机器学习大作业展示

基于一系列监督学习的电力部门的二氧化碳排放量回归预测

李远铄 (19211332), 李潮乐 (20281140), 赵东阳 (20281029)

北京交通大学计算机与信息技术学院

2022 年 12 月 15 日

数据集基本信息和目标

数据分布 数据总共有 9 个碳排放的指标，其中 7 个指标有 1973 年 1 月起，一直到 2021 年 12 月的所有月份的数据；剩下两个指标则有 1989 年 1 月起，一直到 2021 年 12 月的所有月份的数据。

数据量 数据量不大，不应该选择复杂的模型。

目标 使用机器学习方法，预测出 2022 年 1 到 6 月这 9 个指标的值。

数据集的进一步分析-周期性

分析 我们使用自相关系数来判定指标是否有 12 这个周期。

- 对于每一个指标 y 年 m 月的数据，我们都将它除以这个指标 y 年的数据和。这样可以一定程度上排除趋势性影响。
- 在经过上述处理后，计算这些指标的数据序列在 12 这个间隔下的自相关性。程序计算后，结果如下：

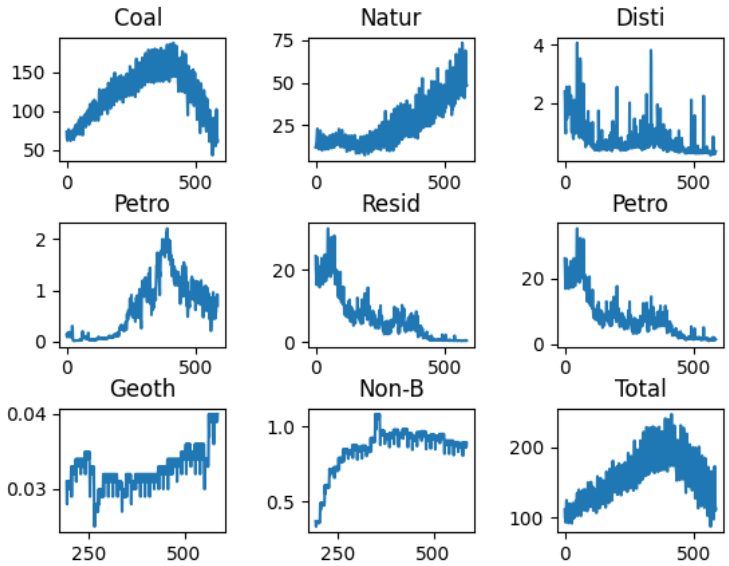
指标名（前 5 字符简写）	间隔为 12 的自相关性
Coal	0.9202
Natur	0.9279
Disti	0.8330
Petro	0.9421
Resid	0.9204
Petro	0.9128
Geoth	0.6345
Non-B	0.7385
Total	0.9369

