# 逻辑回归

逻辑回归（Logistic Regression）是机器学习中的一种二分类模型（主要为二分类应用，Softmax 回归是直接对逻辑回归在多分类的推广，即多元逻辑回归），由于算法的简单和高效，在实际中应用非常广泛。

主要用途：

寻找危险因素：什么样的用户通常会是“bad”用户？

预测判别：用户是 “good”还是“bad”? 多大的概率是好用户，多大的概率是差用户？

二分类问题及应用：

获客评估：判断用户是否会成为平台客户；

交易评估：判断用户申请贷款时是否能通过；

风险评估：判断用户还款时是否会预期；

客户留存：判断用户是否会流失；

客户价值：判断用户是否能产生预期的价值；

用户画像：判断用户是否具备某一属性（如性别、品类偏好等）

这些问题都可以看成是二分类问题，这些问题本身有很重要的价值，能够帮助我们更好的了解我们的用户，服务我们的用户。

逻辑回归的适用性

1） 可用于概率预测，也可用于分类。

2） 仅能用于线性问题

只有在feature和target是线性关系时，才能用Logistic Regression（不像SVM那样可以应对非线性问题）。这有两点指导意义，一方面当预先知道模型非线性时，果断不使用Logistic Regression； 另一方面，在使用Logistic Regression时注意选择和target呈线性关系的feature。

3） 各feature之间不需要满足条件独立假设，但各个feature的贡献是独立计算的。

逻辑回归不像朴素贝叶斯一样需要满足条件独立假设（因为它没有求后验概率）。但每个feature的贡献是独立计算的，即LR是不会自动帮你combine 不同的features产生新feature的 (这是决策树,LSA, pLSA, LDA或者你自己要干的事情)。举个例子，如果你需要TF\*IDF这样的feature，就必须明确的给出来，若仅仅分别给出两维 TF 和 IDF 是不够的，那样只会得到类似 a\*TF + b\*IDF 的结果，而不会有 c\*TF\*IDF 的效果。

逻辑回归一般过程

收集数据：主要是根据业务目标收集相关的数据，模型效果往往和所用特征密切相关

准备数据：因为需要进行距离计算，因而要求数据类型为数值型。另外，结构化数据格式最佳

分析数据：分析并进行数据的预处理

训练算法：大部分时间将用于训练，目的是为了找到最佳的分类系数

寻找h函数（即hypothesis）；

构造J函数（损失函数）；

想办法使得J函数最小并求得回归参数（θ）

测试算法：模型评估

使用算法：获取新数据，转化成对应的结构化数值，基于训练好的回归系数就可以对这些数值进行简单的回归计算，判定新数据属于哪个类别；在这之后，可以在输出的类别上做一些其他分析工作

逻辑回归的关键概念

sigmod函数（hypothesis）

1.数学形式：

g(x)=11+e−x

逻辑回归简介

逻辑回归（Logistic Regression）是机器学习中的一种二分类模型（主要为二分类应用，Softmax 回归是直接对逻辑回归在多分类的推广，即多元逻辑回归），由于算法的简单和高效，在实际中应用非常广泛。

主要用途：

寻找危险因素：什么样的用户通常会是“bad”用户？

预测判别：用户是 “good”还是“bad”? 多大的概率是好用户，多大的概率是差用户？

二分类问题及应用：

获客评估：判断用户是否会成为平台客户；

交易评估：判断用户申请贷款时是否能通过；

风险评估：判断用户还款时是否会预期；

客户留存：判断用户是否会流失；

客户价值：判断用户是否能产生预期的价值；

用户画像：判断用户是否具备某一属性（如性别、品类偏好等）

这些问题都可以看成是二分类问题，这些问题本身有很重要的价值，能够帮助我们更好的了解我们的用户，服务我们的用户。

逻辑回归的适用性

1） 可用于概率预测，也可用于分类。

2） 仅能用于线性问题

只有在feature和target是线性关系时，才能用Logistic Regression（不像SVM那样可以应对非线性问题）。这有两点指导意义，一方面当预先知道模型非线性时，果断不使用Logistic Regression； 另一方面，在使用Logistic Regression时注意选择和target呈线性关系的feature。

3） 各feature之间不需要满足条件独立假设，但各个feature的贡献是独立计算的。

逻辑回归不像朴素贝叶斯一样需要满足条件独立假设（因为它没有求后验概率）。但每个feature的贡献是独立计算的，即LR是不会自动帮你combine 不同的features产生新feature的 (这是决策树,LSA, pLSA, LDA或者你自己要干的事情)。举个例子，如果你需要TF\*IDF这样的feature，就必须明确的给出来，若仅仅分别给出两维 TF 和 IDF 是不够的，那样只会得到类似 a\*TF + b\*IDF 的结果，而不会有 c\*TF\*IDF 的效果。

