气象人员如何上手学习 AI（机器学习）

——知识点和视频推荐

1）什么是机器学习（个人总结）？

传统的线性统计模型（例如：多元线性回归模型）中的参数可以通过最小二乘法直接计算求解得出。非线性复杂模型（例如：可以提取非线性特征的浅层神经网络模型）中的参数不能通过某类算法直接计算求解得出。需要使用已有样本数据训练模型，通过不断微调模型中的参数，使模型能够越来越好地拟合样本数据，最后得到一套模型参数。这个训练过程，称之为机器学习。

入门了解机器学习概念，推荐视频

https://www.bilibili.com/video/BV1vJ41147QU/?spm_id_from=333.788.recommend_more_video.-1

2）了解下面几个常用机器学习模型的基本原理

分类模型：K 近邻、决策树、随机森林；

回归模型：浅层神经网络、卷积神经网络、循环神经网络；

学习策略及算法：监督学习、损失函数、梯度下降法。

重点体会每种模型的适用范围，具备根据问题特点选择合适模型的能力。

了解上述知识点，推荐视频：

https://space.bilibili.com/10781175/video

https://www.bilibili.com/video/BV1z5411f7Bm?spm_id_from=333.337.search-card.all.click

3）机器学习模型的性能评估

模型相关概念：模型容量（可调参数个数）、超参数。

样本数据划分：训练集、测试集、验证集。

模型性能评价标准：泛化能力、过拟合、稳定性、可解释性。

了解上述知识点（稳定性、可解释性除外），推荐视频：

https://www.bilibili.com/video/BV1df4y117TT/?spm_id_from=333.788.recom

mend_more_video.-1

https://www.bilibili.com/video/BV1EK411s7hG?spm_id_from=333.337.search-card.all.click

https://www.bilibili.com/video/BV1P64y1F7pj?spm_id_from=333.337.search-card.all.click

https://www.bilibili.com/video/BV1f64y1D7gb?spm_id_from=333.337.search-card.all.click

4）**气象领域使用 AI 技术将面临的挑战**

不同与广泛使用机器学习方法的“大数据”领域（例如：图像识别）。AI 在气象领域应用将会面临两类特殊问题。第一类：气象观测样本数据不够多。气象观测也就百年左右。若分析气候问题，样本量十分有限。天气过程事件虽然相对较多些。但是，其中极端事件所占比例很少。因此，训练模型时往往能抓住一般天气事件的特征，而不能抓住极端事件的特征。然而，对极端事件的预报，才是社会公众最需要的气象服务。若仅使用极端事件训练模型，**由于样本量少，往往导致模型过拟合、不稳定等问题。第二类：气象领域已经积累了很多理论研究，具有大量的气象背景知识。机器学习得到的模型不能违背已有气象背景知识的。若想将 AI 成功用于气象服务，一定要结合气象领域的这两类特殊问题进行有针对性的技术处理。**针对上述问题的研究目前还很少见，无合适的相关资料。这里给出作者本人的研究分析。

在训练机器学习模型过程中，过拟合是一种普遍的现象。过拟合是指能够很好地拟合训练数据，而对除训练数据之外的未知数据却不能进行很好地拟合。例如：记录 10 个人近 10 天每人每天零点零分的心情指数和近 10 天某地区逐日气温。通过求解 10 元一次方程组，就可以得到用这 10 人心情指数完全拟合当天气温的多元线性回归方程。这个多元线性回归方程仅仅对这 10 天数据样本适用，丝毫不具备对今后气温预测的能力。模型训练的最终目的是提高其泛化能力，使其对未知数据也能够较好地拟合。如果训练样本量不足，训练样本代表的特征与总样本特征偏差较大，就会导致过拟合（Goodfellow et al., 2015）。这里采用机器学习网络教程中常用的手写数字识别数据集和相应模型展示过拟合现象。这套

