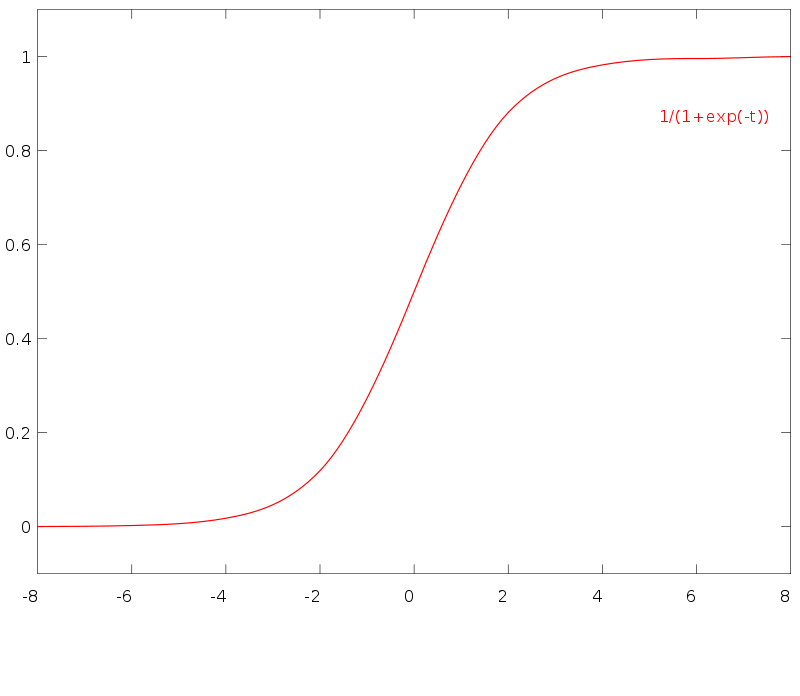
Logistic Regression

逻辑回归

Logistic Regression是一种Parametric Learning Algorithm。即给定固定数量的参数来拟合数据。

Logistic Regression的中文翻译是“逻辑回归”。名字中的“逻辑”，即logistic指的是logistic function，又名sigmoid function（“S”型函数）。

其表现形式为：

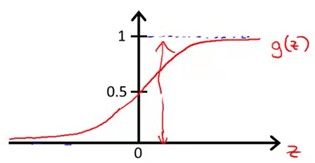


不要问我为啥取名叫“逻辑”，因为我也不知道。

线性回归是解决数据拟合问题的，逻辑回归是用来解决分类问题的。其实质也是一种数据拟合。

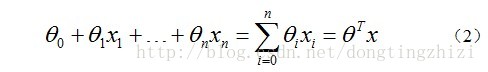
所谓分类问题，归根到底是一个二分类问题，这张图片是不是猫脸？这个肿瘤是不是恶性的？这封邮件是不是spam？Yes or No, 简单干脆。（多分类问题也是由二分类问题组合而生的。）我们希望能有一个模型，训练完成后能直接告诉我们分类的概率（yes 的概率），比如一张图片是猫脸的概率。（大于百分之五十，我们认为是yes，小于为no）