结论 大多数指标周期性显然，因此我们需要考虑周期性对时间序列分析的影响。

数据集的进一步分析-相关性

分析 我们先可视化一下：

图示



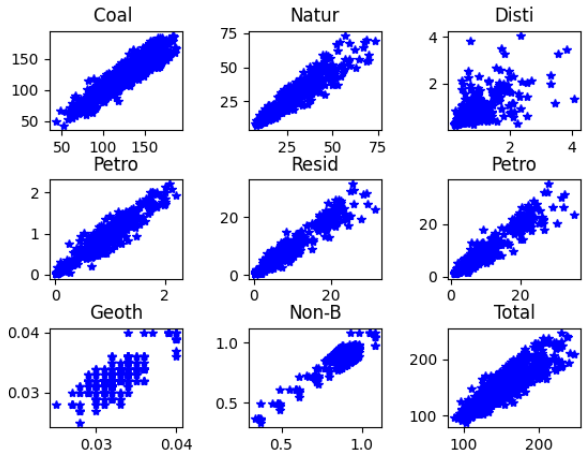
初步结论 看起来少部分数据有相关性。

进一步 实际上分析相关性后，发现部分数据相关性确实很高。比如"Resid" 和"Petro" 的相关系数达到了 0.9951。

结论 我们需要考虑相关性对我们建模的影响。

很有意思的性质-前一月与后一月关系为线性

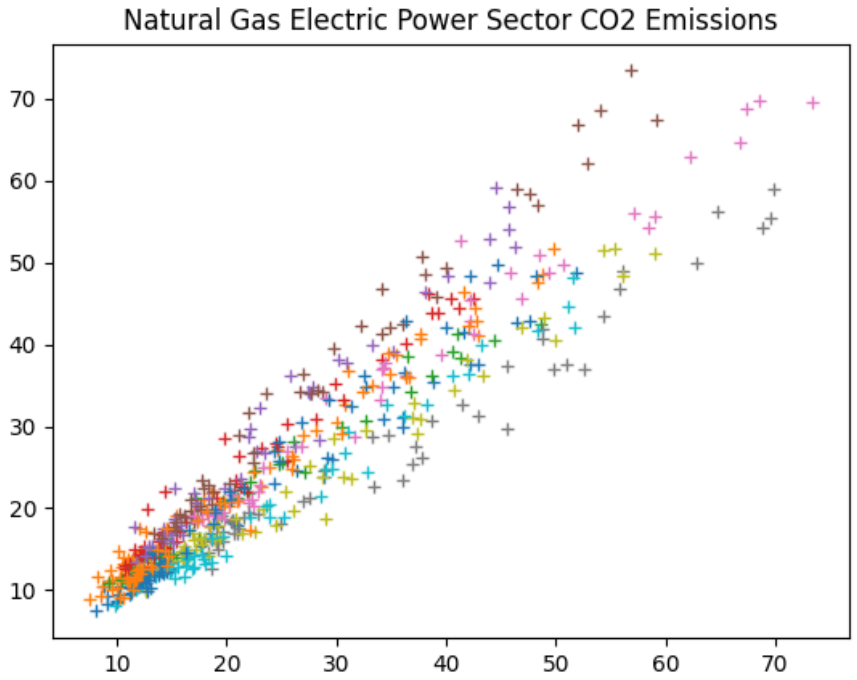
- 一般会将时间序列分析问题转化成监督学习问题，即用前一(几)月的数据预测后一(几)月的数据。
- 所以我们想看看“一个月数据与下一个月数据的关系”。
- 直接上图：



- 令人惊讶的发现：无论哪个指标“一个月数据与下一个月数据”都有不小的线性关系！
- 我们称这个关系为“隔月线性关系”。

“隔月线性关系”与哪个月也有关

- 在“隔月线性关系”进一步看，“这个月数据与下一个月数据”的具体线性关系也和“这个月”是哪个月有关。
- 以"Natur" 为例，按照“这个月”涂色后，它的“这个月数据与下一个月数据”关系图如下，可以看出不同颜色的点在不同线上。



“隔月线性关系”的启发

- 以“时间序列预测转监督学习”为基础，在“隔月线性关系”的启发下，我们可以得到这么一个想法：
  根据输入的月份区分不同的模型，多个模型结合起来做预测。
- 这么做的原因：
  1. “隔月线性关系”表明，“时间序列转监督学习”这个思路很可能有效。
  2. “‘隔月线性关系’与哪个月也有关”表明，不同的月份“监督学习”的模型也需要不同。
  3. 实际上，这种预测方式在原理上就利用了“周期性”，同时也兼顾了“趋势性”

