逻辑回归模型

0、概述

逻辑回归模型（Logistic Regression）借回归之名，行分类之实，历来为人所不齿。最基本的逻辑回归模型用于二类分类，它属于概率模型，且属于判别模型。它是在线性回归模型的外面套了一个logistic函数（也叫sigmoid函数），将输出映射到[0,1]区间上，从而赋予了输出以概率的意义。

注意，逻辑回归与线性回归回归模型一样，适用于样本的特征都是连续型特征的情况。在这种情况下，样本特征的取值的大小会影响输出结果。

用于解决二类分类问题的模型有很多，它们的基本思想都是求解一个超平面，使该超平面能够尽量将两类样本分到其两侧，如下图所示：

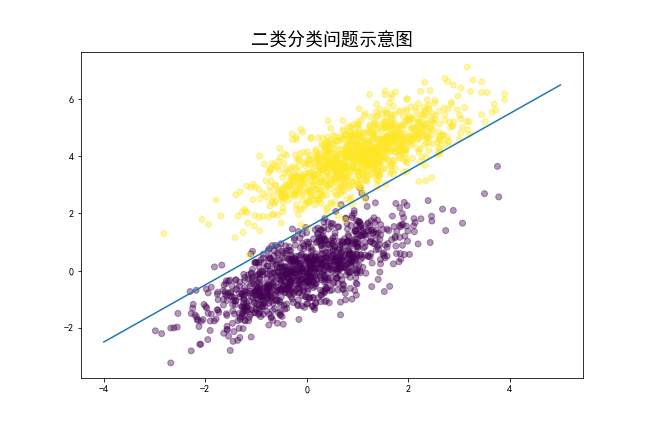


图1 二类分类问题示意图

在这些分类模型中，逻辑回归模型的优点是：它不仅能给出一个合适的超平面，还能对于每一个样本给出它属于每一类的“可能性”或者说“置信度”。如何给出“置信度”呢？从直观上看，与超平面越近的样本越可能分到另一类（置信度越低）；而与超平面距离越远，则该样本属于它本身的类的确定性就越高（置信度越高）。逻辑回归模型用最简单的方式实现了这一“置信度”的量化。

对于分类模型，有一个非常基本且重要的结论，须记住：只有在数据是可分的情况下，考虑对其进行分类才有意义，否则是无意义的。当然，数据的“可分性”也是有程度之分的。如下图所示，基本上可以认为这两类样本是不可分的：

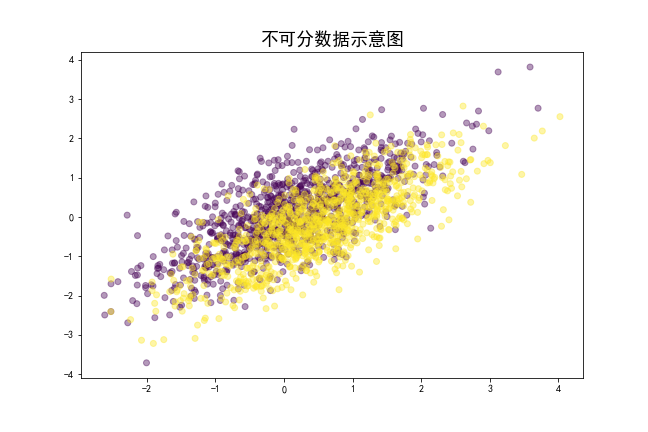


图2 不可分数据示意图

一、逻辑回归模型的形式

首先介绍一下logistic函数及其特点。我们用g(z)表示logistic函数：

|  |  |
| --- | --- |
|  | (1) |

它的图像如下：

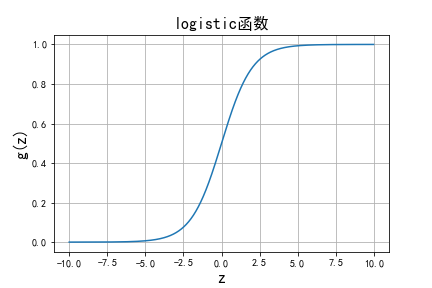


图3 logistic函数示意图

可以看出，logistic函数将区间映射到[0,1]区间，且该函数中心对称，当z=0时，g(z)=0.5。

logistic函数的导数具有以下性质：

|  |  |
| --- | --- |
|  | (2) |

为什么要使用logistic函数呢？因为它具有几个优点：简单；对称；可以将区间映射到[0,1]区间，从而赋予输出以概率的意义；在中心处（z=0处）斜率很大，这样对于不易区分的值（即靠近中心的值），能有较大的区分度。

