

机器学习实战

Stephen CUI¹

February 17, 2023

¹cuixuanStephen@gmail.com

Contents

I	分类	3
1	Logistic回归	4
1.1	基于 Logistic 回归和 Sigmoid 函数的分类	4
1.2	基于最优化方法的最佳回归系数确定	4
2	树回归	5

Part I

分类

前两部分主要探讨监督学习（supervised learning）。在监督学习的过程中，我们只需要给定输入样本集，机器就可以从中推演出指定目标变量的可能结果。

监督学习一般使用两种类型的目标变量：标称型和数值型。标称型目标变量的结果只在有限目标集中取值，如真与假、动物分类集合 { 爬行类、鱼类、哺乳类、两栖类 }；数值型目标变量则可以从无限的数值集合中取值，如 0.100、42.001、1000.743 等。数值型目标变量主要用于回归分析。

Chapter 1

Logistic回归

假设现在有一些数据点，我们用一条直线对这些点进行拟合（该线称为最佳拟合直线），这个拟合过程就称作回归。利用Logistic回归进行分类的主要思想是：**根据现有数据对分类边界线建立回归公式，以此进行分类**。这里的“回归”一词源于最佳拟合，表示要找到最佳拟合参数集。

1.1 基于 Logistic 回归和 Sigmoid 函数的分类

Logistic回归

优点：计算代价不高，易于理解和实现。

缺点：容易欠拟合，分类精度可能不高。

适用数据类型：数值型和标称型数据。

我们想要的函数应该是，能接受所有的输入然后预测出类别。例如，在两个类的情况下，上述函数输出0或1。或许你之前接触过具有这种性质的函数，该函数称为**海维塞德阶跃函数**（Heaviside step function），或者直接称为单位阶跃函数。然而，海维塞德阶跃函数的问题在于：该函数在跳跃点上从0瞬间跳跃到1，这个瞬间跳跃过程有时很难处理。幸好，另一个函数也有类似的性质¹，且数学上更易处理，这就是Sigmoid函数。Sigmoid函数具体的计算公式如下：

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

为了实现Logistic回归分类器，我们可以在每个特征上都乘以一个回归系数，然后把所有的结果值相加，将这个总和代入Sigmoid函数中，进而得到一个范围在0 1之间的数值。任何大于0.5的数据被分入1类，小于0.5即被归入0类。所以，Logistic回归也可以被看成是一种概率估计。

1.2 基于最优化方法的最佳回归系数确定

Sigmoid函数的输入记为 z ，由下面公式得出：

$$z = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

¹Sigmoid函数是一种阶跃函数（step function）。在数学中，如果实数域上的某个函数可以用半开区间上的指示函数的有限次线性组合来表示，那么这个函数就是**阶跃函数**。而数学中**指示函数**（indicator function）是定义在某集合 X 上的函数，表示其中有哪些元素属于某一子集 A 。

Chapter 2

树回归