基于决策树模型和神经网络模型的降雨量问题研究

三号作品

李沐阳　钟绍恒　易领程

2024 年 5 月 25 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

目录

研究背景

研究背景

现如今, 气候无时无刻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. 将关系简化为 “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.
- 假设降水量只与五个自变量有关.
- 根据气象学的要求，定义降水量预测结果准确的标准为：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

搜集数据及数据整理

数据网站

我们通过 github 上的 awesome-dataset 项目，找到了如下的欧洲气候网站，并从中获取到了降雨量、气温、气压、风速、湿度和云层覆盖的数据.

数据整理方法

为了选取完整、连续、准确的数据，我们抛弃了最近的以及最早的数据，只选取了 1877-2000 年的数据进行分析. 因为以天为单位的数据方差过大, 我们决定将数据以一个月为单位取平均值, 使用的工具为 python 的 pandas.

数据可视化

数据折线图

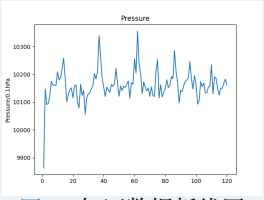

图 1: 气压数据折线图

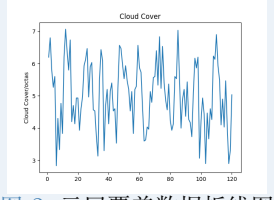

图 2: 云层覆盖数据折线图

数据可视化

数据折线图

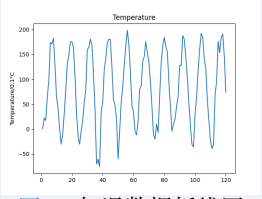

图 3: 气温数据折线图

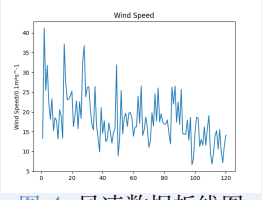

图 4: 风速数据折线图

数据可视化

数据折线图

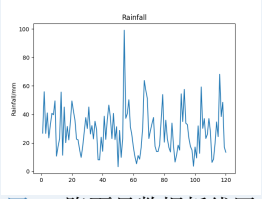

图 5: 降雨量数据折线图

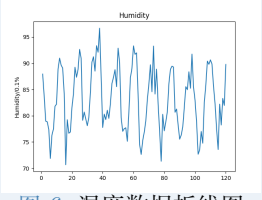

图 6: 湿度数据折线图

模型尝试与论文撰写

尝试模型

- 线性回归模型. 准确率 50% 左右, 不够理想.
- 神经网络. 相较于线性回归模型，准确率更高，但仍然未达到 80%.
- 决策树. 准确率达到 90%. 可用于该模型的精确研究.

论文撰写

我们采用 latex 作为论文撰写的工具,python 生成数据图片,c++ 和 python 用于代码编写, github 存放代码共同整合出了这篇论文.

目录

线性回归模型

研究方法

观察数据发现，降水量与气温、气压、风速、湿度、云层覆盖之间可能存在线性关系. 尝试对该问题建立多元线性回归模型. 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. 为了方便起见，我们使用了 python 的 statsmodels 库.

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:

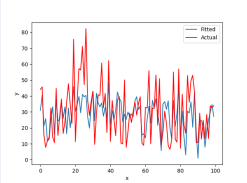

图 7: 线性回归模型拟合降水量和真实降水量的折线统计图

尝试使用神经网络的原因

研究过程

神经网络回归模型作为非参数非线性的模型，拥有强大的自我调整能力，可以在一定程度上很好的拟合数据，并且响应变量的整个条件分布. 经过最初的尝试和对数据特点的分析后，我们使用了最基础的全连接神经网络，首先确定了 6 个神经元层，分别是 5,64,64,64,64,1，若神经元数量再多过拟合的风险很大，若过少也很容易造成对训练数据和测试数据正确率都降低的情况，这样的数据是比较合理的. 训练方法是传统的梯度下降反向传播算法. 在对神经网络激活函数的选择中，我们最初选择了 Sigmoid，后来经过讨论，认为该二分类型激活函数不适合该背景，改为了 Leaky ReLU 函数，大大提升了最初的训练速度. 首先将数据划分为 80% 的训练集和 20% 的测试集，估价函数的设计是简单地在每次梯度下降后对于测试集进行计算正确率.

网络规模

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. Leaky ReLU 函数表达式为: $\mathrm{f(x)} = \max(0, \mathrm{x}) + \alpha \cdot \min(0, \mathrm{x})$, 其中 $0 < \alpha \ll 1$.