我们将二类分类问题的两个类别分别编码为1和0。下面借助logistic函数，给出逻辑回归模型的形式：

给定一个样本的输入，逻辑回归模型首先计算它输出为1的概率：

|  |  |
| --- | --- |
|  | (3) |

其中表示一个样本的输入；和b是模型参数，其意义与线性回归模型的参数类似。

为了方便，我们这里采用与线性回归模型中相同的记法：将b表示成w的一个分量，即，此时样本输入x也相应地变成，则(3)式中的就变成了。我们得到：

|  |  |
| --- | --- |
|  | (4) |

根据(4)式，我们可以直接得到样本输出为0的概率：

|  |  |
| --- | --- |
|  | (5) |

需要注意的是，虽然逻辑回归模型最重要的一步是计算概率，但其最终输出并不是概率值，而是分类结果（0或1），相应地，训练数据的标签也是0/1类别。因此完整的逻辑回归模型的形式为：

|  |  |
| --- | --- |
|  | (6) |

可以看出，给定一个样本，逻辑回归模型预测它的输出的步骤为：首先计算P(Y=1|x)和P(Y=0|x)，然后比较二者大小，将样本分到概率较大的类中。

二、构建逻辑回归模型的目标函数

设给定训练集，训练集中有m个样本，每个样本的输入具有n维连续型特征，并有一个0/1类别标签。我们需要给定一个模型性能评估准则，基于该准则推导出模型的目标函数。

一般来说，回归模型的性能评估准则是MSE，比如线性回归模型的目标函数就是基于MSE推导出来的；而分类模型的性能评估准则就比较多了，如准确率、查准率、查全率、F1-Score、AUC、交叉熵等。

对于任意概率模型，我们都可以基于极大似然估计推导其目标函数。下面我们针对一般的二类分类模型推导其极大似然估计的目标函数。我们将会看到，推导出的目标函数与交叉熵目标函数是一致的。

设，则，则对给定的样本有：

则模型的似然函数为：

|  |  |
| --- | --- |
|  | (7) |

在上式中，对于输出为1的样本，它出现的概率为；对于输出为0的样本，它出现的概率为。上式通过简单的指数操作将分段函数写成一个统一的式子的形式，颇为巧妙。

则对数似然函数为：

|  |  |
| --- | --- |
|  | (8) |

可以看出，这就是交叉熵的相反数。我们知道，“信息熵”的概念表示消除系统不确定性的最小代价；而这里的交叉熵表示用预测的分布去消除具有真实分布的系统的不确定性的代价，显然，代价越小则预测分布越接近真实分布。极大化似然函数就等价于极小化交叉熵。

对于逻辑回归模型，我们把它的模型形式带入上式，有：

|  |  |
| --- | --- |
|  | (9) |

注意，尽管logistic回归模型的最终输出是0/1类别，但是其目标函数的构建是基于作为中间输出的概率的。也就是说，它的目标不仅考虑最终分类的准确率，还考虑作为最终分类的依据的概率分布的好坏。

三、逻辑回归模型的优化

逻辑回归模型的优化通常使用梯度下降法。这里省略推导过程，仅给出梯度下降法的参数迭代格式。设学习率为，有：

|  |  |
| --- | --- |
|  | (10) |

与线性回归模型一样，w的各分量的更新应同时进行。

四、逻辑回归模型的改进

对于分类任务，我们不仅需要解决二类分类问题，常常还需要解决多分类问题。

实际中，除了直接使用多分类的模型，还常常将二分类模型拓展成多分类模型。常见的两种拓展策略是one-vs-rest和one-vs-one。

比如对于K类分类问题，one-vs-rest的基本思路是构造一个二叉树，二叉树的每一层是一个二分类器，它将上一层分下来的样本分成“属于某类”或“不属于某类”这两类，最终便能对所有样本进行划分。

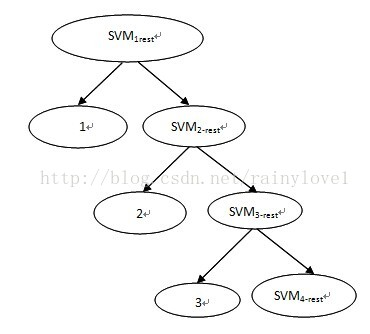


图4 one-vs-rest示意图（SVM）

而one-vs-one是对任意两个类，训练一个二分类器，这样如果有K类，将训练出K\*(K-1)/2个二分类器，然后对一个新样本，这些二分类器的分类结果进行投票以给出最终分类结果。

对于逻辑回归模型，还有一种方法将其扩展为多类分类模型，就是softmax回归，softmax回归模型跟logistic回归模型类似，给定一个样本x，模型计算该样本属于每个类的概率：

|  |  |
| --- | --- |
|  | (11) |

可以看出，这样定义的概率满足归一化条件，是合理的。

计算完概率以后，将样本分到概率最大的那一类中。