整体思路

- 前面提到，我们将时间序列模型转化成监督学习问题来解决。目前我们不考虑多个指标一起预测（后续会再分析这会事）。
- 因为目标为预测 2022 年 1 月到 6 月的数据，所以我们的模型要能预测后 6 个月的数据。
- 我们会训练一系列“‘前一或几个月数据’作为输入，输出‘后一或几个月的数据’”的模型来做这个预测。
- 我们称这一系列模型为 X2Y 模型，比如：
  - 一个模型接受每年 12 月的数据，输出第二年 1 月的数据，那么就称它为 **O2O** 模型，O 表示 One。
  - 一个模型接受每年 12 月、次年 1 月的数据，输出次年 2 月的数据，那么就称它为 **M2O** 模型，且称 M=2，O 表示 One。
  - 一个模型接受每年 12 月的数据，输出次年 1-6 月的数据，那么就称它为 **O2M** 模型，且称 M=6，O 表示 One。
  - 一个模型接受每年 11,12 月的数据，输出次年 1-6 月的数据，那么就称它为 **M2M** 模型，且称两个 M 分别为 2,6。

滚动预测和分别预测

滚动预测

- 对于 X2O 模型（即 O2O 或者 M2O 模型），我们需要滚动预测。
- 比如我们只有 M=6 的 M2O 模型，那么我们就需要：
  1. 先用 2021 年 11,12 月的数据预测 2022 年 1 月的数据。
  2. 再用 2021 年 12 月的数据和第一步预测出的 2022 年 1 月的数据预测 2022 年 2 月的数据。
  3. 再用第一步预测出的 2022 年 1 月的数据、第二步预测出的 2022 年 2 月的数据，来预测 2022 年 3 月的数据。
  4. 以此类推。
- 注意，这里几步用到的模型可以是一个也可以不是一个。

分别预测

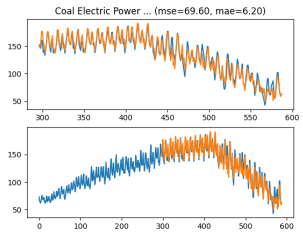
- 对于 X2M 模型（即 O2M 或者 M2M 模型），我们只需要一系列模型做预测就行。
- 比如我们只有 M=2 的 O2M 模型，那么我们只需要使用 2021 年 12 月的数据，使用一系列模型预测 2022 年 1-6 月的数据就行。

初选择

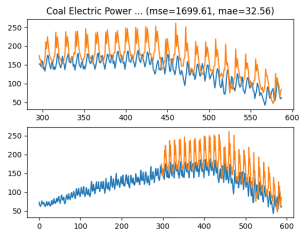
出于“分别预测 3 月和 2 月的关系、2 月和 1 月的关系比直接预测 3 月和 1 月的关系好”的观念，先实现 X2O 模型。

大 失 败

- 因为“隔月线性关系”，我们一开始就想到了使用滚动预测的方式，但是我们发现，无论是 O2O 还是 M2O：
  1. 只往后预测一两个月的效果还行。(Linear SVM，M=6)



  2. 在往后预测六个月的时候效果就不好了。(Linear SVM，M=6)



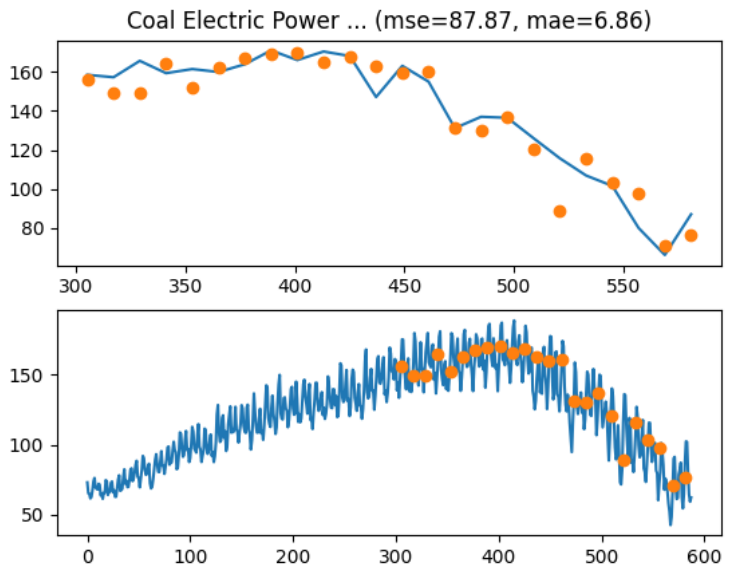
- 而且，这个结果还是在我用全部数据作为训练集，然后抽后 50% 的数据作为测试集的情况下的结果。。。

