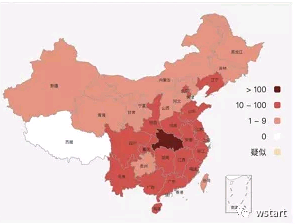
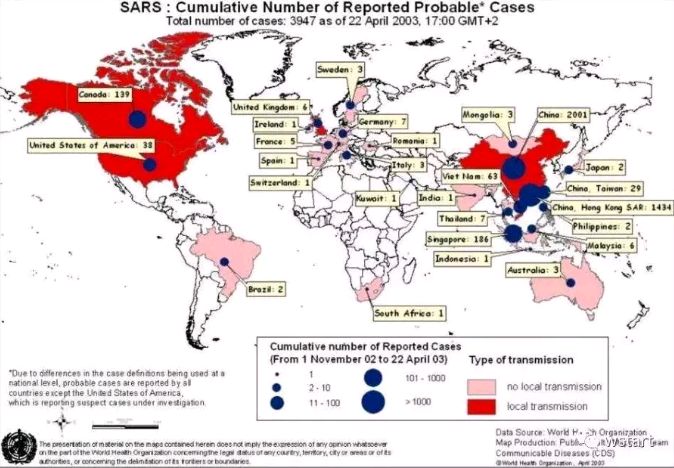
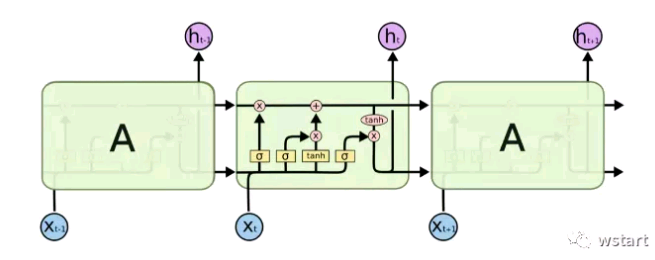
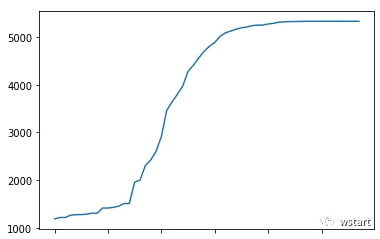
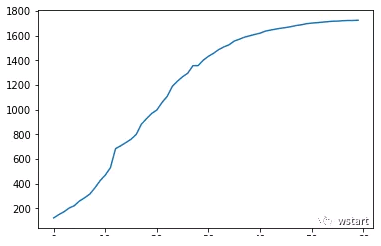
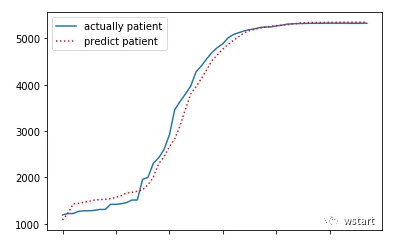
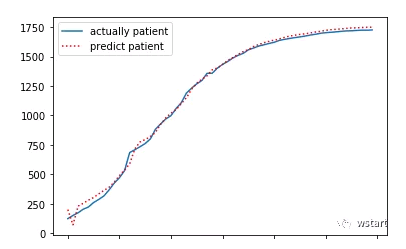
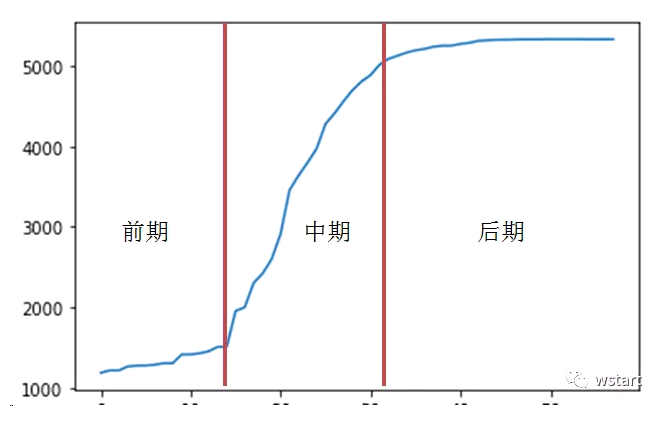
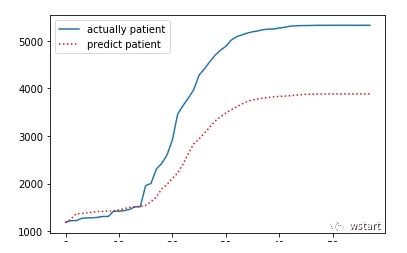
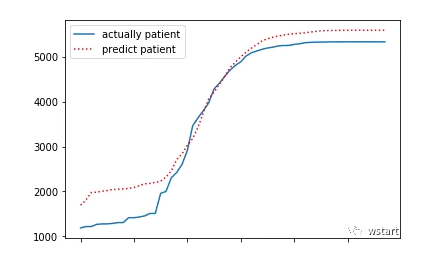
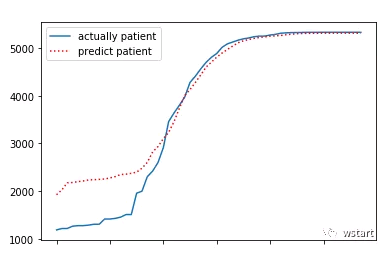
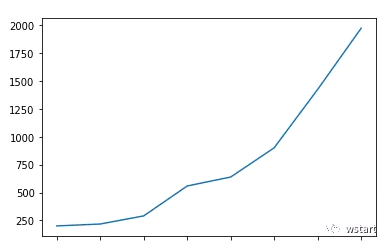
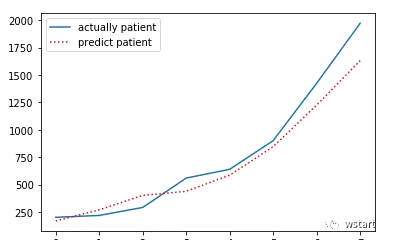
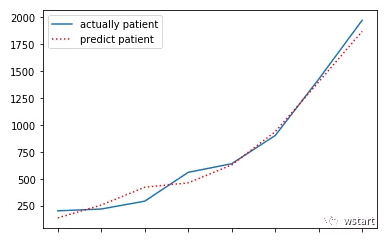
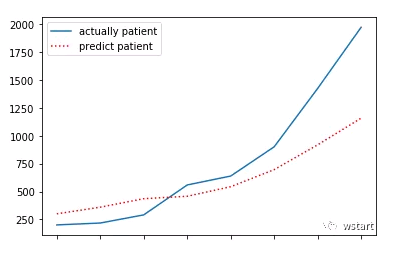
机器学习相信我们很多人都接触过，但是很多大牛利用机器学习来预测新型冠状病毒，当然无论对错，其实这正是我们学习技术的价值，让我们的知识能够应用于生活，对我们的身边的事情能够起一个指导作用。如果大家感兴趣，想练手，其实自己也可以实际尝试下。下面是是大牛的实验。  
目的：  
本文探讨的主要目的是，有没有一种算法可以根据已有的数据对疫情进行预测和评估，可以方便在疫情出现的时候，根据预测的数据来判断当前的疫情状况。  
文章分为 5 段