我们再看看这个logistic function

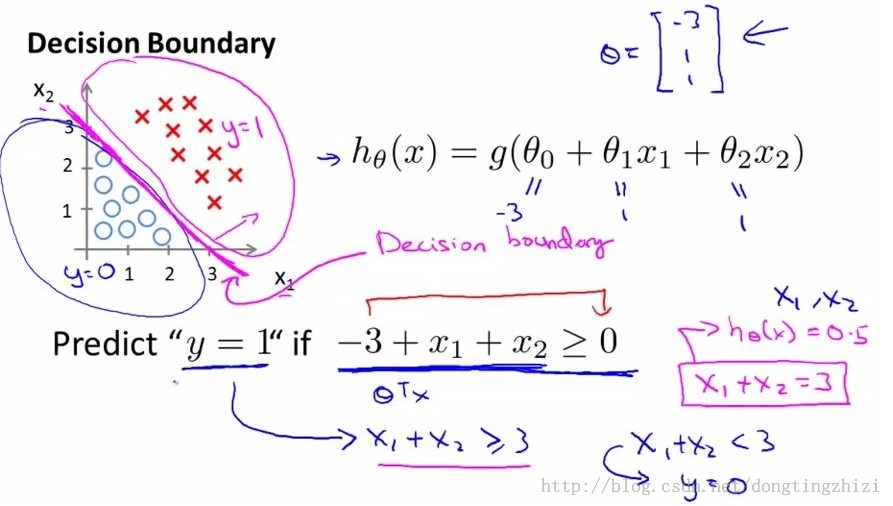


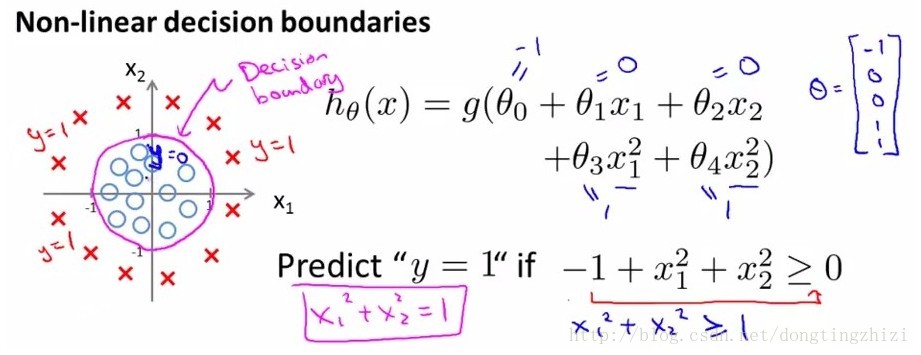
自变量 z取值到，因变量g(z) 取值0到1 。z取0时g(z)取0.5，完美的符合了我们的要求。

我们的目的就变成了，找到decision boundary，即boundary的一侧的数据为一类（yes），在另一侧为一类（no）。decision boundary我们这里用线性方程表示

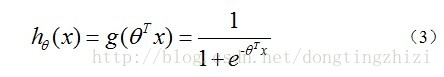


即为decision boundary。这里我们将0设为threshold

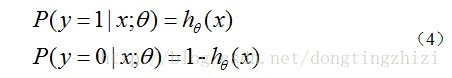


线性方程引入多项式解决非线性分类问题毫无压力。

所以我们将decision boundary 代入到logistic function 中，落在decision boundary上的点，结果为0.5，即yes or no概率各为0.5。距离decision boundary越远，就越趋近于1或0 。这正是我们所需要的预测模型（hypothesis）。



hθ(x)函数的值有特殊的含义，它表示结果取1的概率，因此对于输入x分类结果为类别1和类别0的概率分别为：



看起来很眼熟不是？很显然这是一个贝努力分布，二项式分布。

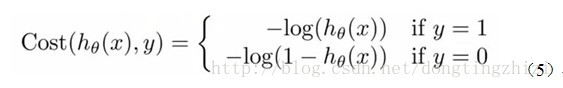
至于为啥用这个sigmoid function 代入boundary function， 就得到我们需要的符合二项式分布的预测模型了，我们之后会证明。

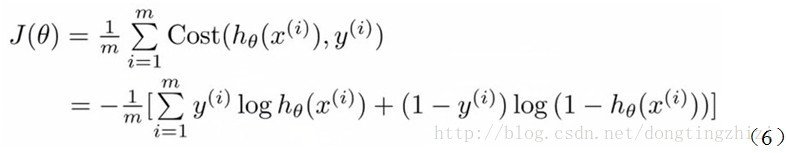
构造好hypothesis，就可以开始训练了。

机器学习的训练一般需要构造一个cost函数，也就是误差函数，告诉我们的模型当前参数是好还是坏。而训练的过程就是不断调整我们当前的模型即hypothesis的参数，使得cost函数越来越小。而一个模型的训练好坏也是根据cost函数的输出结果来评判的。