失败原因分析

- 实际上这里的原因在于：滚动预测，不仅预测值在滚，误差也在滚，而且越滚越大。
- 并且这个问题不能通过改进模型复杂度（以获得更精细的模型）解决。。。这个问题的数据量压根不支持这么做。
  （在分月预测后，每个子模型只有 50 条数据。）
- 怎么办？
  - 我们一开始使用 X2O 的原因是：分别预测 3 月和 2 月的关系、2 月和 1 月的关系比直接预测 3 月和 1 月的关系好。
  - 但是，真的吗？
  - 由于“隔月线性关系”，所以其实也有“隔两月线性关系”、“隔三月线性关系”？
  - 所以其实没必要用滚动预测白白承担“误差滚动”的影响！

大 成 功

- 前面提到，我们实验 X2O 时候，“用全部数据作为训练集，然后抽后 50% 的数据作为测试集”。
- X2O 模型在这种训练集、测试集划分下，往后预测六个月的时候效果还是差的离谱，那么 X2M 模型呢？
- Linear SVM, O2M, for July:



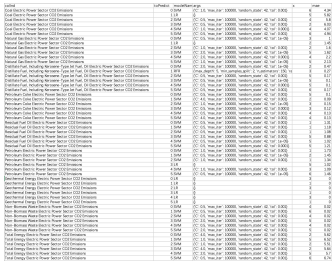
- 差 距 明 显。

模型选择

- 在确定使用 X2M 之后，剩下的就是找合适的模型和参数了。
- 这里，对于每一个指标、1-6 月每一个月，我们都有这些需要选择的参数/模型：
  1. X2M 中的 X 是多少：这里 M 确定是 6（因为要求预测六个月的），但是 X 没有确定。
     比如如果 X 是 5，就表示对于一个指标，最后我们有某个模型，输入是 2021 年 8 月到 12 月的数据，获得 2022 年 1-6 月数据。
     注意，对于相同指标、不同月份的预测，其 X2M 的 X 是可以不同的。
  2. 模型用什么。我们只考虑 Linear SVM、Linear Regression、Random Forest Regression。
  3. 模型参数是什么。
- 我们实现进阶的网格搜索，获得了一系列模型。

模型结果

- 最终结果：



- 预测出来的值是：


```
{
    "Coal Electric Power Sector CO2 Emissions": [
        58.28242855091633, 59.28508067157844, 44.37397442507746, 58.57187792458451, 60.022288864662016, 73.42881691222624
    ],
    "Natural Gas Electric Power Sector CO2 Emissions": [
        48.60946758728248, 47.64954341605172, 50.8750863718924 75, 42.46443608336682, 54.3519791197608, 59.53495676816743
    ],
    "Distillate Fuel, Including Kerosene-Type Jet Fuel, Oil Electric Power Sector CO2 Emissions": [
        0.3978112148583304, 0.6714608888888892, 0.2323169098055728 5, 0.2633059381516299 5, 0.3134267798480481, 0.3009595510860905
    ],
    "Petroleum Coke Electric Power Sector CO2 Emissions": [
        0.8883229494596206, 0.8823398908175851, 0.7902402162365672, 0.7434701609195875, 0.8122806719087994, 0.882619934720472
    ],
    "Residual Fuel Oil Electric Power Sector CO2 Emissions": [
        0.6250397096995013, 0.3589187496529193, 0.2931985881145331, 0.2795968445016286, 0.3651154636481591 7, 0.3927491072831417
    ],
    "Petroleum Electric Power Sector CO2 Emissions": [
        2.060768494488162, 1.614568454188619, 1.5970329424307663, 1.1646791833449617, 1.6344255485315895, 1.3064157550832323
    ],
    "Geothermal Energy Electric Power Sector CO2 Emissions": [
        0.04041857923497266, 0.03446357878068092, 0.04041857923497266, 0.0369034045922407, 0.04041857923497266, 0.0369034045922407
    ],
    "Non-Biomass Waste Electric Power Sector CO2 Emissions": [
        0.8957012645802837, 0.8135480570232533, 0.8957012645802837, 0.8668912356137429, 0.8957012645802837, 0.8668912356137429
    ],
    "Total Energy Electric Power Sector CO2 Emissions": [
        123.79468662508768, 103.61823622351255, 117.61217116264152, 107.1165693342472, 108.24285791180694, 123.67079114405452
    ]
}
```


