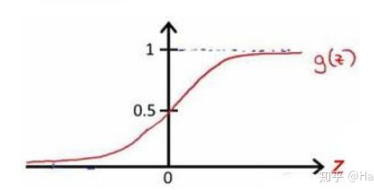
即：

的意义，给定输入变量,，计算输出变量等于1的可能性，也写作：

## 判定边界

Decision Boundary，就是判定的边界。

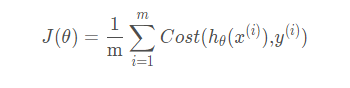
的图像为：



也就是确定后，的位置。

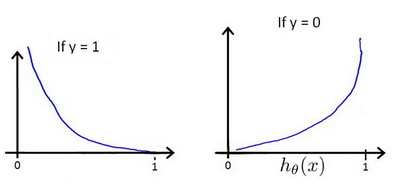
## 代价函数

逻辑回归的原始代价函数：



其中：

图像为：



这个函数的意义在于：

的情况，等于1时误差为0，越远离1，误差越大。

的情况，等于0时误差为0，越远离0，误差越大。

这样的函数特性符合实际预期，Cost二段式函数合并一下：

所以逻辑回归的最终代价函数为：

python代码实现也是根据这一条公式而来：

|  |
| --- |
| # Logistic Regression代价函数  # 返回值[jVal, gradient] - 为了适应fmin\_tcn调用  # 注意这个theta参数类型为数组  # 原因是fmin\_tnc函数自动调用代价函数时，  # 会对自动将theta转换成array\_like，因此参数theta的类型为array\_like  def costFunctionJ(theta, x, y):      #输入转换为矩阵形式      theta = np.matrix(theta)      x = np.matrix(x)      y = np.matrix(y)      #H(x)      hx= sigmoid(x.dot(theta.T))      first = np.multiply(-y, np.log(hx))      second = np.multiply((1 - y), np.log(1 - hx))      #矩阵形式，下面两个都可以，只是grad的形状不一样      #grad = 1.0/(len(X)) \* (hx - y).T.dot(x)      grad = 1.0/(len(X)) \* x.T.dot(hx - y)      return np.sum(first - second) / (len(X)), grad |

后续中， 这个代价函数会作为scipy库中的truncated newton (TNC) 算法的参数进行最优计算，为了适配TNC函数，代价函数的返回值参数表是[jVal, gradient]。

## 梯度下降

梯度下降公式：

代入后化简，得到的结果和线性方程的一致：

其中：

实际工程中基本不会使用这个算法，而是使用现成的算法库。

## TNC算法库

直接将代价函数、初始化值（[0, 0, 0]）、原始数据集传入fmin\_tnc函数进行调用。

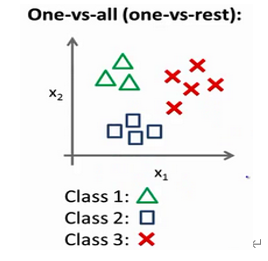
|  |
| --- |
| import scipy.optimize as opt  result = opt.fmin\_tnc(func = costFunctionJ, \  x0 = theta, args=(x, y), messages = 0) |

返回值：

|  |
| --- |
| 返回值[array, nfevel, rc]  @array : 最优theta数组  @nfevel: 函数评估的数量  @rc : 返回码  返回码rc定义：  -1 : Infeasible (lower bound > upper bound)  0 : Local minimum reached (|pg| ~= 0)  1 : Converged (|f\_n-f\_(n-1)| ~= 0)  2 : Converged (|x\_n-x\_(n-1)| ~= 0)  3 : Max. number of function evaluations reached  4 : Linear search failed  5 : All lower bounds are equal to the upper bounds  6 : Unable to progress  7 : User requested end of minimization |