这个cost 函数的表示为：

构造预测函数为：





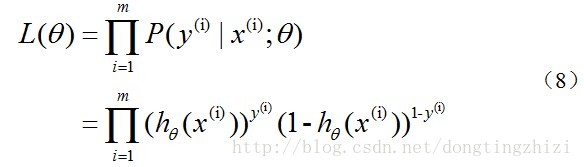
cost function 一般有两个参数，一个是 hypothesis，一个是实际结果（supervised learning的标注值）

logistic regression的cost function 在这里被设为一个区间函数，但可以用一条方程来表达。则如有m条训练数据，cost function 为m条cost function 的平均值。用J（theta）来表示。

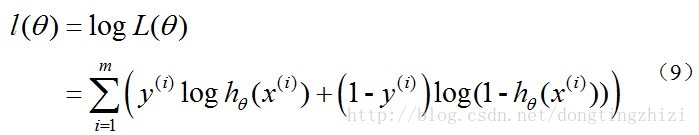
这里的cost和jfuntion 是由极大似然估计得出的（MLE）

由





取对数后：

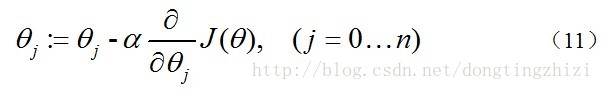




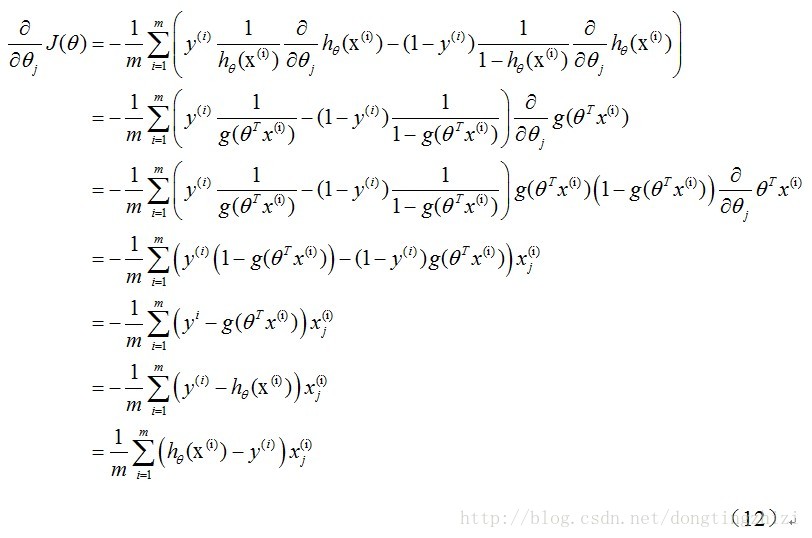
其中的系数－1/m可以自由选取， Andew用的是这个。。。。

有了cost function， 下一步要做的就是通过改变参数找到cost function的最小值，最传统的方法就是梯度下降法，当然还有牛顿法，L－B法

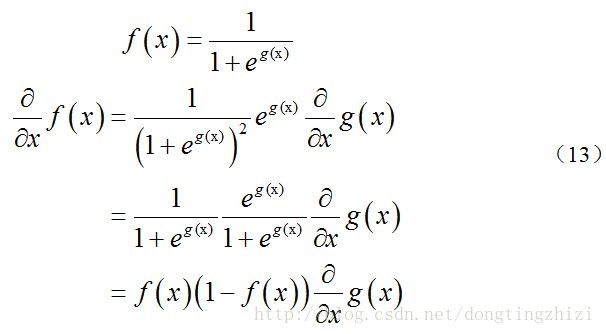
梯度下降法，就是沿着梯度减小的方向改变参数值，如果cost function是convex的（凸）的，我们肯定能找到它的全局极值。



