逻辑回归和线性回归

2019年5月15日 19:13

1.二者区别

线性回归模型：$Y = WX$(假设$W > 0$)

线性回归是根据样本X各个维度的X_i的线性叠加（线性叠加的权重系数wi就是模型的参数）来得到预测值的Y，然后最小化所有的样本预测值Y与真实值y'的误差来求得模型参数。从这里可以看出模型的值Y是样本X各个维度的X_i的线性叠加，是线性的。

进行数值预测

逻辑回归模型：$Y = \frac{1}{1+e^{-W^T X}}$(假设$W > 0$)

Y的值大小不是随X叠加和的大小线性的变化了，而是一种平滑的变化，这种变化在x的叠加和为0附近的时候变化的很快，当X各维度叠加和取无穷大的时候，Y趋近于1，当X各维度叠加和取无穷小的时候，Y趋近于0.

在logistic回归中，X各维度叠加和（或X各维度）与Y不是线性关系，而是logistic关系。而在线性回归中，X各维度叠加和就是Y，也就是Y与X就是线性的了。

进行类型分类

2.联系

逻辑回归与线性回归都属于广义线性回归模型

逻辑回归原理

2019年5月15日 19:33

逻辑回归是一个分类算法，他是在线性回归的基础上加入了sigmoid函数，将线性回归的结果输入至sigmoid函数中，并且设定一个阈值，如果大于阈值为1，小于阈值为0。

逻辑回归损失函数的推导与优化

2019年5月15日 19:57

1.交叉熵损失函数的推导

将模型用概率表示：

$P\left(y^{(i)}=1\middle|x^{(i)};\theta\right)=h_\theta\left(x^{(i)}\right)$

$P\left(y^{(i)}=0\middle|x^{(i)};\theta\right)=1-h_\theta\left(x^{(i)}\right)$

将两式整合：

$P\left(y^{(i)}\middle|x^{(i)};\theta\right)=h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta\left(x^{(i)}\right))^{1-y^{(i)}}$

将模型优化问题看作是极大似然估计问题：

$$L(\theta)=P\left(\vec{y}\middle|x^{(i)};\theta\right)=\prod_{i=1}^{m}P\left(y^{(i)}\middle|x^{(i)};\theta\right)=\prod_{i=1}^{m}h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta\left(x^{(i)}\right))^{1-y^{(i)}}$$

取对数似然：

$$\log L(\theta)=\sum_{i=1}^{m}y^{(i)}log\ h_\theta\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)$$

则得证交叉熵函数

2.优化

使用梯度下降算法、牛顿法或者拟牛顿法对交叉熵损失函数进行优化

正则化与模型评估指标

2019年5月15日 20:25

1.正则化

（1）找一个合适的预测函数（Andrew Ng的公开课中称为hypothesis），一般表示为h函数，该函数就是我们需要找的分类函数，它用来预测输入数据的判断结果。这个过程时非常关键的，需要对数据有一定的了解或分析，知道或者猜测预测函数的“大概”形式，比如是线性函数还是非线性函数。

（2）构造一个Cost函数（损失函数），该函数表示预测的输出（h）与训练数据类别（y）之间的偏差，可以是二者之间的差（h-y）或者是其他的形式。综合考虑所有训练数据的“损失”，将Cost求和或者求平均，记为J(θ)函数，表示所有训练数据预测值与实际类别的偏差。

（3）显然，J(θ)函数的值越小表示预测函数越准确（即h函数越准确），所以这一步需要做的是找到J(θ)函数的最小值。找函数的最小值有不同的方法，Logistic Regression实现时有的是梯度下降法（Gradient Descent）。
https://blog.csdn.net/qq_38923076/article/details/82925183

逻辑回归优缺点

2019年5月15日 20:49

1.优点：

1）预测结果是界于0和1之间的概率；

2）可以适用于连续性和类别性自变量；

3）容易使用和解释；

2.缺点：

1）对模型中自变量多重共线性较为敏感，例如两个高度相关自变量同时放入模型，可能导致较弱的一个自变量回归符号不符合预期，符号被扭转。需要利用因子分析或者变量聚类分析等手段来选择代表性的自变量，以减少候选变量之间的相关性；

2）预测结果呈“S”型，因此从log(odds)向概率转化的过程是非线性的，在两端随着log(odds)值的变化，概率变化很小，边际值太小，slope太小，而中间概率的变化很大，很敏感。 导致很多区间的变量变化对目标概率的影响没有区分度，无法确定阀值。

样本数据不平衡

2019年5月15日 20:53

严格地讲，任何数据集上都有数据不平衡现象，这往往由问题本身决定的。不平衡程度相同（即正负样本比例类似）的两个问题，解决的难易程度也可能不同，因为问题难易程度还取决于我们所拥有数据有多大。

1.采样
采样方法是通过对训练集进行处理使其从不平衡的数据集变成平衡的数据集，在大部分情况下会对最终的结果带来提升。

采样分为上采样（Oversampling）和下采样（Undersampling），上采样是把小众类复制多份，下采样是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。

2.数据合成
采样方法是通过对训练集进行处理使其从不平衡的数据集变成平衡的数据集，在大部分情况下会对最终的结果带来提升。

采样分为上采样（Oversampling）和下采样（Undersampling），上采样是把小众类复制多份，下采样是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。

3.加权
对不同类别分错的代价不同

4.一分类
对于正负样本极不平衡的场景，可以换一个完全不同的角度来看待问题：把它看做一分类（One Class Learning）或异常检测（Novelty Detection）问题。这类方法的重点不在于捕捉类间的差别，而是为其中一类进行建模，经典的工作包括One-class SVM等。
https://blog.csdn.net/jemila/article/details/77992967