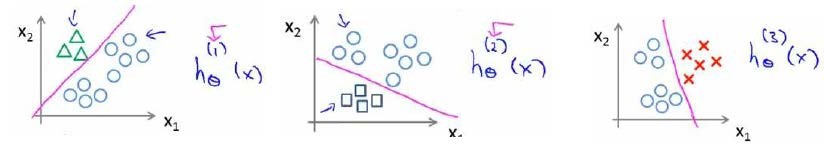
## 一对多问题

前面的例子只有两个取值的二元分类问题，如果是有多个取值，其图形是这样的：



其实可以转换思想，将“一对多”转换成“一对余”问题，上面图形的分类模型可以分为3个二元分类模型：

，其中



对每一个分类，将标记的一类记为1，其余标记都是0。分别计算、、。

预测时，给定，选择一个让 最大的，即，其中。

# 正则化

正则化用以解决过拟合问题。

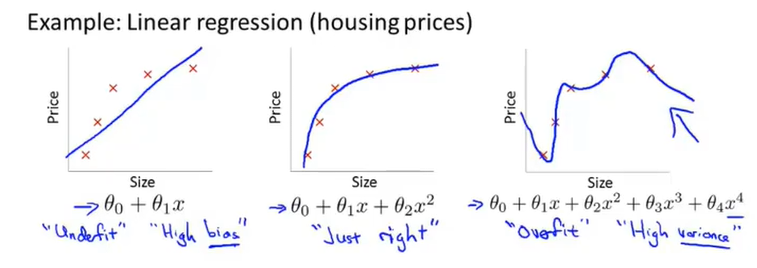
这里说的正则化指是L2正则化（还有L2正则）。

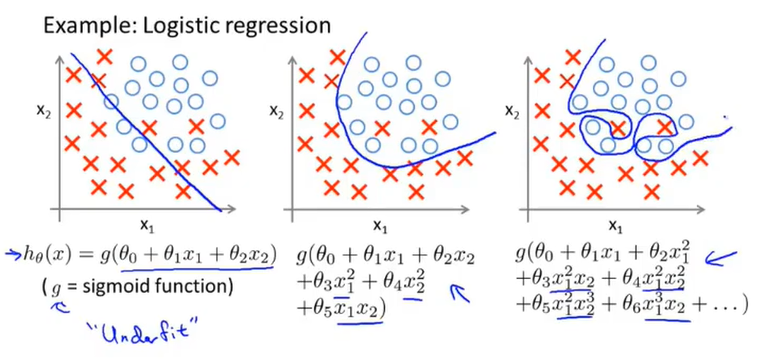
## 过拟合

为了满足现有的训练集，盲目加特征次数，搞的太复杂，结果泛化效果可能会不好。

（泛化，panalize，一个假设模型能够应用到新样本的能力）。

下面两组图，左一是欠拟合/高偏差，右一是过拟合/高方差。





解决办法：

1. 减少变量的个数：舍弃一些变量，保留更为重要的变量。但是，如果每个特征变量都对预测产生影响。当舍弃一部分变量时，也就舍弃了一些信息。所以，希望保留所有的变量。
2. 正则化：保留所有的变量，但是会减小特征变量的数量级（参数数值的大小θ(j)）。

## 线性回归正则化

线性回归正则化的代价函数就是在原来基础上加上正规项：

正规项是从1开始，即不进行处理。

### 梯度下降

线性回归正则化的梯度下降函数推导：

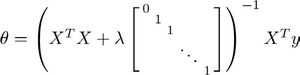
其中不处理：

其他的：

调整一下：

其中是一个略小于1的数，跟普通线性方程中的一样。

### 正规方程



其中矩阵的尺寸为。

这个公式还解决了普通线性回归正规方程的不可逆问题。

## 逻辑回归正则化

### 代价函数

代价函数中同样是加上正规项：

梯度下降不作处理：

之后的：

调整一下：

其中。

python代码，为了使用高级优化算法，同样要在代价函数里返回gradient：

|  |
| --- |
| #sigmoid函数  #参数可以是矩阵，或者实数  def sigmoid(z) :  return 1 / (1 + np.exp(-z))  def gradient(theta, x, y):  theta = np.matrix(theta)  x = np.matrix(x)  y = np.matrix(y)  hx= sigmoid(x.dot(theta.T))  return 1.0 / (len(x)) \* (x.T.dot(hx - y))  #这里要保证参数是矩阵，最好还是在函数内部进行一下处理  def gradientReg(theta, x, y, lam=1):  theta = np.matrix(theta)  x = np.matrix(x)  y = np.matrix(y)  #实数\*矩阵  reg=(lam/len(x))\*theta.T  reg[0] = 0 # 第一项没有惩罚因子  return gradient(theta, x, y) + reg  #代价函数  #相比于非正则化的逻辑回归代价函数，就是多了reg项  #其中reg项中的j是从1开始, 而不是0  def costReg(theta, X, y, learningRate):  theta = np.matrix(theta)  X = np.matrix(X)  y = np.matrix(y)  first = np.multiply(-y, np.log(sigmoid(X \* theta.T)))  second = np.multiply((1 - y), np.log(1 - sigmoid(X \* theta.T)))  #正规项中j从1开始  reg = (learningRate / (2 \* len(X))) \* np.sum(np.power(theta[:,1:theta.shape[1]], 2))  return np.sum(first - second) / len(X) + reg, gradientReg(theta, X, y) |