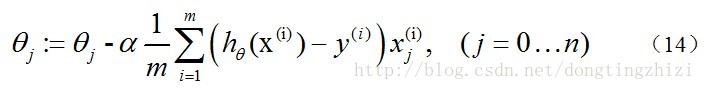
其中的偏导展开为：



上式中用到的公式：



所以参数更新方程为：



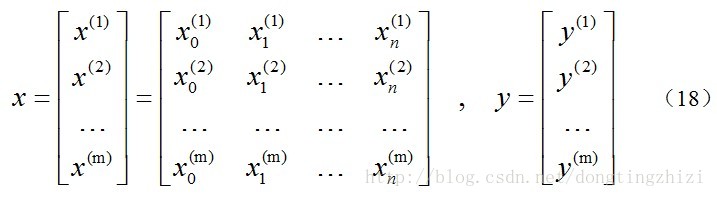
1/m和步长同为系数，省略之。

之后的过程就是不断的迭代，直到参数不再更新，即达到收敛（converge）。

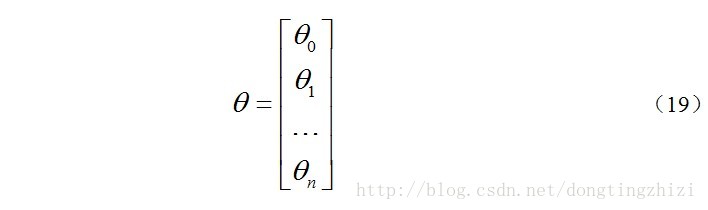
Vectorization

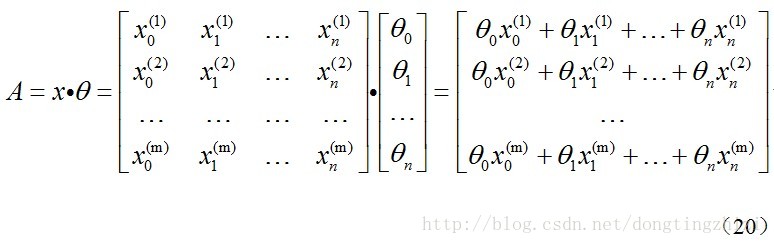
由以上内容来看，整个更新方程比较简单，训练的主要问题在于数据量大，参数多，为了计算方便，最好将计算过程矩阵化（vectorizaiton）

