**逻辑回归方法简介**

# 1. 什么是回归分析法

回归，最初是遗传学中的一个名词，是由生物学家兼统计学家高尔顿首先提出来的。他在研究人类的身高时，发现高个子父母的孩子回归于人口的平均身高，而矮个子父母的孩子则从另一个方向回归于人口的平均身高。

预测是我们日常生活中离不开的活动，因而在机器学习中借助于历史数据构建高精确度的预测模型是一项基本任务。如果预测结果是离散的分类标号，我们称之为分类，如果是连续值则称之为预测，因而分类与预测是预测技术的两种类型。

回归分析是一种预测性的建模技术，研究自变量和因变量之间数量变化关系的一种分析方法。它主要是通过建立因变量Y与影响它的自变量X之间的回归模型，衡量自变量X对因变量Y的影响能力，进而可以预测因变量Y的发展趋势。例如，研究个人饮酒量与肝功能之间的关系，最好的研究方法就是回归。

按照自变量的多少，回归分析分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

# 2. 线性回归简介

## 2.1 线性回归模型

预测训练数据集： ，即*D*中有*m*个实例（或称之为点）。，即每个实例中包含*n*个自变量。

线性回归假设数据特征间满足线性关系，即根据训练数据集*D*，构建一个自变量的线性组合模型，并用此模型进行因变量的预测。其函数形式为：

(1)

令*x*0=1，即，，该线性函数可表示为：

(2)

因而确定线性回归模型，就是确定权重系数向量*W*。

## 2.2 线性回归模型训练

模型根据什么进行训练呢？定义一个合理的目标函数（或称之为损失函数），使得损失函数取值最小。线性回归模型所用的损失函数一般是平方误差，即：

(3)

基于最小二乘法，按下面的公式求解即可得W的值。

(4)

# 3. 逻辑回归简介

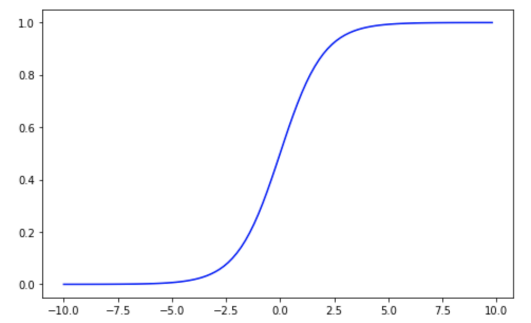
## 3.1 逻辑回归模型

Logistic回归为概率型非线性回归模型，是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法。通常的问题是，研究某些因素条件下某个结果是否发生，比如医学中根据病人的一些症状来判断它是否患有某种病。

Logistic回归是一种广义线性回归，是在所建立的线性模型的基础上在进行Logistic函数的处理。Logistic函数是最典型的Sigmoid函数，其公式如下：

　　　　　　　 　 (5)

　　其图像如下：



由于Logistic函数的定义域为(-∞,+∞)，值域为(0, 1)，因此最基本的LR分类器适合对两类目标进行分类。当其输出大于0.5时，可以认为该实例属于甲类；小于0.5时，认为其属于乙类。

分类训练数据集： ，即*D*中有*m*个实例。若，，则可建立逻辑回归模型为：

(6)

　上式逻辑回归分析模型的关键为权重向量*W*的取值，只有公式中的*W*已知，才能对一个未分类的数据进行二分类，那么该如何求得*W*呢？

## 3.2 逻辑回归模型的训练

　通常使用极大似然估计来建立目标函数，求得*W*的值。使用样本（即已知分类的数据），进行一系列的估算，得到*W*。这个过程在概率论中叫做参数估计。

　　(1) 首先我们令：

(2) 将上述两式整合：

(3) 求其似然函数：

(4) 对其似然函数求对数：

　　(5) 当似然函数为最大值时，得到的*W*即可认为是模型的参数。

### 3.2.1 基于梯度下降法的逻辑回归模型训练

求似然函数的最大值，可以使用一种方法，梯度上升，但我们可以对似然函数稍作处理，使之变为梯度下降，然后使用梯度下降的思想来求解此问题，变换的表达式如下：

在梯度下降法中，要使用当前的*W*（优化问题中，参数通常表示为*θ*）值更新得到新的*W*值，因而需要知道*W*更新的方向（即梯度）。对J(*W*)求偏导数可得到在当前点下的梯度。求梯度如下：

得到更新方向后便可使用下面的式子不断迭代更新得到最终结果。

### 3.2.2 基于牛顿法的逻辑回归模型训练

(1) 基于逻辑回归模型第4步中定义的似然函数，求得一个维度的梯度计算公式为：

(2) n+1维的梯度向量为：

(3) 根据*L*(*W*)计算Hessian矩阵*H*，其中*j*行*k*列的值为：

(4) (n+1)\*(n+1)的Hessian矩阵为：

(5)   更新参数*W*，*W* = *W* – *H*-1*g*