### 特征映射

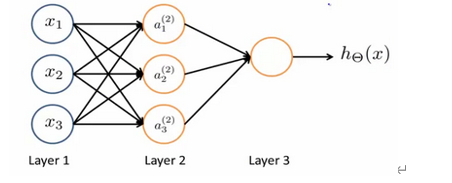
原案例中，只给了两个特征项，不好演示正则化，因此需要扩展特征项，使之包含一系列特征高次项。

|  |
| --- |
| degree = 5  x1 = data2['Test 1']  x2 = data2['Test 2']  data2.insert(3, 'Ones', 1)  for i in range(1, degree):  for j in range(0, i):  data2['F' + str(i) + str(j)] = np.multiply(np.power(x1, i-j) ,np.power(x2, j)) |

# 神经网络

## 原理

神经网络似乎是一个多层的逻辑回归，每一层增加一个1作为偏差单元（bias unit）,每层之间使用一组作为权重矩阵，



每一层的运算都是几组sigmoid函数运算，前后层的单元数量（也叫神经元数量）可以不同，考虑到前一层的最前面要补‘1’，偏差单元的维度应该是：后一层单元数量x（前一层单元数量+1）。

这个有点FPGA流水线的意思，每一层进行一个特定运算，向后传递，这个算法也确实被称为前向传播算法（FORWARD PROPAGATION）。

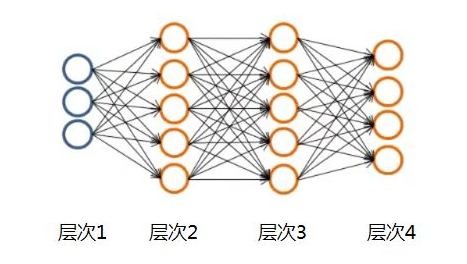
神经网络就是多层的逻辑回归，但是相比逻辑回归，比更先进。

## 代价函数

逻辑回归的代价函数：



神经网络就是对逻辑回归增加了维度的扩展：



用来标记层数，标记层的单元数，表示最终的分类数量（）。来标记某一层的行，来标记的列。

那么神经网络的代价函数为：

第一部分：

其实就是把各个分类的计算结果进行累加。也就是说，我们需要求得的参数，应该对每一个分类计算代价函数，并使得加总之后的结果最小。

第二部分：

表示各层的累加。

表示对某一层各行各列的累加。

表示行，第层的行数为下一层单元数量，即。从1开始因为不参与正则化，网上这么多的分析过程，你们就不舍得在这里加个括号吗。混淆视听！

表示列，从1开始是为了去除偏差项。

## 反向传播算法

### 理解

无论是前向传播，还是反向传播，一个重要计算项是偏导数，每一次迭代，都是计算所有训练实例的，然后使用高级优化算法计算最优。

但是对于神经网络，直接计算梯度会导致计算过于庞大，不适用于工程。而应当使用反向传播算法作为神经网络的训练算法。

所以引入了一个神经元误差的概念，定义第层第个神经元的误差为：

那么可以由神经元误差推导代价函数偏导数，已知：

那么：

### 神经元误差公式

以下图的4层神经算法为例：

图示

描述已自动生成

各层的误差值：

小tips（对Sigmoid函数求导）:

### 推导过程

推导过程就是围绕神经网络误差定义计算而来，对中间误差的计算会用到求导的链式法则：

表格, 信件

中度可信度描述已自动生成

**的推导过程**：

**的推导过程**（使用了链式法则，需要耐心）：

上面的转换运用了链式法则，其中有：

继续化简：

将结果向量化，考虑矩阵相乘的维度后，即得到的结果：

### 实施

1. 参数随机初始化

不能像以前初始化为0了，应该是随机值。随机初始化一个，后续使用时再根据神经元数量拆解出各层的。

1. 构造反向传播函数

构造反向传播函数，这个函数类似正向传播的代价函数，要同时返回正则化的代价函数，和正则化梯度。

1. 正向传播计算

在反向传播函数中，先进行一次正向传播，由此可以获得各层的和最终输出。根据最终输出可以算出代价函数值，作为第一个返回值。

1. 反向传播计算神经元误差

在反向传播函数中，对于每一个训练实例，通过最后一层的神经元误差，计算前面各层神经元误差。

1. 计算代价函数偏导

在反向传播函数中，通过计算各层偏导，结果要除以m，然后加上正则项。

1. 返回梯度

在反向传播函数中，将各层的梯度一维化，作为第二个返回值。

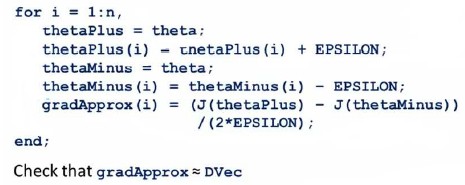
1. 使用高级优化算法

将构造好的反向传播函数，传入高级优化算法函数中，获取最优。

1. 梯度检测

(计算耗时高。使用高级优化算法后，这一步似乎不需要，高级优化函数的返回值会包含收敛信息)

1. 使用最优再计算一次反向传播，得到偏导数Dvec。
2. 计算近似梯度gradApprox，假设最优长度为，一般取epsilon=0.001:



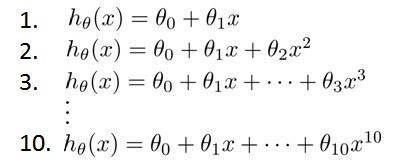
1. 通过范数比较两个矩阵的相似度，结果应小于1e-9。

diff = norm(numgrad-grad)/norm(numgrad) + norm(grad);

# 评估

## 模型选择

假设我们要在10个不同次数的二项式模型之间进行选择：



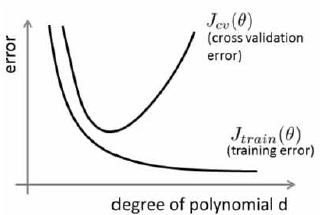
那么需要要样本分为3份，分别作为Train Set（训练集，60%）、Cross Validation Set（交叉验证集20%）和Test Set（测试集20%）。然后：

1. 分别使用这10个模型对Train Set进行训练，得到10组训练参数。
2. 分别使用这10组训练参数，对Cross Validation Set的数据计算代价函数的值。选择结果最小的训练参数对应的模型，作为最佳模型。
3. 使用最佳模型参数对Test Set计算代价函数值，得出推广误差。

## 偏差和方差

高偏差(Bias)代表欠拟合，高方差(Variance)代表过拟合。

随着多项式次数d的增加，训练集的代价函数会越来越小，而交叉验证集的代价函数会形成如图曲线。



对于训练集，当 较小时，模型拟合程度更低，误差较大；随着 的增长，拟合程度提高，误差减小。

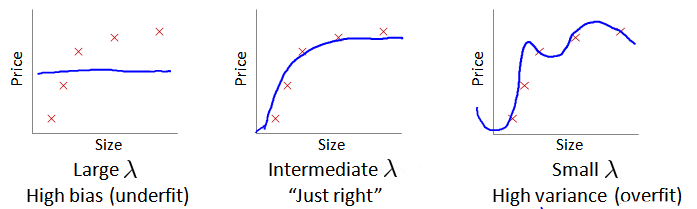
对于交叉验证集，当 较小时，模型拟合程度低，误差较大；但是随着 的增长，误差呈现先减小后增大的趋势，转折点是我们的模型开始过拟合训练数据集的时候。