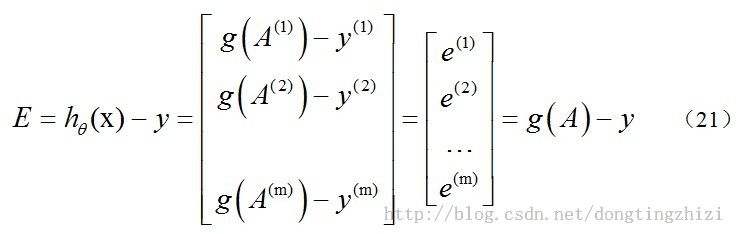
训练数据矩阵化：

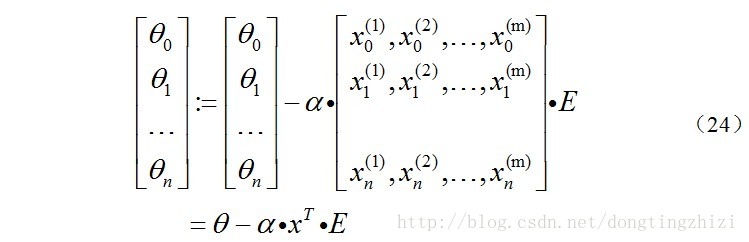


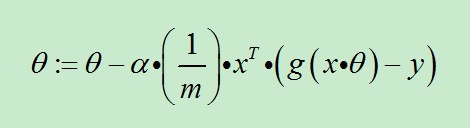
参数矩阵化：



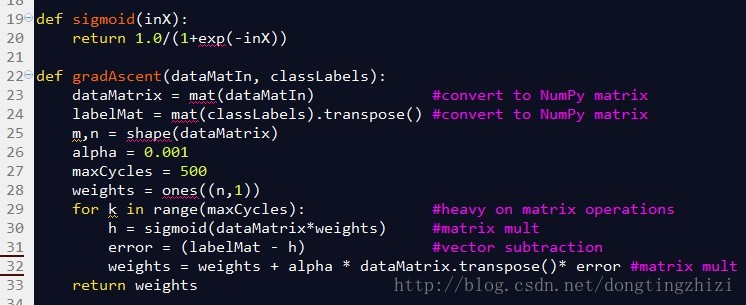








python实例：



经典线性模型自变量的线性预测就是因变量的估计值。 广义线性模型：自变量的线性预测的函数是因变量的估计值。常见的广义线性模型有：probit模型、poisson模型、对数线性模型等等。对数线性模型里有：logistic regression、Maxinum entropy。本篇是对逻辑回归的学习总结，以及广义线性模型导出逻辑回归的过程。下一篇将是对最大熵模型的学习总结。本篇介绍的大纲如下：

1、逻辑斯蒂分布，logit转换

2、在二分类问题中，为什么弃用传统的线性回归模型，改用逻辑斯蒂回归？

3、逻辑回归模型的求解过程？

4、实际应用逻辑回归时数据预处理的经验总结。但经验有限，如果有哪位网友这块经验丰富，忘指教，先谢过

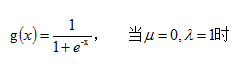
5、为什么我们在实际中，经典线性模型的优化目标函数是最小二乘，而逻辑回归则是似然函数

6、从最根本的广义线性模型角度，导出经典线性模型以及逻辑回归

**1、逻辑斯蒂分布，logit转换**

 一个连续随机变量X，如果它的分布函数形式如下，则X服从逻辑斯蒂分布，F（x）的值在0~1之间，它的的图形是一条S型曲线。





**2、在二分类问题中，为什么弃用传统的线性回归模型，改用逻辑斯蒂回归？**

      线性回归用于二分类时，首先想到下面这种形式，p是属于类别的概率：



      但是这时存在的问题是：

      1）等式两边的取值范围不同，右边是负无穷到正无穷，左边是[0,1]，这个分类模型的存在问题

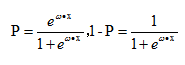
      2）实际中的很多问题，都是当x很小或很大时，对于因变量P的影响很小，当x达到中间某个阈值时，影响很大。即实际中很多问题，概率P与自变量并不是直线关系。

      所以，上面这分类模型需要修整，怎么修正呢？统计学家们找到的一种方法是通过logit变换对因变量加以变换，具体如下：





        从而，



        这里的P完全解决了上面的两个问题。

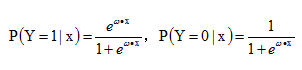
**3、逻辑回归模型的求解过程？**

      1）求解方式

        逻辑回归中，Y服从二项分布，误差服从二项分布，而非高斯分布，所以不能用最小二乘进行模型参数估计，可以用极大似然估计来进行参数估计。

      2）似然函数、目标函数

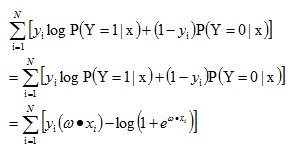
        严谨一点的公式如下：



        似然函数如下：



        对数似然函数，优化目标函数如下：



         整个逻辑回归问题就转化为求解目标函数，即对数似然函数的极大值的问题，即最优化问题，可采用梯度下降法、拟牛顿法等等。

**4、实际应用逻辑回归时数据预处理的经验总结，但经验有限，如果有哪位网友这块经验丰富，忘指教，先谢过**

      1）枚举型的特征直接进行binary

      2）数值型特征，可以：标准化、根据分布进行binary

      3）进行pairwise

**5、为什么我们在实际中，经典线性模型的优化目标函数是最小二乘，而逻辑回归则是似然函数**

      下面公式直接从Ng notes里面复制过来。

     1） **经典线性模型**的满足下面等式：



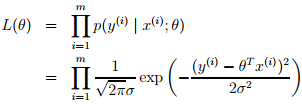
       这里有个假设，即最后这个误差扰动项独立同分布于均值为0的正态分布，即：



