回归模型实验报告

卢珍妮

摘要： 回归模型对统计关系进行定量描述的一种数学模型。回归分析是研究一个变量（被解释变量）关于另一个（些）变量（解释变量）的具体依赖关系的计算方法和理论。 从一组样本数据出发，确定变量之间的数学关系式对这些关系式的可信程度进行各种统计检验。利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。本文通过对回归模型中线性回归模型与多项式回归模型的实现，详细介绍回归模型的思想、原理以及具体的代码实现。

关键词： k 近邻算法；kNN 算法；分类；基本要素

1 简介

线性回归(Linear Regression)是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。 线性回归利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归,大于一个自变量情况的叫做多元回归。

多项式回归(Polynomial Regression)是研究一个因变量与一个或多个自变量间多项式的回归分析方法。如果自变量只有一个时，称为一元多项式回归；如果自变量有多个时，称为多元多项式回归。在一元回归分析中，如果依变量 y 与自变量 x 的关系为非线性的，但是又找不到适当的函数曲线来拟合，则可以采用一元多项式回归。 多项式回归的最大优点就是可以通过增加 x 的高次项对实测点进行逼近，直至满意为止。事实上，多项式回归可以处理相当一类非线性问题，它在回归分析中占有重要的地位，因为任一函数都可以分段用多项式来逼近。

2 算法

2. 1 介绍

梯度下降法是一个最优化算法，通常也称为最速下降法。最速下降法是求解无约束优化问题最简单和最古老的方法之一，虽然现已不具有实用性，但是许多有效算法都是以它为基础进行改进和修正而得到的。最速下降法是用负梯度方向为搜索方向的，最速下降法越接近目标值，步长越小，前进越慢。

2. 2 流程

2.2.1 模型

线性回归：

$$f(x) = w_0 + w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + w_4 \times x_4 + w_5 \times x_5$$

其中，$x_i=(x_1, x_2, x_3, x_4, x_5)$，$i=(1, 2, \cdots, m)$

多项式回归模型：

$$f(x) = w_0 + w_1 \times X + w_2 \times X^2 + w_3 \times X^3 + \cdots + w_n \times X^n$$

其中，$Xi=(x_1, x_2, x_3, x_4, x_5)$，$i=(1, 2, \cdots, m)$，且 W_0 是一个常数，W_i（i>1）是与 X^i 一一对应的五维向量。

损失函数：

$$\mathrm{Loss}(w_0, w_1, \ldots, w_n) = {}^1\!/_m \sum_{i=1}^{m} (f(x^{(i)}) - y^i)^2$$

2.2.2 梯度下降

$$W_0 = W_0 - \alpha\, {}^1\!/_m \sum_{i=1}^{m} (f(x^{(i)}) - y^i)$$

$$W_1 = W_1 - \alpha\, {}^1\!/_m \sum_{i=1}^{m} (f(x^{(i)}) - y^i) X^1$$

$$W_2 = W_2 - \alpha\, {}^1\!/_m \sum_{i=1}^{m} (f(x^{(i)}) - y^i) X^2$$

...

$$W_n = W_n - \alpha\, {}^1\!/_m \sum_{i=1}^{m} (f(x^{(i)}) - y^i) X^n$$

2.3 算法优缺点

回归分析为分析多因素模型时，更加简单和方便。运用回归模型，只要采用的模型和数据相同，通过标准的统计方法可以计算出唯一的结果，但在图和表的形式中，数据之间关系的解释往往因人而异，不同分析者画出的拟合曲线很可能也是不一样的。回归模型可以准确地计量各个因素之间的相关程度与回归拟合程度的高低，提高预测方程式的效果。由于实际一个变量仅受单个因素的影响的情况极少，所以一元回归分析法适用于确实存在一个对因变量影响作用明显高于其他因素的变量时。多元回归分析法比较适用于实际问题，受多因素综合影响时使用。但在回归模型中，选用哪个因变量以及该因变量采用何种表达式只是一种推测，这影响了因变量的多样性和某些因变量的不可测性，使得回归模型在某些情况下受到限制。

3. 实验

实验部分，采用 NACA0012 airfoils 数据集，展示通过梯度下降法求解的不同参数向量对测试数据集预测的误差平方和。

3.1 算法流程

算法
输入：训练数据集，测试数据集，模型的维度 输出：测试数据集上预测值与真实值的误差平方和，迭代次数 步骤： 1.读取数据集并对其进行处理，在本实验中，对数据集的 X 特征向量进行了特征缩放，使特征的范围缩放到接近的范围，否则当多个特征的范围差距过大时，损失函数的轮廓图会非常的偏斜这会导致梯度下降函数收敛的非常慢。 2.使用训练集的数据进行梯度下降，不断的更新参数向量。 3.对于不断生成的参数向量，代入测试集上求解模型预测值。计算预测值与真实值误差的平方和。 4.设置合适的学习率，正则化因子，以及迭代次数。 5.比较不同的参数向量对于误差值的影响。

3.2 实验设置

3.2.1 评价指标

为了观察学习率、正则化因子以及模型维度对于最优模型的误差值影响。下面的 3 张表分别展示了不同的学习率、模型维度以及正则化因子下的迭代次数以及训练误差和测试误差。其中，表 1 展示了学习率对于模型的影响，其中模型维度为 3，正则化因子为 0.1。表 2 展示了模型维度对于模型的影响，其中学习率为 0.5，正则化因子为 0.1。表 3 展示了正则化因子对于模型的影响，其中学习率为 0.5，模型维度为 3。

可以看出学习率对于迭代次数的影响较为明显，但对于训练误差及测试误差的值影响较小。因此说明模型性能对于学习率并不是很敏感。学习率只是会影响迭代停止的次数。模型维度对于迭代次数的影响不是很大，但是对于测试误差与训练误差的值有影响，当维度为 1 时与维度为 5 时的训练误差与测试误差的相差较大。然而，正则化因子对于迭代次数的影响很小，对于训练误差的影响较大，当正则化因子增大时，训练误差增大，测试误差也在小幅度的增大。

因此，选择一个合适的学习率，会影响求解最优模型的时间，而正则化因子与模型维度是否合适，会直接影响到最终的误差值。

表 1.展示了学习率对于模型的影响

学习率	迭代停止次数	训练误差	测试误差

0.1	5285	11.473880	11.811303
0.2	3322	11.376356	11.717978
0.3	2497	11.341620	11.686157
0.4	2029	11.323534	11.670397
0.5	1725	11.312135	11.660924

表 2.展示了模型维度对于模型的影响

模型维度	迭代停止次数	训练误差	测试误差
1	628	12.065086	12.405812
2	2408	11.385404	11.706111
3	1725	11.312135	11.660924
4	1489	11.330409	11.687392
5	1626	11.339185	11.6762492

表 2.展示了正则化因子对于模型的影响

正则化因子	迭代停止次数	训练误差	测试误差
0.1	1725	11.312135	11.660924
0.2	1653	12.289302	11.663418
0.3	1589	13.255432	11.674503
0.4	1530	14.211660	11.693021
0.5	1476	15.158704	11.717761

4 总结

本文对回归模型的思想，在 NACA0012 airfoils 数据集上实现了回归模型，通过选择不同的模型维度、正则化因子以及学习率，对比训练误差与测试误差的变化，从而发现参数的选择对于模型的性能至关重要。