Logistic Regression

基础概念

Logistic Regression 虽然被称为回归，但其实际上是分类模型，并常用于二分类。**Logistic Regression** 因其简单、可并行化、可解释强深受工业界喜爱。

Logistic 回归的本质是：假设数据服从这个分布，然后使用极大似然估计做参数的估计。

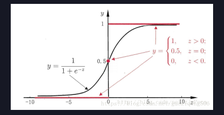
阶跃函数Sigmoid Function

直接依靠拟合曲线的函数值是不能得到类标号的，还需要一种理想的"阶跃函数"，将函数值按照正负性分别映射为**0**，**1**类标号。

Sigmoid函数：

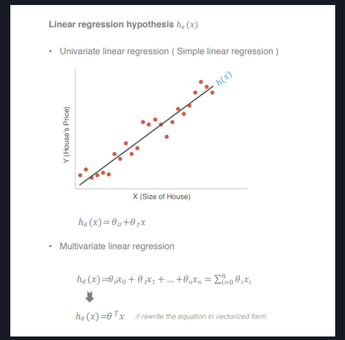
$$\phi(z) = \frac{1}{1+e^{-z}}$$

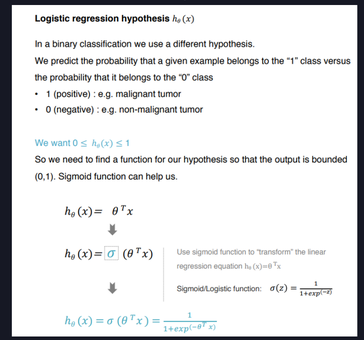
Sigmoid函数的图像如下图所示，当**z>0**时，**Sigmoid**函数大于**0.5**；当**z<0**时，**Sigmoid**函数小于**0.5**。所以，我们可以将拟合曲线的函数值带入**Sigmoid**函数，观察**ϕ(z)**与**0.5**的大小确定其类标号。



Sigmoid函数还有一个好处，那就是因为其取值在**0，1**之间。所以可以看做是测试元组属于类**1**的后验概率，即**p(y=1|X)**。其实这一点从图像也可以看出来：**z**的值越大，表明元组的空间位置距离分类面越远，他就越可能属于类**1**，所以图中**z**越大，函数值也就越接近**1**；同理，**z**越小，表明元组越不可能属于类**1**.

Linear regression vs Logistic regression



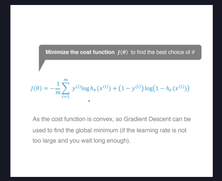


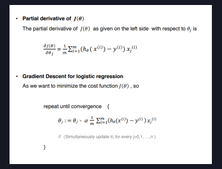
代价函数cost function

$$p(y|X;W) = \phi(z)^y (1-\phi(z))^{1-y}$$

在参数W下，元组类标号为y的后验概率

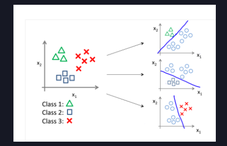
上述公式表达的含义是在参数W下，元组类标号为y的后验概率。假设现在已经得到了一个抽样样本，那么联合概率的大小就可以反映模型的代价：联合概率越大，说明模型的学习结果与真实情况越接近；联合概率越小，说明模型的学习结果与真实情况越背离。而对于这个联合概率，我们可以通过计算参数的最大似然估计的那一套方法来确定使得联合概率最大的参数W，此时的W就是我们要选的最佳参数，它使得联合概率最大（即代价函数最小）。





多分类问题

用one vs all



逻辑回归的优缺点

LR优点：

1.直接对分类的可能性建模，无需事先假设数据分布，避免了假设分布不准确带来的问题
2.不仅预测出类别，还可得到近似概率预测
3.对率函数是任意阶可导凸函数，有很好得数学性质，很多数值优化算法可直接用于求取最优解
4.容易使用和解释，计算代价低
5.LR对时间和内存需求上相当高效
6.可应用于分布式数据，并且还有在线算法实现，用较小资源处理较大数据
7.对数据中小噪声鲁棒性很好，并且不会受到轻微多重共线性影响
8.因为结果是概率，可用作排序模型

LR缺点：

1.容易欠拟合，分类精度不高
2.特征有缺失或特征空间很大时效果不好