神经网络

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. 为了方便起见，我们使用了 python 的 tensorflow 框架进行神经网络的简单设置.

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:

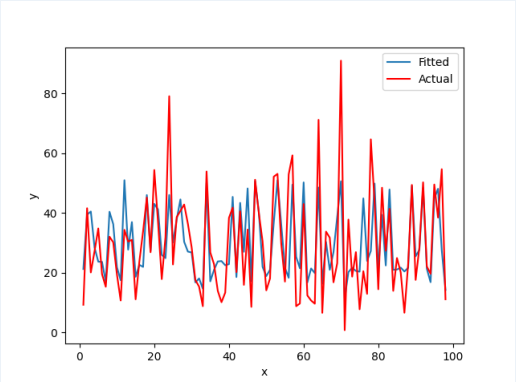

图 8: 神经网络拟合降水量和真实降水量的折线统计图

模型结果

非常明显，相比与之前的线性回归模型，神经网络回归模型的拟合准确度明显大大增强，但是对于某些特别极端的值似乎依旧有一些缺失，不过总体拟合的还算准确.

遗传算法训练神经网络尝试

模型误差

猜测用遗传算法训练网络会有更好的结果，我们也展开了尝试，但经过切换各种激活函数和初始化方法，发现网络得到的值方差较小，有靠拢趋势，在经过 1000 代后，所有的个体的适应度都逐步收敛到一个值上，同时所有个体的策略都是无论输入，输出都是相同的值，并不符合现实情况，对现实指导的意义不大. 针对这种情况，我们猜想是因为遗传算法对局部搜索能力不够强大，导致了网络的学习效率低下. 同时我们采用的算法也不太适合用于神经网络模型训练，我们只将其作为一种特殊的模型进行尝试，提供新的思路.

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.

尝试使用决策树模型的原因

从理论上来看，决策树模型通过连续的逻辑判断树来预测，适合用于数据非线性或数学关系不明显的情况，但数据间并非存在明显的非线性关系，这样的问题应当不太适合使用决策树模型刻画，但是对于降水量这样取值范围较为固定和小，呈现周期性，且同时具有自变量之间相互影响特性的问题，用决策树模型可能也会有特殊的优异效果

决策树

实际训练方法

实际上采用 tensorflow 进行离散决策树模型的简单设置

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. 基尼系数代表了模型的不纯度，基尼系数越小，则不纯度越小。基尼系数反映了从数据集合中随机选取两个样本，其类别不一样的概率，因此，基尼系数越小代表数据集越纯

- 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:

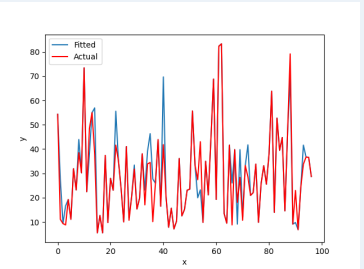

图 9: 决策树拟合降水量和真实降水量的折线统计图

决策树模型总结

模型总结

由于我们的数据范围较大，决策树模型的训练算法会尽量在较大尺度上寻找数据的普遍规律，具有一定的准确度，决策树模型通过对于给定数据的每一个取值都产生一个相应的分支来拟合数据. 虽然数据中明显存在属性的交叉，但是由于数据范围比较大，统计特征是充分的，所以大致上不会影响决策树的准确性和可理解性.

仔细分析，可以得知也许是因为欧洲气象较为平均，没那么极端，决策树模型虽然尝试在较大的尺度上寻找数据普遍的分类依据, 究其本质还是一种另类的分类讨论，若是在气象多变的广东地区，数据量相同的情况下，效果也许反而不如神经网络模型. 神经网络模型和线性回归模型从本质上探索他们之间的数量关系，但决策树模型只是分类讨论罢了，对于出现次数少的预测能力不足，但是实践证明，也是一种非常切实可行的方案了.

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.

目录

模型检验

模型检验

我们用 1、5、17 号气象站的数据测试了线性回归模型和神经网络模型, 由于考虑不周全，也因经验不足，所以我们并没有测试决策树模型，这有待后人进一步研究

目录

意义

研究价值

我们今天在这里仅仅是抛砖引玉, 展现了神经网络以及决策树在该问题下蕴藏的巨大潜力. 在全球气候越来越极端、环境越来越恶化的情形下, 研究气候因素之间的关系尤为重要. 可以为我们寻求可持续发展道路指明方向, 照亮未来. 我们学习数学建模不仅仅是学习一门科学, 更是让我们新时代青年为新时代贡献力量.

END

END