基于决策树模型和神经网络模型的降雨量问题研究

李沐阳　　钟绍恒　　易领程

2024 年 5 月 9 日

基于决策树模型和神经网络模型的降雨量问题研究

论文标题: 基于决策树模型和神经网络模型的降雨量问题研究

代码链接: https://github.com/limuy2022/math_model

发表年份: 2024

作者信息: 李沐阳 [1] , 钟绍恒 [2] , 易领程 [3]

1. 东莞市东华高级中学 120 班学生
2. 东莞市东华高级中学 120 班学生
3. 东莞市东华高级中学 120 班学生

目录

研究背景

研究背景

现如今, 气候无不影响着人类的生活, 探寻其中各种因素的关系成为了当务之急. 在此前提下, 我们决定着手降水量的研究, 试图为气象研究提供参考. 考虑到现实因素的复杂性, 我们决定简化问题, 将其转化为 5 个自变量和 1 个因变量之间的函数关系. 本篇论文主要研究欧洲 “降水量 “与 “气温 “、“海平面气压 “、“风速 “、“湿度 “、“云层覆盖 “之间的关系. 同时, 为了方便表达, 我们规定了一下符号以及其中的单位, 如下表所示:

符号说明

表 1: 符号说明

符号	说明	单位
r	降雨量	mm
t	温度	0.1°C
f	风速	$0.1\,\mathrm{m\,s^{-1}}$
h	湿度	0.1%
c	云层覆盖	octas
p	气压	0.1hPa

前提假设

前提假设

- 排除一切人为影响气候因素，如工业排放, 热岛效应等.
- 排除次要因素对降水量的影响, 如辐射、空气污染等.
- 假设降水量只与气温、气压、风速、湿度、云层覆盖有关.
- 降水量准确的标准为：得出的降水量 r 和正确的降水量 r_0 满足关系 $r_0 - 10 \leq r \leq r_0 + 10$.

目录

线性回归模型

模型假设

- 降水量与气温、气压、风速、湿度、云层覆盖之间存在线性关系.
- 线性关系式导致的误差可忽略.

研究方法

考虑到拟合线性关系, 我们决定采用较为常见的最小二乘法进行模型拟合.

最小二乘法简介

基本概念

最小二乘法是一种求解线性方程组的方法, 该方法的基本思想是将线性方程组表示为如下形式的最小二乘方程组:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{1}$$

其中 x_1、x_2、x_3、x_4、x_5 分别代表云层覆盖、湿度、气温、气压、风速.

求解公式

最小二乘法的求解公式如下所示:

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x_1} - \beta_2 \bar{x_2} - \beta_3 \bar{x_3} - \beta_4 \bar{x_4} - \beta_5 \bar{x_5} \\ \beta_1 = \frac{\sum_{i=1}^{n}(x_{1_i} - \bar{x_1})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1_i} - \bar{x_1})^2} \\ \beta_2 = \frac{\sum_{i=1}^{n}(x_{2_i} - \bar{x_2})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{2_i} - \bar{x_2})^2} \\ \beta_3 = \frac{\sum_{i=1}^{n}(x_{3_i} - \bar{x_3})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{3_i} - \bar{x_3})^2} \\ \beta_4 = \frac{\sum_{i=1}^{n}(x_{4_i} - \bar{x_4})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{4_i} - \bar{x_4})^2} \\ \beta_5 = \frac{\sum_{i=1}^{n}(x_{5_i} - \bar{x_5})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{5_i} - \bar{x_5})^2} \end{cases} \tag{2}$$

其中 $\bar{y}$、$\bar{x_1}$、$\bar{x_2}$、$\bar{x_3}$、$\bar{x_4}$、$\bar{x_5}$ 分别代表降水量的均值、云层覆盖的均值、湿度的均值、气温的均值、气压的均值、风速的均值.