* 猜想
* 数据准备
* 猜想验证
* 新型冠状病毒确诊数据的预测和评估
* 总结  
  Ps : 心急的朋友可以直接拖到末尾 看结论。  
    
    
  0x01 猜想  
  由于当前新型冠状病毒的数据并不多（从 11 号-26 号），所以仅仅是新型冠状病毒的数据来做研究是不够的，必须还要一组参考验证组。（当然是越多越好，但是精力有限，所以只找了一组）。  
  首先明确一下判断疫情的严重情况的指标，目前疫情一般会公告这四个数据：
* 确诊病例
* 疑似的病例
* 感染的地区
* 死亡的人数  
  由于我们把地区限定在了中国（虽然各个地区不一样），所这个地区的指标就不用了，死亡人数在疫情的前中后期都不太一样，这个也不用。疑似病例这个和各地的收诊情况有关系，比如香港的疑似病例就非常多，确诊很少，所以这个指标也不采用，剩下就以确诊病例作为唯一的参考指标。  
  猜想 1  
  假设病毒的传播和地区，气候以及病毒自身的特征有非常大的关系，那么只要找到一个和新型冠状病毒这三个属性非常相似的病毒，只要算法能够对这种病毒的确诊病例进行预估和判断，那么算法也可以对新型冠状病毒进行预估和判断。  
    