逻辑回归一般过程

收集数据：主要是根据业务目标收集相关的数据，模型效果往往和所用特征密切相关

准备数据：因为需要进行距离计算，因而要求数据类型为数值型。另外，结构化数据格式最佳

分析数据：分析并进行数据的预处理

训练算法：大部分时间将用于训练，目的是为了找到最佳的分类系数

寻找h函数（即hypothesis）；

构造J函数（损失函数）；

想办法使得J函数最小并求得回归参数（θ）

测试算法：模型评估

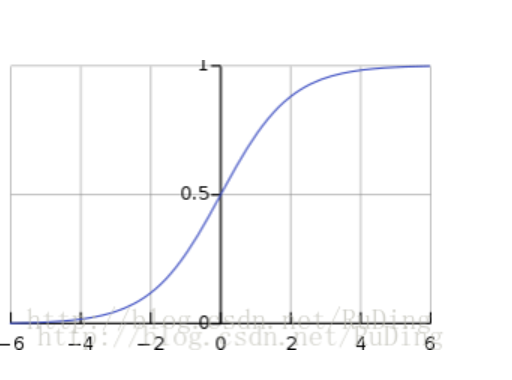
使用算法：获取新数据，转化成对应的结构化数值，基于训练好的回归系数就可以对这些数值进行简单的回归计算，判定新数据属于哪个类别；在这之后，可以在输出的类别上做一些其他分析工作

逻辑回归的关键概念

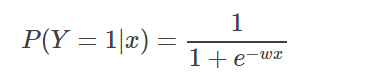
sigmod函数（hypothesis）

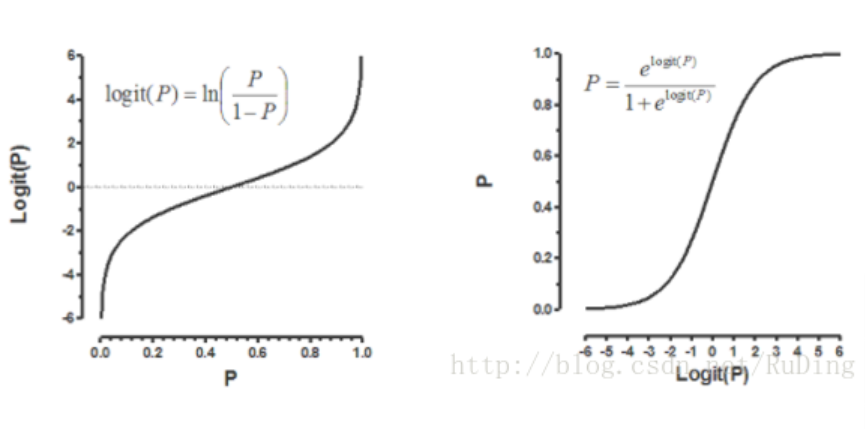
1.数学形式：

g(x)=11+e−x



2.逻辑回归的sigmod函数



最优化理论（损失函数）

1.介绍：逻辑回归的优化目标函数：

①LR输出的是分到每一类的概率，则参数估计的方法为最大似然估计 (MLE) 。

②假设训练的每个样本独立，输出为 y = {0, 1}，样本的似然函数就是将所有训练样本 label 对应的输出节点上的概率相乘, 令 p = P(Y=1|x) ,如果 y = 1, 概率就是 p， 如果 y = 0, 概率就是 1 - p , 将这两种情况合二为一，得到似然函数：

∏[P(Y=1|xi)yi][1−P(Y=1|xi)1−yi]∏[P(Y=1|xi)yi][1−P(Y=1|xi)1−yi]

对应的似然函数如下：

L(w)=∑[yi(w∗xi)−ln(1+exp(w∗xi))]L(w)=∑[yi(w∗xi)−ln(1+exp(w∗xi))]

3.难点：全局最优VS局部最优

梯度上升/梯度下降最优化算法：（通常采用梯度下降最优化算法）

这里主要介绍两种梯度优化算法，其中批量梯度优化算法每一次迭代需要扫描所有的样本数据；随机梯度下降法每一次迭代只需要考虑一个样本即可。明显的，随机梯度优化算法在计算消耗上由于批量梯度优化算法，但是可能实际的求取最优值得效果比不上批量梯度优化算法。对于凸函数，随机梯度下降算法最终的结果与批量梯度优化算法差不多。

目标函数：J(w)=−1nL(w)J(w)=−1nL(w)

批量梯度下降法：

wi=wi−∑α∂∂wJ(w)wi=wi−∑α∂∂wJ(w)

随机梯度下降法：

for j=1 to m {

wi=wi−∑α∂∂wJ(w)wi=wi−∑α∂∂wJ(w)

(for all i)

}

正则化

过拟合问题：逻辑回归的过拟合问题主要是特征太多造成的，一般不存在过拟合问题，特征过多可以采取范式的方式加以解决

解决方法

多分类：

可以利用逻辑回归进行多分类，与上述二分类逻辑回归的主要区别在于假设函数，上述为sigmod函数，多分类问题可以采用softmax函数进行映射。

总结

逻辑回归的优点：

1）实现简单，对因变量的分布没有要求；

2）对构建的模型有较强的解释性；

3）可以适用于连续性（特征离散化处理）和类别性自变量；

4）分类时计算量小，速度快，存储资源低

逻辑回归的不足：

1）处理海量、多维、非线性的数据时效果不是很好；

2）算法的范化能力、灵活性比较差；

3）算法处理数据的不对称性问题能力不足；

4）容易欠拟合，一般准确度不太高

逻辑回归的数学模型和求解都相对比较简洁，实现相对简单。通过对特征做离散化和其他映射，逻辑回归也可以处理非线性问题，是一个非常强大的分类器。因此在实际应用中，当我们能够拿到许多低层次的特征时，可以考虑使用逻辑回归来解决我们的问题。