C1-2

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

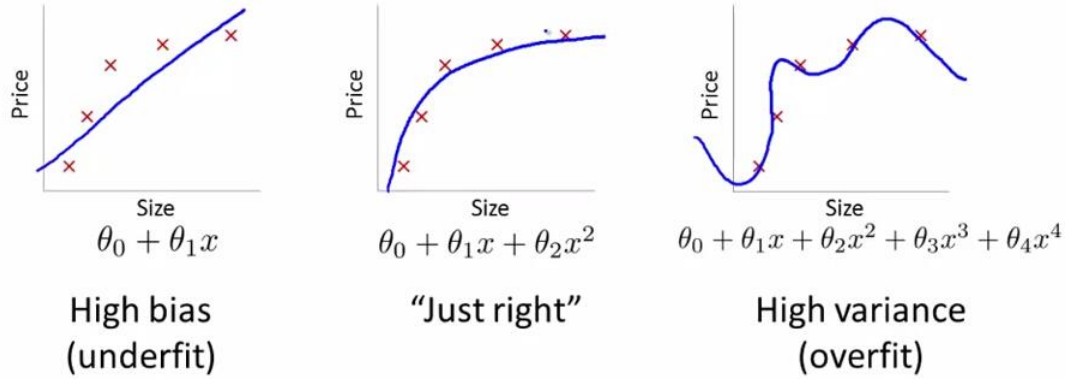
2016.8.7

第三集：欠拟合与过拟合的概念

- 欠拟合与过拟合(Under-fitting and Over-fitting)
- 局部加权回归(Locally Weighted Regression)
- 线性回归的概率学解释(Probabilistic Interpretation)
- Logistic回归(Logistic Regression)
- 感知机(Perception)

第三集：欠拟合与过拟合的概念

- 欠拟合与过拟合
  - 回到上次说到的房屋大小和房费的例子，假设我们忽略房数，只考虑用房屋大小去预测房费的话，如何对这个问题进行建模
    - 上次见到一种最简单的建模方法：线性回归，用一条直线尽可能描述样本点的特征
    - 然而对于m组数据样本来说，用m-1阶曲线可以进行完美的拟合
      - 比如说给定两个样本点，用一个1阶曲线（直线，线性函数）可以完美拟合，给定三个样本点，用一个2阶曲线（抛物线，二次函数）可以完美拟合
    - 但是模型不是越准确就越好的，太过于准确的模型会遇到过拟合的问题，参数过多模型训练复杂，而且容易受到噪声点的影响；但是完全不准确的模型会遇到欠拟合的问题，无法合理地表示出样本的内在特征

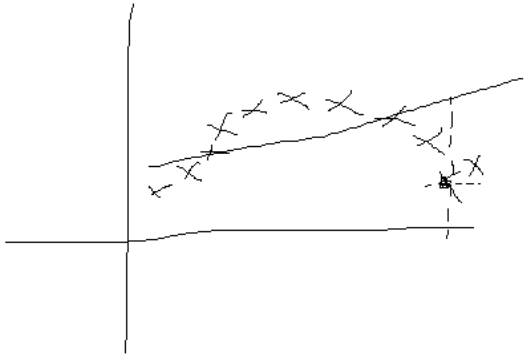


第三集：欠拟合与过拟合的概念

- 欠拟合与过拟合
  - 欠拟合和过拟合的问题关键在于到底应该选择多少个参数对问题进行建模求解
    - 比如说现在有m个样本，那我们是应该选择1阶曲线进行拟合，还是2阶曲线进行拟合，还是m-1阶曲线进行拟合
      - 1阶曲线对应2个参数($h(x)=\theta 1x+\theta 0$)， 2阶曲线对应3个参数($h(x)=\theta 2x^2+\theta 1x+\theta 0$)
    - 对此主要有两种解决办法
      - 采用特征选择算法(Feature Selection Algorithm)，自动化地选择所需的参数个数和拟合曲线的阶数
      - 采用非参数学习算法(Non-parametric Learning Algorithm)，比如下面说到的局部加权回归算法（这一方法并不能完全解决欠拟合与过拟合的问题，但是在一定程度上缓解了这个问题）