  （ 03 年 4 月 22 日 SRAS，图源..不太清楚）  
  目前和新型冠状病毒最像的应该就是 SARS：气候和现在差不多，1 月和 3 月。地区主要是是中国。病毒类型：都是冠状病毒，都有发烧咳嗽等特征。  
  那么基于上述的条件，采用了 SARS 作为参考组。  
  猜想 2：  
  可以给定一个 Δ,
* 如果 当前确诊的病例 - 预测的确诊病例的比率 >Δ，那么当前的疫情变异了，进一步扩大影响的范围。
* 如果 当前确诊的病例 - 预测的确诊病例的比率 < -Δ，那么当前的疫情已经基本被控制，已经减弱了。
* 如果 -Δ < 当前确诊的病例 - 预测的确诊病例的比率 < Δ，那么当前的疫情还在平稳的发展中，没有进一步变异。

0x02 数据的准备  
确诊病例的数据

* SARS 的数据来源于 www.who.int , 数据从 03 年的 3 月 17 日-5 月 30 日 。
* 新型冠状病毒的数据来源于腾讯新闻和国家卫生健康委员会 http://www.nhc.gov.cn 数据从 20 年的 1 月 15 日到 1 月 26 日。【下载更多数据链接https://www.who.int/data/gho】  
  如果可以拿到 SARS 在 1 月份的数据就好了，可惜没有一个比较完整的地方有记录，基本上卫生部发布的好像是从四月份开始。于是只能采用四月份的数据了。  
  评估的算法  
  采用了 LSTM(长短期记忆网络)，搭建的话 Keras。LSTM 已经是个老算法了，经常用于量化投资(简单点来说就是炒股之类，预测股票的增长)，非常擅长于时间序列的预测。最终调教后的算法模型是，根据过去三天的数据，可以预估最新一天的数据。  
    
  LSTM  
  实验的环境  
  数据量比较少，直接租了一台 16 核 8G 的云主机进行测试。  
  0x03 猜想验证  
  猜想验证部分直接全部使用 WHO 的 SARS 的数据进行验证。采用的字段是 Cumulative number of case(s)，累计病例数，我这里就当它是确诊人数了。  
  SARS 的数据分别有两部分，一部分是中国大陆的，一部分是中国香港的。在 5 月份后基本已经趋于平缓，所以只取了前两个月（约 60 天左右的数据）。  
    
  中国大陆确诊人数变化曲线  
    
  中国香港确诊人数变化曲线  
  可以看到数据涵盖了病毒爆发期，病毒平稳增长期，以及病毒被消灭期，比较合适用来验证算法是否可以拟合整体的曲线。  
  0x031 猜想一  
  这里采用前 50 天的数据做训练样本，训练的结果用来预测 60 天的情况, 预测和实际的对比如下：  
    
  中国大陆 SARS 确诊人数变化曲线(预测和实际)  
    
    
    
  中国香港 SARS 确诊人数变化曲线(预测和实际)  
    
  从图中我们可以看到，在使用了 90%的数据量下，数据是拟合的比较好的，肉眼看上去香港的数据基本可以用算法完全预估。这里给一下数据误差的对比情况：
* 误差：|（实际值-预估值）/ 实际值 |
* 中国大陆：误差的平均值 0.057 ， 误差的均方差 0.060 。
* 中国香港：误差的平均值 0.053 ， 误差的均方差 0.137 。  
  说人话就是，模型预测的数据和实际的数据最多差了 6%。效果是非常理想的。这里我们可以证明第一个猜想，即 LSTM 可以用来拟合 SARS 确诊人数的变化情况。  
  0x032 猜想二  
  实际上虽然猜想一得到了证实，但是依旧没啥用，因为这里的模型是使用了 90%的数据，基本已经到了传播的后期。那么前期是否可以使用同一个模型进行验证呢？这里只取中国大陆的数据进行试验。  
    
  根据三个阶段（这三个阶段是肉眼判断截取的，是否合理当时没有考虑）截取对应的数据进行建模，建模后的预测效果如下：  
    
  前期的数据建模人数变化曲线(预测和实际)  
    
  中期的数据建模人数变化曲线(预测和实际)  
    