最小二乘法结果

结果展示

最小二乘法的求解结果如下所示：

$$\begin{cases} \beta_0 = 315.80363888548624 \\ \beta_1 = 6.197181723408759 \\ \beta_2 = 0.15411116711110595 \\ \beta_3 = 0.16901104754961455 \\ \beta_4 = -0.03432027705042855 \\ \beta_5 = 0.17773547386441546 \end{cases} \tag{3}$$

模型误差

正确率为 60% 左右, 最小二乘法的拟合图像如下所示:

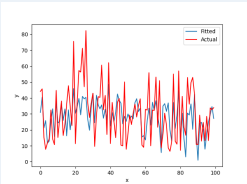

图 1: 线性回归模型拟合降水量和真实降水量的折线统计图

神经网络

网络规模

本网络采用的规模为 $5 \times 64 \times 64 \times 64 \times 64 \times 1$ 的全连接神经网络 [1]. 其中, 第 1 层为输入的五个自变量, 第 6 层的值为降水量的预测值. 神经元采用的激活函数为 Leaky ReLU 函数 [2].

1. 全连接神经网络 (Fully Connected Neural Network, 简称 FCNN) 是一种最基础的人工神经网络结构, 也称为多层感知器 (Multilayer Perceptron,MLP).
2. 函数表达式为: $\mathrm{f(x)} = \max(0, \mathrm{x}) + \alpha \cdot \min(0, \mathrm{x})$.

训练方法

本网络采用较为普遍的梯度下降 [1] 和反向传播算法 [2] 进行训练.

1. 梯度下降, 英文名 Gradient Descent, 是一种用来最小化函数的优化算法. 它的工作原理就像一个探险家在山上寻找最低点. 探险家在每一步都会选择下坡的方向, 直到他找到一个地方, 无论向哪个方向走, 都是上坡, 那么他就知道他已经找到了最低点.
2. 反向传播, 英文名 Backpropagation, 是一种在神经网络中用来调整权重和偏置的方法. 它的工作原理就像一个电影导演, 指导演员们 (神经元) 如何更好地表演 (调整权重和偏置), 以达到最好的观众反馈 (最小化损失函数)

神经网络结果

模型误差

正确率为 65% 左右, 神经网络的拟合图像如下所示:

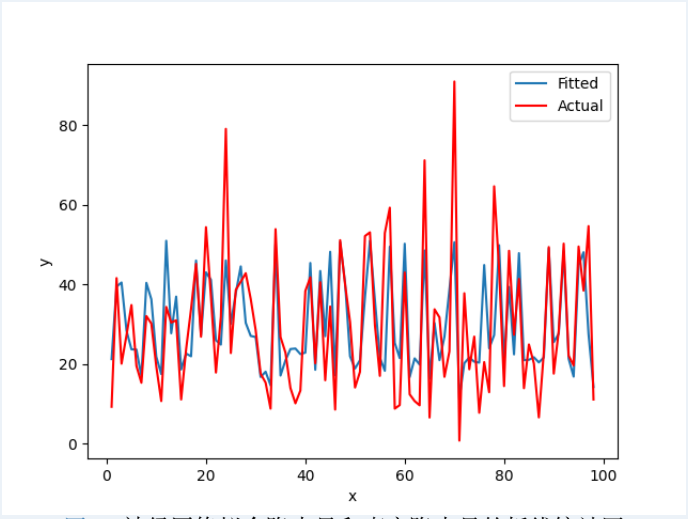

图 2: 神经网络拟合降水量和真实降水量的折线统计图

决策树

决策树基本介绍

决策树是一种用于分类和回归的非监督学习算法. 它的基本思想是将数据集划分为若干个子集，每个子集都是一个独立的决策树. 通过递归地构造决策树，可以得到一颗最优的决策树.