数据集共有 6 万个样本，通常随机抽取 5 万个作为训练数据、剩下的 1 万个作为测试数据。在 5 万个训练样本情况下，训练若干步后，模型对训练集的拟合精度和对测试集的拟合精度几乎相同（图略）。在“大数据”样本充足情况下，过拟合问题可以忽略。若从 6 万个样本中仅抽出 200 个样本作为训练数据。模型对训练集的拟合精度和对测试集的拟合精度呈现显著差异（图 1）。从图中可以看出，当训练迭代次数超过 100 步时，模型对训练集的识别精度几乎达到 100%。但是，对于测试集来说，识别精度不超过 80%。主要原因是模型拥有大量可调参数（大于 1 万），在训练数据偏少情况下，导致模型过多地去努力抓住训练样本独有的特征，而不是数据总体的共同特征。两个数据集上识别精度的明显差异说明过拟合问题不可忽略。总之，**在样本量不足（样本量远小于模型可调参数个数）的情况下，机器学习方法训练得出的复杂模型（模型容量较大）往往会产生严重的过拟合现象。若不分析了解过拟合问题，往往会导致机器学习模型在汇报介绍时效果很好、但用于实际业务后效果下降。**

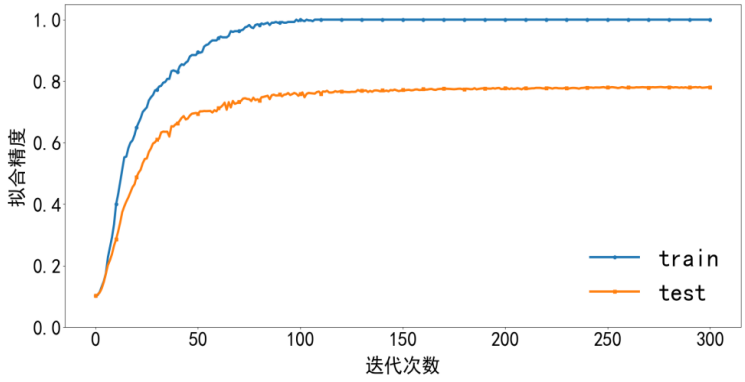

图 1　拟合精度随迭代次数的变化。两条曲线分别为训练集(train，蓝色)和测试集(test，红色)。

机器学习模型在训练过程的参数初始化、样本使用顺序等方面常常具有一些随机性。因此，基于完全相同的训练样本，每次训练得到的模型参数不同。在“大数据”情况下，这种随机性有利于抑制模型过拟合、便于训练找到最佳模型参数。但是，在小样本情况下，每次训练得到的模型在测试样本上的计算输出结果往往存在较大偏差，即模型不稳定。这里采用机器学习网络教程中常用的房屋价格预测模型及其数据集来举例说明模型稳定性问题。这套数据共有 506 个样本。在小样本情况下，设置训练样本量为 40 个，测试样本 20 个。为对比小样本产生的影

响，也开展正常的全样本试验（即大样本情况，除 20 个测试样本外的数据全部用于训练模型）。图 2 展示了小样本情况下模型预测结果的稳定性情况。预测结果的稳定性由 10 个集合成员（即 10 次训练得到的模型）的标准差来衡量（即图中的 error bar）。小样本情况下模型预测结果的标准差在 0.3 左右，远高于大样本情况下的标准差（0.05 左右）。小样本情况下模型预测技能（用预测值和实际值的相关系数 Cor 衡量）为 0.58±0.11，而大样本情况下模型预测技能为 0.77±0.02。由于过拟合现象，由小样本训练得出的模型预测技能明显低于大样本训练得出的模型。此外，由小样本训练得出的模型预测技能的稳定性（用 10 个成员计算得到 Cor 的标准差衡量，0.11）也明显弱于大样本训练得出的模型（0.02）。总之，**由于样本量少，机器学习模型性能存在明显的不稳定性。一个不稳定的机器学习模型，肯定不能用于气象服务**。