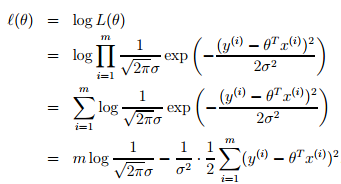
      从而：



      由于有上面的假设，从而就有下面的似然函数：



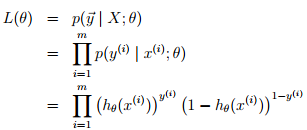
      从而这线性回归的问题就可转化为最大化下面的对数似然估计，由于下面公式前面的项是常数，所以这个问题等价于最小化下面等式中的最后一项，即least mean squares。

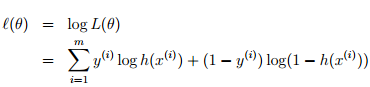


      2）**逻辑斯蒂回归**中，因变量y不再是连续的变量，而是二值的{0,1}，中间用到logit变换，将连续性的y值通过此变换映射到比较合理的0~1区间。在广义线性回归用于分类问题中，也有一个假设（对应于上面回归问题中误差项独立同分布于正态分布），其中h(x)是logistic function



      即，给定x和参数，y服从二项分布，上面回归问题中，给定x和参数，y服从正态分布。从而。





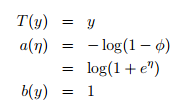
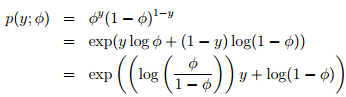
      问题不同（一个是分类、一个是回归）对应假设也就不同，决定了logistic regression问题最优化目标函数是上面这项，而非回归问题中的均方误差LMS。

**6、从最根本的广义线性模型角度，导出经典线性模型以及逻辑回归**

**1）指数家族**

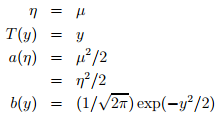
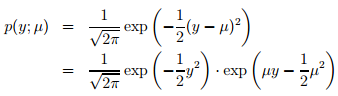


        当固定T时，这个分布属于指数家族中的哪种分布就由a和b两个函数决定。下面这种是伯努利分布，对应于逻辑回归问题

          注：从上面可知 ，从而，在后面用GLM导logistic regression的时候会用到这个sigmoid函数。

        下面这种是高斯分布，对应于经典线性回归问题

**2）GLM（广义线性模型）**

        指数家族的问题可以通过广义线性模型来解决。如何构建GLM呢？在给定x和参数后，y的条件概率p(y|x,θ) 需要满足下面三个假设：

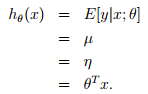
        assum1)      y | x; θ ∼ ExponentialFamily(η).

        assum2)      h(x) = E[y|x]. 即给定x，目标是预测T(y)的期望，通常问题中T(y)=y

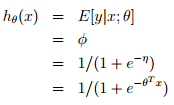
        assum3)       η = θTx，即η和x之间是线性的

**3）经典线性回归、逻辑回归**

       经典线性回归：预测值y是连续的，假设给定x和参数，y的概率分布服从高斯分布（对应构建GLM的第一条假设）。由上面高斯分布和指数家族分布的对应关系可知，η=µ，根据构建GLM的第2、3条假设可将model表示成：



        逻辑回归：以二分类为例，预测值y是二值的{1,0}，假设给定x和参数，y的概率分布服从伯努利分布（对应构建GLM的第一条假设）。由上面高斯分布和指数家族分布的对应关系可知，，根据构建GLM的第2、3条假设可model表示成：



        可以从GLM这种角度理解为什么logistic regression的公式是这个形式~

      参考资料：

ou﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽所以