第三集：欠拟合与过拟合的概念

- 局部加权回归
  - 上一节中的线性回归算法是参数学习算法(Parametric Learning Algorithm)的一个例子
    - 参数学习算法：在参数学习算法中有固定数目的参数来进行数据的拟合，在进行数据拟合前先要根据全部的样本点计算出参数然后才能完成预测
      - 比如说上一节讲到的线性回归中的θ
    - 非参数学习算法：在非参数学习算法中参数的个数会随着样本数目的增长而增长
      - 因此不需要一开始就指定参数的个数
  - 局部加权回归算法则是一种非参数学习算法

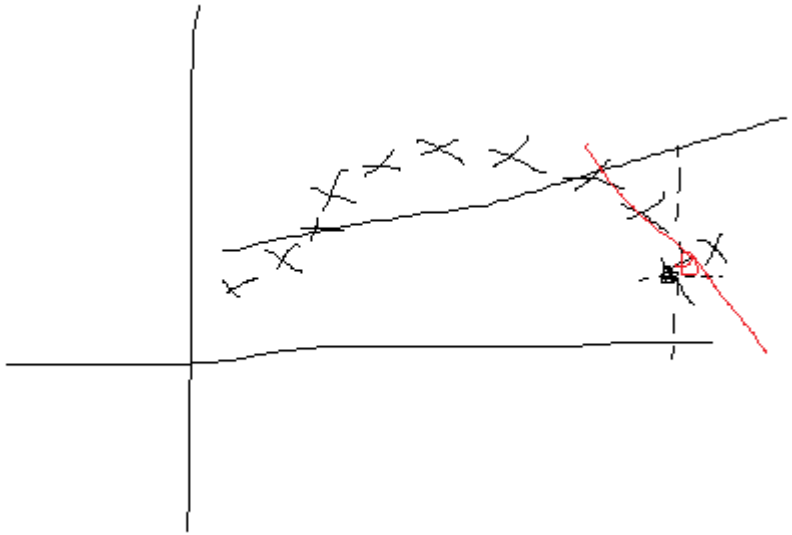


假设左边这图，x表示样本点，如果我们直接使用下面的线性回归很可能得到图中的直线。但很明显，这条直线并不能很好地对数据进行拟合。现在的问题是怎样对这组数据进行建模？

1. Fit θ to minimize $\sum_i (y^{(i)} - \theta^T x^{(i)})^2$.
2. Output $\theta^T x$.

第三集：欠拟合与过拟合的概念

- 局部加权回归
  - 在局部加权回归中我们的做法略有不同
    - 但我们要预测给定输入x的对应输出y时，会先检查训练数据集合中那些x附近（附近的程度需要人为控制）的样本点，然后只对这一区域内的样本点进行线性回归。求出这一区域样本点的θ值后，再预测出给定的输入x对应的输出y



比如左图中我们要预测虚线给出的点x对应的预测值y，我们只需要考虑x附近的样本点，然后对于这些点进行线性回归从而得到图示的红色直线，再进行预测明显比原来直接对于全部样本使用线性回归预测更为准确。

第三集：欠拟合与过拟合的概念

- 局部加权回归
  - 实际上我们将上面这个例子看得更为广义，可以看成是在原来的线性回归中加上了一个权值
    - 上面这个例子相当于在输入变量x附近的样本点的权值为1，而其他的点的权值为0
  - 由此我们可以归纳出局部加权回归算法的一般形式

1. Fit θ to minimize $\sum_i w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2$.
2. Output $\theta^T x$.

这里面的w为权值，每个样本的权值很可能不同，原本的线性回归相当于是w全为1的特例。这方法的一个问题是对于样本集合很大的情况学习速度可能很慢。

  - 其中w非负，对于w取值的定义方法有多种，一般要使得附近的样本的权值尽量大，原理的样本的权值尽量小。下面展示一种经典的w值选择方法。

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

在这个例子里面，我们发现当样本点在输入变量x的附近时，分子的值很小，最后结果近于1. 如果很远则近于0.τ称为带宽(bandwidth)系数，控制了远离输入变量x的样本权重下降的程度，需要根据实际情况进行确定。

第三集：欠拟合与过拟合的概念

- 线性回归的概率学解释
  - 在线性回归里面我们采用了最小二乘衡量方法设计目标函数，为什么我们的目标是一个这样的函数，而不是用1-范数进行距离衡量（绝对值距离）或是用其它的距离衡量办法？
  - 首先我们将原问题表示成下面的形式