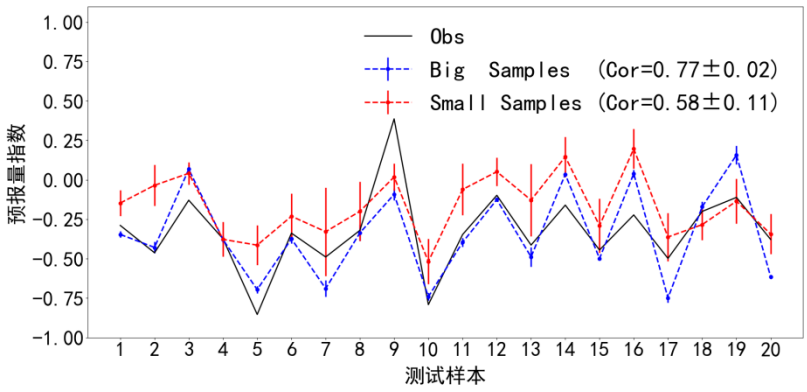

图 2 基于测试样本得出的模型预测结果和相应实际值。黑色线为实际观测值；红色线为小样本下模型的预测结果；蓝色线为大样本下预测结果；竖线表示预测结果的不确定性范围（即 error bar，10 个成员的均方差）。

如何抑制小样本带来的上述困扰，还有待进一步深入研究。这里仅仅给出两点思路，以供读者参考。第一，人为扩充样本。仿照其他机器学习应用领域样本扩充技术，对气象事件样本进行一定程度的扩充，增加样本量。例如：借鉴图像识别领域中的图像数据扩充方法（如：翻转、旋转、变形、裁剪、添加噪声和像素缩放等），利用气象背景知识对气象事件样本数据进行合理扩充，即增加一些符合气象背景知识的人为制造的气象事件。**第二，尽量使用可调参数较少（即模型容量低）的简单模型。**图 3 展示了，小样本情况下，复杂神经网络和简单神经网络的模型性能差异。为模拟小样本情况，采用 60 个数据样本开展试验（训

练样本 40 个，测试样本 20 个）。复杂模型所使用的神经网络模型有 8 个隐藏层，每层 50 个节点；简单模型为有 1 个隐藏层的浅层神经网络模型，节点数设为 20。相比复杂模型，简单模型预测结果的标准差明显减小。此外，简单模型的预测技能（0.60）略高于复杂模型（0.58）；预测技能的不确定性范围（0.09）也明显优于复杂模型（0.16）。

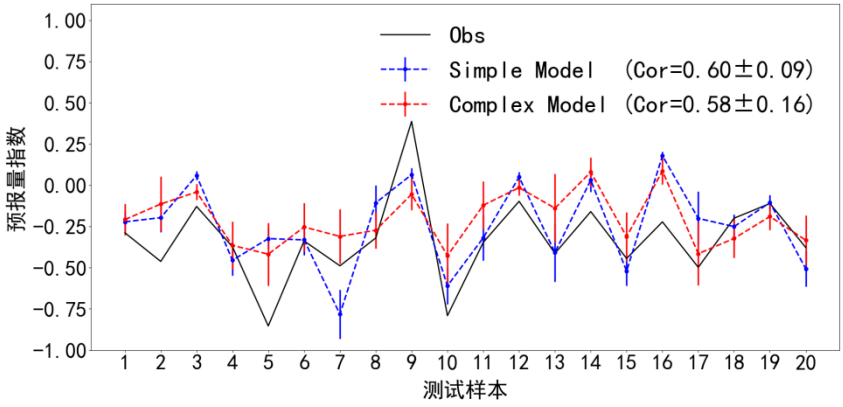

图 3 同图 2，小样本情况下，复杂神经网络模型和简单神经网络模型对测试样本预测结果的对比。

机器学习方法在气象领域中的应用越来越多，正在成为一个强有力的工具。然而，机器学习模型通常被比喻为“黑箱子”。人们很难理解或者解释机器学习模型到底学到了什么以。**对模型结果无法解释，就难以根据模型结果进行决策。**因此，分析评估模型可解释性十分重要。在分析机器学习模型可解释性的众多方法中，有一个几乎通用的方法——“梯度扰动”方法。该方法把统计模型当作“黑箱子”，通过给输入数据施加小扰动，分析模型输出结果对各个输入数据的敏感性。**“梯度扰动”方法可以了解模型中各个输入量对输出结果的贡献。然后将之与气象背景知识进行对比，判断模型是否具有可解释性。凡是与气象背景知识相悖的模型，肯定不能使用。**

