# LR与线性回归的区别与联系

联系：都是线性模型，在线性回归模型应用sigmoid函数转化为对数几率模型；2. 回归vs 分类 3. 损失函数： 最小平方差；最小对数似然

区别：(1)线性回归要求变量服从正态分布，logistic回归要求变量服从伯努利分布。(2)线性回归要求因变量是连续性数值变量，而logistic回归要求因变量是分类型变量。(3)线性回归要求自变量和因变量呈线性关系，而logistic回归不要求自变量和因变量呈线性关系(4)线性回归是直接分析因变量与自变量的关系，而logistic回归是分析因变量取某个值的概率与自变量的关系。

LR优势：线性回归在整个实数域范围内进行预测，敏感度一致，而分类范围，需要在[0,1]。逻辑回归都是以线性回归为理论支持的。但线性回归模型无法做到sigmoid的非线性形式，sigmoid可以轻松处理0/1分类问题。逻辑回归就是一种减小预测范围，将预测值限定为[0,1]间的一种回归模型，因而对于这类问题来说，逻辑回归的鲁棒性比线性回归的要好。

# 逻辑斯特回归为什么要对特征进行离散化。

在工业界，很少直接将连续值作为逻辑回归模型的特征输入，而是将连续特征离散化为一系列0、1特征交给逻辑回归模型，这样做的优势有以下几点：

a. 离散特征的增加和减少都很容易，易于模型的快速迭代；

b. 稀疏向量内积乘法运算速度快，计算结果方便存储，容易扩展；

c. 离散化后的特征对异常数据有很强的鲁棒性：比如一个特征是年龄>30是1，否则0。如果特征没有离散化，一个异常数据“年龄300岁”会给模型造成很大的干扰；

d. 逻辑回归属于广义线性模型，表达能力受限；单变量离散化为N个后，每个变量有单独的权重，相当于为模型引入了非线性，能够提升模型表达能力，加大拟合；

e. 离散化后可以进行特征交叉，由M+N个变量变为M\*N个变量，进一步引入非线性，提升表达能力；

f. 特征离散化后，模型会更稳定，比如如果对用户年龄离散化，20-30作为一个区间，不会因为一个用户年龄长了一岁就变成一个完全不同的人。当然处于区间相邻处的样本会刚好相反，所以怎么划分区间是门学问；

g. 特征离散化以后，起到了简化了逻辑回归模型的作用，降低了模型过拟合的风险。

# LR和SVM的区别与联系

联系：LR和SVM都可以处理分类问题，且一般都用于处理线性二分类问题（在改进的情况下可以处理多分类问题）。

①都是线性分类器。本质上都是求一个最佳分类超平面。

②都是监督学习算法。

③都是判别模型。判别模型不关心数据是怎么生成的，它只关心信号之间的差别，然后用差别来简单对给定的一个信号进行分类。常见的判别模型有：KNN、SVM、LR，常见的生成模型有：朴素贝叶斯，隐马尔可夫模型。

区别：1) 本质上是损失函数不同 。LR的损失函数是交叉熵。逻辑回归基于概率理论，假设样本为正样本的概率可以用sigmoid函数（S型函数）来表示，然后通过极大似然估计的方法估计出参数的值。 支持向量机基于几何间隔最大化原理，认为存在最大几何间隔的分类面为最优分类面。

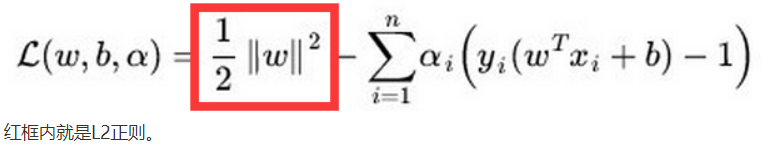
2) 两个模型对数据和参数的敏感程度不同 。SVM考虑分类边界线附近的样本（决定分类超平面的样本）。在支持向量外添加或减少任何样本点对分类决策面没有任何影响； LR受所有数据点的影响。直接依赖数据分布，每个样本点都会影响决策面的结果。如果训练数据不同类别严重不平衡，则一般需要先对数据做平衡处理，让不同类别的样本尽量平衡。

3) SVM 基于距离分类，LR 基于概率分类。 SVM依赖数据表达的距离测度，所以需要对数据先做 normalization；LR不受其影响。

4) 在解决非线性问题时，支持向量机采用核函数的机制，而LR通常不采用核函数的方法。 SVM算法里，只有少数几个代表支持向量的样本参与分类决策计算，也就是只有少数几个样本需要参与核函数的计算。 LR算法里，每个样本点都必须参与分类决策的计算过程，也就是说，假设我们在LR里也运用核函数的原理，那么每个样本点都必须参与核计算，这带来的计算复杂度是相当高的。尤其是数据量很大时，我们无法承受。所以，在具体应用时，LR很少运用核函数机制。

5) 在小规模数据集上，Linear SVM要略好于LR，但差别也不是特别大，而且Linear SVM的计算复杂度受数据量限制，对海量数据LR使用更加广泛。

6) SVM的损失函数就自带正则，而 LR 必须另外在损失函数之外添加正则项。



# L1、L2正则化

<https://www.julyedu.com/question/big/kp_id/23/ques_id/988>

# 梯度下降法找到的一定是下降最快的方向么？

https://www.julyedu.com/question/big/kp\_id/23/ques\_id/1017

# 说说常见的优化算法及其优缺点？

https://www.julyedu.com/question/big/kp\_id/23/ques\_id/1524

# 常见的分类算法有哪些？他们各自的优缺点是什么？

https://www.julyedu.com/question/big/kp\_id/23/ques\_id/1366

# 请写出你了解的机器学习特征工程操作，以及它的意义

https://www.julyedu.com/question/big/kp\_id/23/ques\_id/2116

1 Logistic Regression 的前世今生（理论篇）：http://blog.csdn.net/cyh\_24/article/details/50359055

2 机器学习算法与Python实践之（七）逻辑回归：http://blog.csdn.net/zouxy09/article/details/20319673