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)},$$

其中ε为线性拟合时所产生的误差，这一误差项我们可以看作是数据集内在的未捕捉到的特征（比如说房屋的其它特征），或者看成随机的噪声（比如当天售卖的天气，人的心情）

  - 现在我们假设原来的这个误差项满足某一个概率分布，根据生活实际和中心极限定理可以假设这个概率分布为正态分布，从而可以得到误差项的概率
    - 中心极限定理：多个离散独立随机变量的和趋于正态分布

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

第三集：欠拟合与过拟合的概念

- 线性回归的概率学解释
  - 接下来我们将下面这两个式子组合起来即可得到y关于x的概率密度函数

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)},$$

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right) \longrightarrow p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

其中θ为参数，为确定值，并不是随机变量

  - 假设εi相互之间是独立的，也就是说每一组样本(x,y)之间毫无关联。那么现在我们要问一个问题，y关于x的概率密度函数究竟是什么样子的才能正确拟合出y和x之间的联系？
    - 这个问题反过来是简单的，如果给你这样的一个概率密度函数p和输入变量x0，问这个输入为x0的情况下经过某种拟合对应可以得到y0的概率是多少——很简单直接将x0代入概率密度函数p即可得到输入为x0的情况下经过某种拟合对应可以得到y0的概率
    - 但现在出现了一个问题，我们现在得到了一系列数据样本，这数据样本不足够多，就要问你总体的数据的概率密度函数p应该是什么样子的——这就涉及到一个问题：如何用邮箱的不足的数据去估计总体的分布

第三集：欠拟合与过拟合的概念

- 线性回归的概率学解释
  - 由于我们现在并不能拿到足够多的数据，因此我们并不能准确回答出总体的数据的概率密度函数p应该是什么样子的，但是我们可以用这些数据提供的信息来推测出概率密度函数p更有可能是什么样子的
  - 现在的问题是怎么衡量概率密度函数p的可能性？
    - 我们可以从样本的角度进行考虑，由于这组样本是真实的，因此我们有理由相信这组样本是我们经过某一概率密度函数p最有可能抽取出来的结果。那我们应该怎样衡量抽取的过程？
      - 由于在之前说到每一组样本(x,y)之间是毫无关联，因此从这一概率密度函数p中抽取得到的结果就相当于每次抽取的结果的乘积，即p(y1|x1;θ)*p(y2|x2;θ)*...*p(ym|xm;θ)
      - 由于这个结果反映的是在不同参数θ的取值下，取得当前这个样本集的概率，因此称其为参数θ相对于样本集X的似然函数(Likelihood Function)，记为L(θ)

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

第三集：欠拟合与过拟合的概念

- 线性回归的概率学解释
  - 由于我们要表示这组样本是我们经过某一概率密度函数p最有可能抽取出来的结果，因此我们要做的就是最大化似然函数，我们称这个过程为极大似然(Maximum Likelihood)
    - 对于线性回归的例子来说我们需要极大化下面这个函数

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

    - 由于求极大值一般采取求导的办法，而在这里乘法的求导比较复杂。因此我们可以对上面这个函数加上一个对数以后再求导（由于log是单调增的，所以log L(θ)的最大值和L(θ)的最大值对应的θ相同，因此不会改变原结果）。好处在于log可以使得乘法变加法使得求导变得简单。 log L(θ)被称为对数似然(Log Likelihood)

