线性回归模型

0、概述

线性回归模型是机器学习中形式最简单的回归模型，但也是很常用的一个模型，它假设了输入和输出具有线性关系，并求解出最佳的拟合式。

线性回归模型通常是入门机器学习的第一个模型，因为它简单，所以用它为作为例子能清楚地说明很多所有机器学习模型中共同的问题，比如机器学习中的过拟合和欠拟合、机器学习中的优化方法、机器学习中的正则化等。

通常对于任何一个机器学习模型的了解，都是从四个方面对其进行研究：模型的形式、模型的目标函数、模型的优化算法、模型的正则化策略。下面我们依次介绍这四个方面。

一、线性回归模型的形式

线性回归模型的形式为：

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

其中，表示一个样本的输入，它是n维向量，每个分量是样本的一个特征；和b是模型参数，前者又称为（特征的）系数，后者又称为截距或偏置；是模型对该样本的输出，是一个实值标量。

一般情况下，我们表示的向量都默认是列向量，因此在式子中要对转置，变成行向量，再与相乘。

可以看到，x和w上面有小箭头，是为了强调它们是向量。然而大多情况下，小箭头被省略，即写作：

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

但一定要注意区分表达式中的向量和标量。

线性回归模型的形式还有一种更简洁的表示方法，就是将b表示成w的一个分量，即，此时样本输入x也相应地变成，模型的形式就写成：

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

但是注意，这里的w和x，跟(2)式中的w和x都不相同。

后面我们就使用这种简化的形式。

可以看出，线性回归模型对一个样本x的预测值，就是将该样本的各维特征的取值加权求和，因此求和项中的某一项（对应于某一维特征）的值相对其他项的值越大，则该项对预测结果的影响就越大。由于每一项由某一维特征的取值和相应的权重相乘得到，因此特征取值大小和相应的权重大小都对预测结果有影响。如果各维特征取值的值域都相同，则权重就能公平地衡量各维特征的变化对于预测结果的变化的重要程度。权重越大，对应的特征的变化对于预测结果的变化来说越重要。

二、构建线性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维特征，并对应一个实值的标签。注意，对于样本数据，上标表示第几个样本，下标表示某个样本的第几维特征（或说分量）。我们的目标是在训练集上拟合一个线性模型。

拟合一个线性模型，需要设定一个最优准则（准则定义了在训练集上有怎样表现的模型是最优的模型），依照这个准则可以推导出模型的“目标函数”（也称为损失函数），求解目标函数的最值，得到使模型成为最优模型的参数w和b，就得到了最优模型。

对于线性回归模型，我们有两种准则：

**1、最小均方误差（MSE）准则**

MSE准则的思路是，认为最优的模型是使模型在样本集上的预测的均方误差最小，即：

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

注意，其中每个样本输入都是一个n维向量：。

因此求解最优模型就变成了求解使函数取得最小值的参数w，函数称为模型在训练集T上的目标函数。

**2、极大似然准则**

该准则的思路是，从统计学中的抽样理论的角度来看，给定的训练集相当于是随机向量X和随机变量Y联合抽样得到的样本的观察值。设，Z是X和Y的函数，w是未知参数。我们假设Z服从均值为0的正态分布（这样假设是朴素而合理的），则Z的概率密度函数为：

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

其中z是随机变量Z的观察值。根据Z与(X,Y)的关系，可以得到：

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

其中x和y分别是X和Y的观察值且满足。

Z的m个样本的联合概率密度函数，也就是其似然函数为：

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

其中，w是未知参数。根据极大似然估计的思想，最佳的w是使样本观察值出现的概率最大的w，也就是使似然函数最大的w。所以似然函数就可以作为目标函数。

为了化简目标函数，我们取对数似然函数：

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

可以看出，前两项都是常数，第三项的求和项前的系数也是常数，因此极大化似然函数就相当于最小化，这与MSE准则推导出的目标函数是相同的。

上面我们介绍了两种构建目标函数的准则，它们殊途同归，导出了同样形式的目标函数，然而这种形式看起来不够“简洁”。实际上，目标函数也可以完全用矩阵和向量来表示，从而避免求和符号。

对m个样本输入数据构建X矩阵：

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

对m个样本标签数据构建y向量：

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

注意，这里y是一个向量，表示所有样本的标签，而不是一个样本的标签；另外，也要注意样本标签与模型对样本的输出的区别，后者通常上面有个“小帽子”（）。

则目标函数变为：

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

不知读者有没有注意到，这实际上是在求解一个线性的超定方程组。

另一个需要注意的地方是，在很多书籍和教程中。以上公式中的某些符号经常会混用，比如 “L”有时表示目标函数，有时表示似然函数；f有时表示概率密度函数，有时表示输入与输出的映射关系（即代表模型的形式）。对于这些混用的情况，我们需要根据上下文进行辨别。

三、线性回归模型的优化

构建好了模型的目标函数以后，下一步就是对目标函数函数进行优化。对于线性回归模型，有两种常用的优化方法：正规方程法和梯度下降法，后者适用于大部分机器学习模型，而前者只适用于线性回归模型。

**1、正规方程**

正规方程求解线性回归模型的优化问题，是基于(11)式的矩阵形式的目标函数。先对目标函数进行化简：

|  |  |
| --- | --- |
|  | (12) |
|  |  |

注意，其中和互为转置并且都是标量，故二者相等。

目标函数是关于参数w的函数，令目标函数对w的导数为0，有：

|  |  |
| --- | --- |
|  | (13) |

即：

|  |  |
| --- | --- |
|  | (14) |

则有：

|  |  |
| --- | --- |
|  | (15) |

**2、梯度下降**

梯度下降法求解线性回归模型的优化问题，是基于(4)式的求和形式的目标函数。一般来说，梯度下降法只能求解出局部极值，但线性回归问题的目标函数是一个凸函数，因此它的局部极值就是全局极值。

首先目标函数对w的每个分量求偏导数，得：

|  |  |
| --- | --- |
|  | (16) |

其中表示第i个样本的输入的第j个分量。

然后按照梯度下降法的一般步骤，给出线性回归的梯度下降流程：

（1）给出参数w的初始值；

（2）给定学习率，更新参数w：

|  |  |
| --- | --- |
|  | (17) |

注意，上式给出了w的每个分量的更新公式，实际中每次更新时，需要同时对这些分量更新。

四、线性回归模型的正则化

模型正则化用于限制模型的复杂度（从而改善过拟合现象），或者用于向模型加入先验信息。模型正则化的方法是在目标函数中加入正则化项。线性回归模型的正则化通常有两种方法：L1正则或L2正则。

**1、L1正则：**

加入L1正则化项后，目标函数变为：

|  |  |
| --- | --- |
|  | (18) |

其中是正则化项的权重，表示正则化的程度。

线性回归模型的L1正则化也称为Lasso回归。

**2、L2正则：**

加入L2正则化项后，目标函数变为：

|  |  |
| --- | --- |
|  | (19) |

线性回归模型的L2正则化也称为岭回归。

正则化的线性回归模型，其优化问题的求解同样可以使用梯度下降，具体过程不再赘述。

可以看出，线性回归模型的正则化的作用是限制w的大小，使它尽量小。

注意，除了限制w大小的功能外，L1正则化还有对系数w稀疏化的作用，而L2正则化没有稀疏化的作用。