逻辑回归始末

同学们，大家好！

这次给大家讲解一下逻辑回归的原理及应用。

今天要讲的逻辑回归，和线性回归有着非常紧密的联系，那么逻辑回归到底是什么呢，它和线性回归解决同样的分类问题吗？

逻辑回归是什么

逻辑回归是机器学习中的一种分类模型。虽然名字中带有回归，但是本质是一种分类算法。之所以没有同线性回归一样解决回归问题，是因为它在线性回归的结果上，加上了sigmoid激活函数，这个我们会在后面慢慢道来。由于算法的简单高效，在实际中逻辑回归应用非常广泛。

遵循先易后难的原则，我们先了解一下逻辑回归在哪些地方有应用。

逻辑回归的应用场景

- 天气晴雨预测
- 是否垃圾邮件
- 是否患病
- 是否金融诈骗
- 是否虚假账号

看到上面的这些例子，我们可以发现其中的特点，那就是它们都属于两个类别之间的判断。逻辑回归就是解决二分类问题的利器。

了解了逻辑回归的几个应用场景，大家一定对其原理非常好奇，下面我们讲解逻辑回归的原理。

逻辑回归的原理

要掌握逻辑回归的原理，必须要解决这两个问题：

- 逻辑回归的输入值是什么
- 如何判断逻辑回归的输出

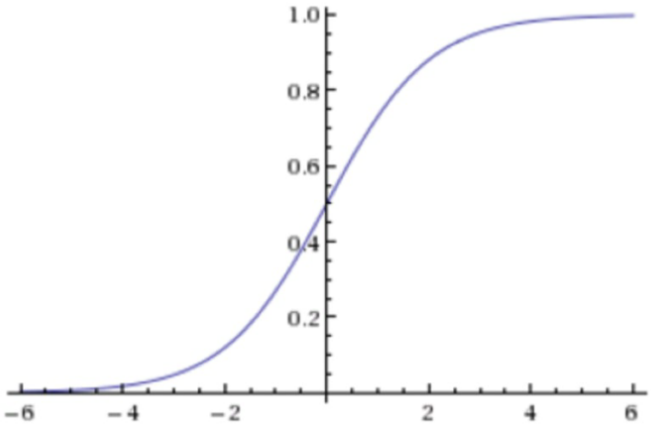
首先看第一个问题，逻辑回归的输入值是什么。
逻辑回归的输入，就是线性回归的输出，就是下面线性回归公式结果。这就呼应了前面我们说的，逻辑回归和线性回归有着非常紧密的联系。

$$h(w) = w_1x_1 + w_2x_2 + w_3x_3 \ldots + \mathrm{b}$$

上面是逻辑回归的输入值，是逻辑回归与线性回归的联系之处，下面是逻辑回归区别于线性回归的地方，也就是我们要解决的第二个问题。我们先来了解一下sigmoid激活函数。sigmoid激活函数公式如下：

$$g(w^T, x) = \frac{1}{1 + e^{-h(w)}} = \frac{1}{1 + e^{-w^T x}}$$

线性回归的结果输入到sigmoid激活函数当中，输出结果是 [0, 1] 区间中的一个概率值，默认0.5为阈值，逻辑回归最终的分类是通过属于某个类别的概率值来判断是否属于某个类别，并且这个类别默认标记为1（1表示正例），另外一个类别会标记为0（0表示反例）。



下面以一个简单例子给大家解释逻辑回归的输出。
假设有两个类别A和B，并且假设我们的概率值为属于A这个类别的概率值，现在有一个样本的输入到逻辑回归输出结果是0.55，那么这个概率值超过0.5，意味着我们训练或者预测的结果就是A类别。那么反之，如果得出结果为0.3，训练或者预测结果就为B类别。关于逻辑回归的阈值是可以改变的，比如上面举例中，如果把阈值设置为0.6，那么输出结果为0.55，就属于B类。

在逻辑回归中，当预测结果不对的时候，我们该怎么衡量他的损失呢?

逻辑回归的损失以及优化

逻辑回归的损失，我们称之为对数似然损失，公式如下：

$$cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if y=1} \\ -log(1 - h_\theta(x)) & \text{if y=0} \end{cases}$$

其中y为真实值，hΘ(x)为预测值。
无论何时，我们都希望损失函数值越小越好。分情况讨论，对应的损失函数值，当y=1时，我们希望hΘ(x)的值越大越好，当y=0时，我们希望hΘ(x)的值越小越好。
综合完整的损失函数：

$$cost(h_\theta(x), y) = \sum_{i=1}^{m} -y_i log(h_\theta(x)) - (1 - y_i) log(1 - h_\theta(x))$$

如果你看懂了上述的逻辑回归原理和损失优化，那么恭喜你get了逻辑回归的六成知识，剩下四成是逻辑回归的评价指标，或者称之为分类任务的评估方法。

分类评估方法

在分类评估方法这一小节中，我们要了解什么是混淆矩阵，要知道分类评估中的精确率和召回率，要知道评价指标F1-score反映了什么特性。
在分类任务下，预测结果与真实结果之间存在四种不同的组合，构成混淆矩阵。听到这里，大家可能有点懵圈，不要慌听我慢慢道来。一个样本通过模型，得到预测结果，这个预测结果可以是正例，也可以是假例。样本本身有一个标签值，这个标签值可以是正例，也可以是假例。当真实结果和预测结果都为正例时，样本为真正例TP；当真实结果为正例，而预测结果为假例，那么样本为伪反例，顾名思义，假的反例；当真实结果为假例，预测结果却为正例，那么样本为伪正例；当真实结果和预测结果都为假例时，样本为真反例。而这些都是为了下面打基础。

预测结果

真实结果	正例	假例
正例	真正例TP	伪反例FN
假例	伪正例FP	真反例TN

下面就是硬菜。

首先是精确率，精确率是预测结果为正例样本中真实为正例的比例，衡量模型查得准不准，计算公式是：精确率 = TP / (TP + FP)

然后是召回率，召回率是真实为正例的样本中预测结果为正例的比例，衡量模型查得全不全，评估模型对正样本的区分能力，计算公式为：召回率 = TP / (TP + FN)

最后是F1-score，它综合反映了模型的稳健性，计算公式如下：

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

上述就是逻辑回归的全部内容，恭喜你又解锁了一个机器学习的经典模型，学习是一场知识的旅行，祝福大家在旅途中披荆斩棘，收获硕果。