第三集：欠拟合与过拟合的概念

- 线性回归的概率学解释
  - 对于线性回归的例子对数似然如下

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2
\end{aligned}
$$

由于这里σ的平方一定非负，因此θ在这里的取值并不会受到σ的影响

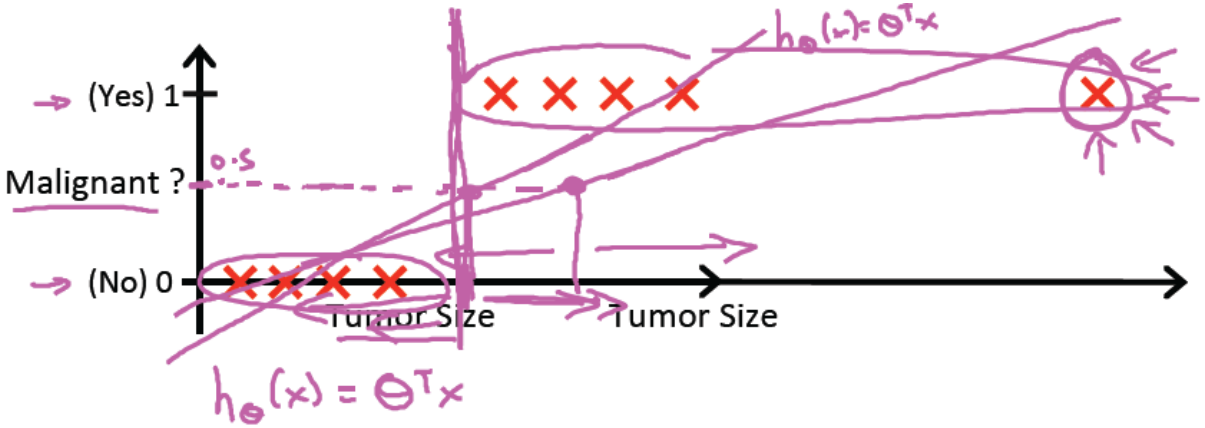
  - 要极大化这个函数要使得后面这项尽可能小，因此我们需要极小化下面这个函数

$$
\frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2
$$

这也就是之前线性回归的那个目标函数

第三集：欠拟合与过拟合的概念

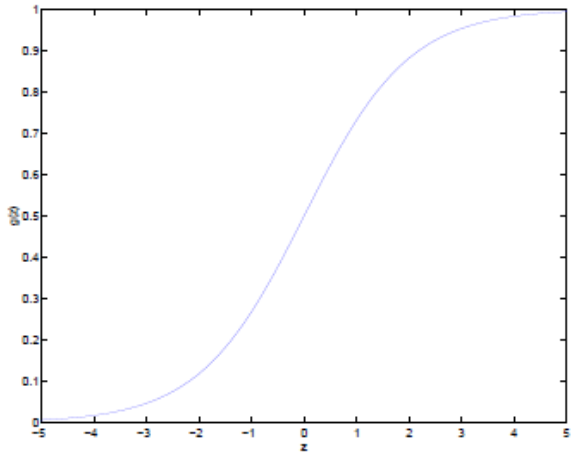
- Logistic回归
  - 之前讲到的线性回归办法是一种回归算法，因为它要预测的值是一个连续的值。那么对于离散的值应该怎样进行处理？
    - 假设Y的取值范围在{0,1}之间，应该如何处理
      - 一种方法，仍然可以用线性回归拟合出一条直线。然后设定一个阈值，大于阈值的判断为1，否则判断为0（但不能保证效果）



比如左边的这例子，如果我给出一个很远的样本点。这样本点其实在印证之前设置的阈值和拟合直线的正确性，但是却会导致了拟合直线的偏离，这时候对应的阈值很可能发生了改变。但不幸的是这是我们一开始难以预料的

第三集：欠拟合与过拟合的概念

- Logistic回归
  - 线性拟合的办法有问题的原因在于如何正确选择一个这样的阈值，由此催生出了一个设想，有没有办法能够不选阈值而直接可以得到非0即1的效果
    - 由此我们需要一个函数g能够在小于某个值x0的输入x的情况下g(x)小于等于0，而在大于某个值x0的输入x的情况下g(x)大于等于1。
      - 一种设想是分段函数，这个后面会提到，称为感知机
      - 还有一种采用sigmoid函数的方法，该函数连续且具有上述属性。与此同时也比较简单，学习起来速度快，其定义如下（具体原理以后会讲到）



$$g(z) = \frac{1}{1+e^{-z}}$$

该函数也被称为logistic函数。因此该分类方法也被称为Logistic回归。其导数也非常好算（见右）

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

第三集：欠拟合与过拟合的概念

- Logistic回归
  - 那么针对我们的例子我们可以令目标函数h(x)为

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

  - 那么下面的问题依然是我们怎样解出这样的θ，在线性回归中我们给出了最小二乘拟合的表达形式得到目标函数J(θ)。这一表达形式并不是通用的。看过前面的内容发现这一表达式是经过极大似然推出来的，只是上一节是直接给出的没有经过证明而已。因此我们同样需要求它的极大似然，在此之前先看一下如何得到y关于x的概率密度函数
    - 在线性拟合中我们假设的是y=h(x)+ε，其中h(x)为$\theta^T x$，ε符合正态分布。而在这里由于是离散的值，因此可以不理会这个ε，直接令y=h(x)，因此有