- 结果会附在代码中～

分析

- 前面提到，我们目前都是将所有指标单独预测，但是是不是说我们可以将多个指标放在一起预测呢？
- 对于相关性强的指标：

原理分析 我们前面的实验结果其实可以看出，这里线性 SVM 和线性回归是比较适合的选择，这种情况下，相关性强的指标放在一起预测只会带来干扰。
毕竟你已经是线性关系了，何必再线性一次呢。

实验分析 不过我们还是补了实验：即使是相关性最高的"Resid"和"Petro" 做网格搜索，maes 也不如我们前面得到的模型。

- 对于相关性不强的指标：

原理分析 这里我想不出原理上反对" 相关性不强的指标做联合预测"的理由。

实验分析 不过还是做了实验，同样，对于相关性不高的模型，做网格搜索，mae 不如我们前面得到的模型。

- 注：这里联合预测的指标都做了归一化。

数据处理和可解释性

数据处理 除了相关性分析那部分的模型外，我们没有对数据做预处理。分别讨论：

1. 异常数据剔除：对于气象数据这种经过确认、科学获得的数据，我觉得剔除反而是不当的。
2. 数据增强：因为“隔月线性关系”，对于我们的预测方式来说，插值类的数据增强看起来不会对训练过程造成决定性影响。

可解释性 虽然使用的备选模型 (Linear SVR, Linear Regression, Random Forest Regression) 中，随机森林的可解释性不太好，但是最终选择的模型中只有一个是随机森林，所以结果整体上说可解释性很高。

趣事

- 我介绍的时候是按照我整理过后的思路介绍的，但是其实我们做的时候是完全不同的节奏。
- 我们最初的进展来自于 bug：
  - 在一开始，我抱着死马当作活马医的心态，写了一个 O2O 代码，然后发现效果巨好，觉得不对，“怎么会这么好呢？”。
  - 隧认为数据集一定是有“玄机”的，于是才发现了“隔月线性关系”这个至关重要的性质，并基于此提出 X2Y 模型。
  - 但是比较搞的是，实际上当时写的代码是有 bug 的，效果好只是那个 bug 会让效果看着很好。
  - 不过最后结果上来说，这个 bug 推进了我们的工作（
- 我们一开始确实不觉得 X2M 模型比 X2O 模型好，所以当我们发现即使是 O2M 都比 M2O 效果好时，有点绷不住。

分工

- 李远铄：
  - 整体程序框架构思。
  - 完成基础模型构思。
  - 完成分析“隔月线性关系”的代码编写。
  - 参与分析数据相关性、周期性代码编写。
  - O2O 模型、M2M 模型的编写。
  - 可视化代码编写。
- 李潮乐：
  - 提供数据相关性、周期性分析思路。
  - 参与分析数据相关性、周期性代码编写。
  - M2O 模型的编写。
  - 利用相关性联合预测的模型的编写。
- 赵东阳：
  - O2M 模型的编写。
  - 利用相关性联合预测的模型的编写。

Thanks!