CART 算法基本原理

- 每个节点都具有一个决策规则.
- 每一层根据不同标签 (即自变量) 计算基尼指数, 公式如下:

$$\mathrm{Gini(D)} = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p(x_i)}^2 \tag{4}$$

  其中,$\mathrm{p(x_i)}$ 是 D 中各个类别的样本数占总样本数的比例.
- 根据基尼指数的大小，决策树会选择具有最小基尼指数的节点.

决策树结果

模型误差

正确率为 90% 左右，决策树的拟合图像如下所示:

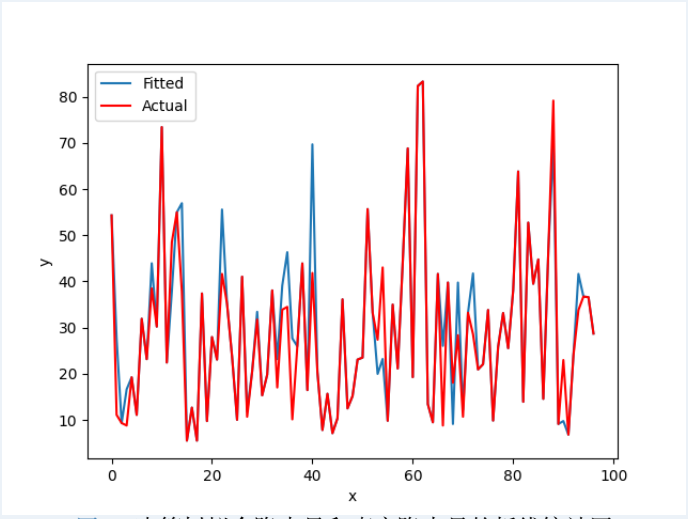

图 3: 决策树拟合降水量和真实降水量的折线统计图

目录

线性回归模型

线性回归优缺点与改进方法

基于线性回归模型在仅研究几个变量的时候准确度和神经网络回归模型相差无几，具有操作十分简单，预测能力也并不算很弱的特点，改进方法也许可以采用并非完全线性的回归模型，对于某些变量进行更加具体的分析，采用其他的基本函数进行描述效果可能会好一点.

神经网络回归模型

神经网络回归优缺点与改进方法

基于神经网络的模型具有普适性更加强的特点，在降雨量受到如此多方面的影响的情况下对其中的几个关键变量进行研究能达到还不错的效果，实践起来也不算特别麻烦.
该方法可以利用神经网络强大的调整和适应能力用来研究各个地区的局部降雨量. 不足之处在于，若是把时间范围拉得很大，会导致模型由于气候变化和人类活动导致不准确性增加，而且模型也并没有针对季节等进行描述.
可以从两方面考虑增强方法，一是结合一定的时间序列模型来解决气候变化和人类活动的问题，因为这两个因素也是按照一定的规律变化的，神经网络可以学习出其中的趋势; 二是考虑季节性的特点，一定程度上可以使用有季节性的时间序列模型来缓解. 但还有一种思路就是添加更多的自变量来间接反应季节的影响，因为气象总体是呈周期性的，例如可以添加风向作为自变量，来间接的反映季风，进而反映出季节的变化，感觉这种思路可能会合适一点.

决策树模型

决策树优缺点与改进方法

由于决策树模型具有可以随意调整叶子节点和非叶子节点的分界点，从而使数据尽可能的良好匹配所有的数据的特性和天气取值范围并不是很大的特点，使得它对于气候这种并非不具有显著函数关系的问题拟合程度非常高，几乎可以说是十分精确的预测了.

另一方面，对于极端情况的预测效果一定是很差的，因为数据被设计的尽量靠近正常的数据了，就会使得偏离不常出现的值，这也体现了决策树模型所具有的对于缺失值得处理功能非常有限的缺点.

但是，若是我们把前两个模型综合起来看，使用线性回归模型和神经网络回归模型生成极端数据，再交由决策树模型拟合，也许就可以在一定程度上缓解这个问题，三个模型可以相辅相成，决策树模型也可以反过来帮助神经网络模型缓解过拟合的问题.

END

END