图的左半部分，很大，，为高偏差。

图的半部分，很小，，为高方差。

## 正则化偏差/方差

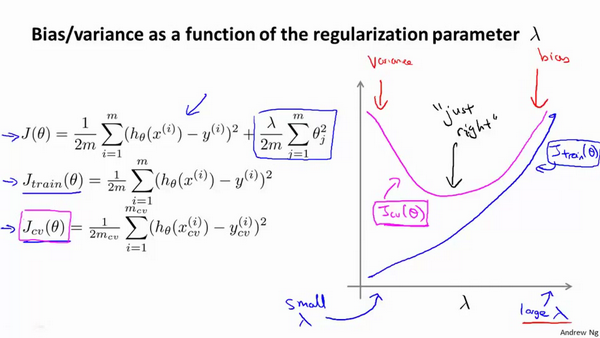
在正则化在选择时同样要考虑偏差方差。



做法：

从0~10选择12个，依次呈2倍关系，然后分别使用训练集训练出12个正则化模型，用这12个模型分别对交叉验证集进行验证，选出交叉验证误差最小的模型。

如果把训练集和交叉验证集的代价函数与绘制在一张图上：



当 较小时，训练集误差较小（过拟合）而交叉验证集误差较大

随着 的增加，训练集误差不断增加（欠拟合），而交叉验证集误差则是先减小后增加。

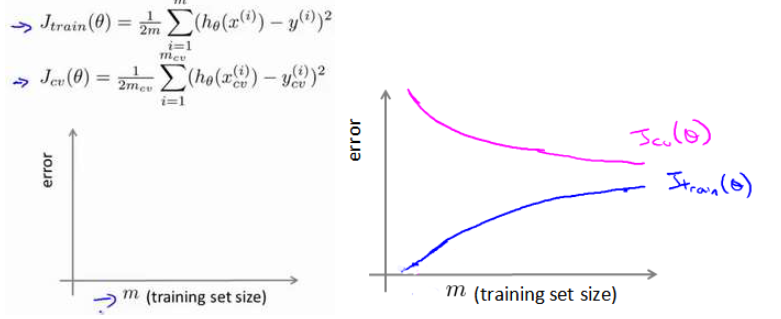
## 学习曲线

学习曲线就是训练集误差和交叉验证集误差作为训练实例数量的函数绘制的图表，可用来判断某个学习算法是否处于偏差、方差问题。

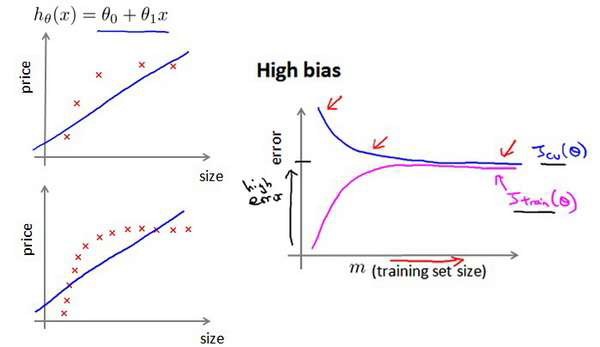
如果我们有100 行数据，我们从1 行数据开始，逐渐学习更多行的数据。思想是：

当训练较少行数据的时候，训练的模型将能够非常完美地适应较少的训练数据，但是训练出来的模型却不能很好地适应交叉验证集数据或测试集数据。

一般情况下，随着训练样本的增加，训练集的平均误差平方和会增加，而相应的验证集上的平均误差平方和会减少。曲线如下图所示：

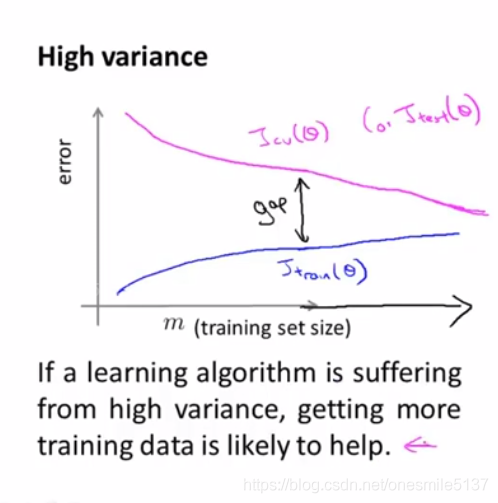


高偏差/欠拟合情况下，比如应该用2次项，却只用了1次项，随着m值增加，和趋于一致，且值较大，如图所示。



这时候就不要盲目增加训练集数量了，而是模型本身有问题。

高方差/过拟合情况下，随着样本数目的增加，交叉验证集的偏差一直与训练集的偏差有很大的差距，训练的偏差在不断地上升；



这种情况下，增大数据集在一定程度上可以优化算法，

这也是为什么一些较复杂的算法，在普通量级的训练样本上表现一般般，一旦到了大数据领域就会有惊艳表现。

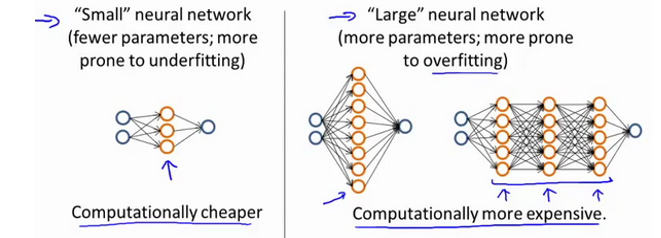
## 结论

一般情况下，下面这些方法，就是在遇到高方差或高偏差时，我们所需要做的事情。

* 得到更多的训练集：适用于**高方差**。
* 减小特征的数目：适用于**高方差**。
* 尝试增加特征：适用于**高偏差**。
* 增加多项式（）：适用于**高偏差**。
* 增大 λ：适用于**高方差**。
* 减小 λ：适用于**高偏差**。

神经网络中，使用简单的神经网络，比如较少的隐藏层数量，较少的神经元数量，容易导致高偏差/欠拟合，计算量小。