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

这个式子很好理解，因为y=h(x)，所以当输入为x结果为1的概率就像原来h(x)所描述的那样，小于某个阈值时几乎为0；当输入为x结果为0的概率就像原来1- h(x)所描述的那样，大于某个阈值时几乎为0

第三集：欠拟合与过拟合的概念

- Logistic回归
  - 我们将上一页的两个式子写在一起可以得到

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$

$$\longrightarrow \quad p(y \mid x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

因为y非0即1

  - 由此可得似然方程和对数似然方程

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)})) \end{aligned}$$

第三集：欠拟合与过拟合的概念

- Logistic回归
  - 我们可以直接将对数似然函数当成目标函数，即$\ell(\theta)$。通过证明可以得知$\ell(\theta)$为凹函数

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)}))\end{aligned}$$

这里我们采用另外一种证明方法来证明$\ell(\theta)$的凸凹性。如果$\ell(\theta)$二阶可导，其为凸函数的充要条件为二阶导数大于等于零，其为凹函数的充要条件为二阶导数小于等于零（其实都等价于凸函数，无非前面是否有正负号）。

因为

$$\begin{aligned}g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)).\end{aligned}$$

所以

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial\theta_j}g(\theta^T x) \\ &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)(1-g(\theta^T x)\frac{\partial}{\partial\theta_j}\theta^T x \\ &= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right)x_j \\ &= \left(y-h_\theta(x)\right)x_j\end{aligned}$$

而二阶导数为

$$\frac{\partial[\ell(\theta)]^2}{\partial^2\theta} = \frac{\partial[\ell(\theta)]^2}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\left(y-h(x)\right)x\right] = -xg\left(\theta\right)\left[1-g\left(\theta\right)\right]x \le 0 \ \ \left(g\left(\theta\right)\ge 0\right)$$

因此原函数为凹函数

第三集：欠拟合与过拟合的概念

- Logistic回归
  - 由于$\ell(\theta)$为凹函数，对于一个凹函数来说具有一个与凸函数相反的一个性质，在凹函数中任何极大值也是最大值
  - 因此我们需要最大化对数似然$\ell(\theta)$相当于求其极大值，求极大值的方法很简单，直接梯度上升(gradient ascent)即可
    - 在之前线性回归里面讲的是求极小值，因此采用的是梯度下降法。梯度上升法和梯度下降法的原理完全相同，只不过反过来做而已，用来求极大值。由此可得下面的算法（这里所采用的是stochastic gradient ascent方法，其与stochastic gradient descent原理上相同）
      $$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$
      - 其中梯度的计算为：

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)} \right) g(\theta^T x)(1-g(\theta^T x)\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x) \right) x_j \\
&= \left(y - h_\theta(x) \right) x_j
\end{aligned}$$

本θ的更新公式在形式上与上一节线性回归中所用的θ的更新公式几乎完全相同，但实际上两者毫无关联。原因在于这个式子里面的h(x)的计算方法是不一样的，变成了Logistic函数

第三集：欠拟合与过拟合的概念

- 感知机
  - 在Logistic回归中我们要使得输出在有限集合（在Logistic回归中为{0,1}之间）取值，但问题是Logistic函数并不是分段的，而是一个连续函数。那么我们怎样强制令结果在{0,1}之间取值？
    - 因此还存在一种解决办法，采用分段函数的办法进行学习。我们可以将此函数定义为：

$$g(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

    - 该函数也可以通过类似Logistic回归的办法进行学习，这称为感知机学习算法(perceptron learning algorithm)

If we then let $h_\theta(x) = g(\theta^T x)$ as before but using this modified definition of g, and if we use the update rule

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}.$$

then we have the **perceptron learning algorithm**.

可以简单看作这个是Logistic回归加上了某个阈值进行转换的结果（以后还会继续讨论）