  后期的数据建模人数变化曲线(预测和实际)  
  在前期的数据的预测中，到后面两个阶段实际的数据都比预测的数据比要大，说明在疫情开始了爆发。  
  在中期的数据的预测中，其他两个阶段预测的数据基本都比实际的数据要大， 说明在疫情在前期的感染速度较低，后期的感染速度也变低 。  
  根据上面两个数据，是可以推测是存在一个 Δ 去判断当前的疫情状况的。但是 Δ究竟是多少呢，这里倒是可以根据好几个因素去设置：
* 斜率的改变
* 误差的改变
* 转折点的个数等  
  这里就偷懒了，可以明显看到各个时期都有一个或者多个转折点（消亡期不算），直接取最接近转折点的误差率来计算：
* 在爆发期里面，最靠近转折点的误差值是 0.21 。
* 在控制期里面，最靠近转折点的误差值是 0.14 。  
  两个取平均作为 Δ的值，也就是 0.175 。  
  0x04 新型冠状病毒确诊数据的预测和验证  
    
  （新型冠状病毒从 16 号至 26 号确诊人数变化曲线）  
  分别取前 6 天的数据（到 20 号），取前面 9 天的数据（到 25 号）的数据代入模型进行预估  
    
  到 20 号的数据模型确诊人数变化曲线（预估和实际）  
    
  到 25 号的数据模型确诊人数变化曲线（预估和实际）  
  为什么选 20 号呢，20 号主要是央视有采访钟南山大佬，而且第一次说明了人传人的存在，以及 14 名医护被感染。我觉得应该算是开始重视这个病毒的一个点，随后各地开始上报疫情，发布以及响应等。  
  再往前的时间点就数据就比较缺失了，模型精准度已经失去了使用的价值，所以采取了 20 号的这个时间点作为分隔来建造模型。  
  如果以到 20 号的模型来估算的话，第一次误差偏离 0.17 的时间是 23 号，误差为 0.19 。23 号，广东，浙江启动了一级响应。确诊的病人增多。但个人感觉应该是检测的方法变强导致的，侧面来看，模型的确检测除了异常值，不过并不能够说明病毒变强了。  
  如果以到 24 号的模型来看，预估的 25 号的数据是 1877，实际的数据 1972，误差率在 0.04 ，未到偏离值。所以根据这个模型推断，当前疫情还在持续扩散中，但是并没有进一步恶化，可能国家的高度重视和很多省份都开始参与其中有关系。  
  依照当前的模型，明天的确诊人数的预估值是 2770，加上 0.17 的误差的话
* 如果实际的确诊人数 > 3337 说明疫情加重
* 如果实际的确诊人数 < 2367 说明疫情正在被控制
* 如果处于中间，那还是依旧在扩散发展中，但并未恶化。

【以上只是根据模型得到的，并不是标准，仅是实验数据】  
再次声明：  
以上的观点和结论都属于依据当前的确诊人数所作的推断，毫无参考的意义和价值。如果数据刚好对了，不能说明什么，如果数据错了，证明这个模型真的不行  
那么，如果我们带入 SARS 最严重的的时候，会怎么样 ？  
  
还是按照误差 0.17 来算，在 23 号，误差达到了 0.18 ，也就是 3 天前，按照这个模型的数据推算，新型冠状的病毒传染性比 SARS 要强。但也可能是和春运有关系，毕竟数据的时间不一致。  
  
再再次声明：  
以上的观点和结论都属于依据当前的确诊人数所作的推断，毫无参考的意义和价值。如果数据刚好对了，不能说明什么，如果数据错了，证明这个模型真的不行。  
0x05 再多讲几句  
写这个的目的纯粹是为了探讨和研究，实际上，真的要写一个这种疫情预测和评估的系统，需要非常多的参考的数值，比如医护人员的数量，检测手法准确率，时间，病毒的潜伏期等等，需要非常多的人力物力才可以编写完成，像本文这种只靠单一维度的输入去判断和预估的模型和玩具没什么区别。  
但是在编写的过程中也发现很多问题,，疫情的相关数据还是多方寻找才找到一部分   
例如支援的物品的统计，损耗和使用情况等  
例如病人的数量对物品的需求等  
例如疫情各个时期需要多少的物资统计等  
希望在这次的疫情中保留比较详细的数据 这些都可以作为系统的的输入，借助这些数据，打造一个强大完整的疫情监控和评估系统，甚至不是疫情，而是重大事件的监控和评估系统，待下次再出现疫情、灾情，可以快速的调度全国的资源进行处理和援助。