这里以一个基于前期海温预测副高指数的神经网络模型进行讲述。图 4 展示了模型输入量（前期秋、冬海温）对模型输出量（6 月西太副高指数）的贡献。通过相关系数可以初步了解前期秋、冬海温对 6 月西太副高指数的影响（图 4 上）。前人大量研究工作已经对这种线性统计关系给出了合理的物理解释，并得到广泛认可。在这里，将其当作气象背景知识。与线性相关性不同，非线性统计模型（神经网络模型）在某区域（如北太平洋）的敏感性会受到其他区域（如东

太平洋）海温状况的影响。因此，由“梯度扰动”方法计算得出的神经网络模型敏感性空间分布图具有年份依赖性。图 4 给出了四个样例年（1962、1974、1983 和 2004 年）的输出结果对输入数据的敏感性空间分布图。这 4 年神经网络模型敏感性空间分布特征与气象背景知识（线性相关性）基本一致。这四年实际观测的副高指数分别为：−0.15（正常）、0.70（高），−0.58（低）和 1.00（高）；模型预测的副高指数为：−0.04（正常）、0.51（高），−0.56（低）和 −0.38（正常）。1962 年、1974 年和 1983 年的副高指数预测基本正确，而 2004 年的预测显然错误。相对于预测正确的年份，预测失误年份（如 2004 年）模型输出结果对输入数据的敏感性整体相对较弱，其空间特征与线性相关性的差异较大。大多数预测错误年份都有这种特征（图略）。也就是说，**通过分析模型可解释性还可以在一定程度上了解当年预测的可信度。若敏感性整体偏弱、空间分布与气象背景知识差异较大通常预示该年份的预测可信度较低。**

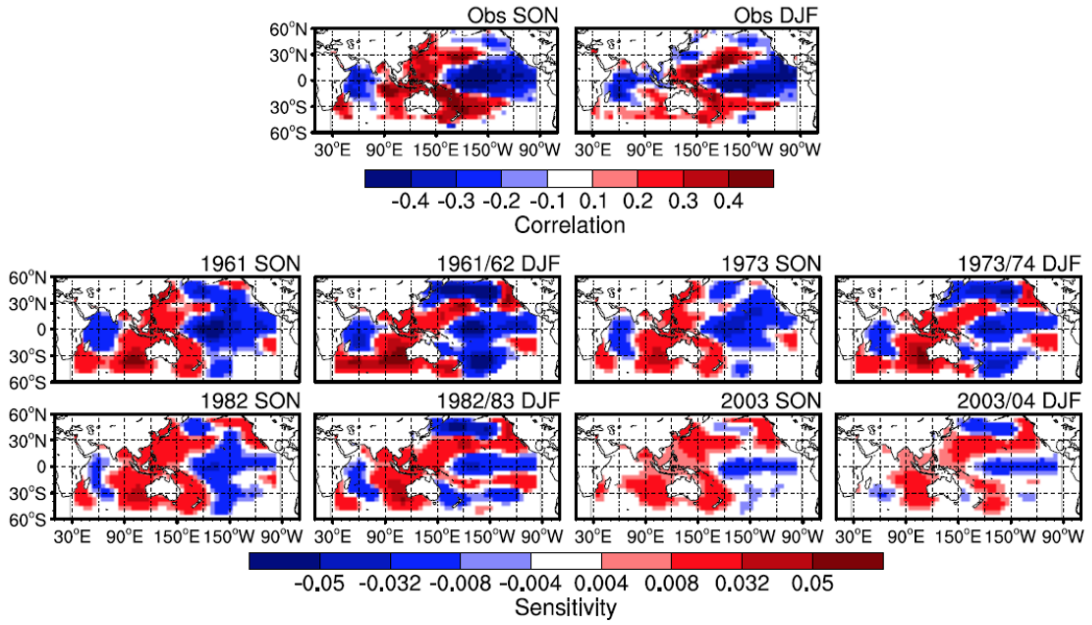

图 4　模型输入量（前期秋、冬海温）对模型输出量（6 月西太副高指数）的贡献。上面板为基于观测数据得出的输入量与输出量的线性相关；下面板为神经网络模型基于某年样本计算得出的输出量对输入量的敏感性。