Python实现Logistic回归算法(对应LR文件里的代码）

[**Python**](http://lib.csdn.net/base/11)实现算法的过程依然直接对J(θ)采取梯度上升法或者随机梯度上升法求解，LRTrain对象同时实现了采取梯度上升法或者随机梯度上升法求解过程。

LR分类器学习包中包含lr.py/object\_json.py/test.py三个模块。lr模块通过对象logisticRegres实现了LR分类器，支持gradAscent('Grad') and randomGradAscent('randomGrad')两种求解方法（二选一，classifierArray只存储一种分类求解结果，当然你也可以定义两个classifierArray同时支持两种求解方法）。

test模块中是利用LR分类器根据疝气病症预测病马死亡率的应用。该数据存在一个问题--数据由30%的丢失率，这里采用特殊值0替代，因为0不会影响LR分类器的权值更新。

训练数据中样本特征值的部分缺失是很棘手的问题，很多文献致力于解决该问题，因为数据直接丢掉太可惜，重新获取代价也昂贵。一些可选的数据丢失处理方法包括：

□使用可用特征的均值来填补缺失值；

□使用特殊值来±真补缺失值，如-1;

□忽略有缺失值的样本；

□使用相似样本的均值添补缺失值；

□使用另外的[**机器学习**](http://lib.csdn.net/base/2)算法预测缺失值。

LR分类器算法学习包下载地址是：

[machine learning Logistic regression](http://download.csdn.net/detail/suipingsp/8239259)

Logistic回归应用

Logistic回归的主要用途：

* 寻找危险因素：寻找某一疾病的危险因素等；
* 预测：根据模型，预测在不同的自变量情况下，发生某病或某种情况的概率有多大；
* 判别：实际上跟预测有些类似，也是根据模型，判断某人属于某病或属于某种情况的概率有多大，也就是看一下这个人有多大的可能性是属于某病。

Logistic回归主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率，等等。例如，想探讨胃癌发生的危险因素，可以选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群肯定有不同的体征和生活方式等。这里的因变量就是是否胃癌，即“是”或“否”，自变量就可以包括很多了，例如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的

(<http://blog.csdn.net/suipingsp/article/details/41822313>)