使用复杂的神经网络，容易高方差/过拟合，计算量大。



一般做法：选用大的神经网络，然后用正则化处理来解决高方差问题。

对于神经网络中的隐藏层的层数的选择，通常从一层开始逐渐增加层数，为了更好地作选择，可以把数据分为训练集、交叉验证集和测试集，针对不同隐藏层层数的神经网络训练神经网络， 然后选择交叉验证集代价最小的神经网络。

# 系统设计

## 设计顺序

* 简单、快速实现算法，并用交叉验证集数据测试
* 绘制学习曲线，决定是增加更多数据，或者添加更多特征，还是其他选择
* 进行误差分析：人工检查交叉验证集中我们算法中产生预测误差的实例，看看这些实例是否有某种共有特征或规律。再进行算法中的特征修正。

这里的误差分析，通常要分析准确率（Precision）和查全率（Recall）这两个因素。

## 查准率&&查全率

偏斜类问题：训练集一种实例占比很大，另外一种实例所占比例很小。

针对偏斜问题，有4中预测结果：

* 正确肯定（True Positive,TP）：预测为真，实际为真
* 正确否定（True Negative,TN）：预测为假，实际为假
* 错误肯定（False Positive,FP）：预测为真，实际为假
* 错误否定（False Negative,FN）：预测为假，实际为真

那么：

查准率（Precision）：TP/(TP + FP)，即，所有预测为真的样例中，实际为真的比例。

查全率（recall）：TP/(TP +FN)，即，所有实际为真的样例中，被正确检测出来的比例。

## 均衡

查准率和查全率都是越大越好，但是也需要权衡。

以肿瘤预测为例，一般的作法是当 0.5时，预测值，其中。

如果为了更高的查准率Precision，即预测结果的准确度高（如果查准率低，就是误诊，告诉没病的人，你有肿瘤。。。。。。），

这时候可以适当增加的阈值，比如当 0.7时，才把预测值。这样也会降低查全率(Recall)。

如果为了更高的查全率Recall，类比为将更多的疑似病例作为确诊病例，可以接受进一步的检查、诊断（如果查全率低，会使某些肿瘤病人没有被检测出来，耽误治疗）。

这时候可以适当降低的阈值，比如当 0.3时，就把预测值。这样也会降低查准率Precision。

因此引入了F1值的概念：

选择不同阈值，得到不同的算法，算出各自的准确率P值和查全率R值，然后算F1值：

F1值越大，算法越优。

## 算法提升和训练样本增加

如果满足以下两个条件，可以增加训练样本：

* 一个人类专家看到特征值x，可以预测出y，即特征向量x完备性。
* 有条件获取足够多的样本，并训练一个复杂、多参数的算法。

如果没有以上条件，已经选择其他高性能算法。