中国最糟糕的奥赛：NOAI

作者 ducati 寄

发布时间 2024-06-21 16:06 分类 生活·游记

文章内容较长，可以直接跳到【考试后的思考】部分。

前言

个人经历

在省选结束并退役后，我开始一门心思地学习 ai，准备参加 NOAI 中国站的比赛。

之前做过的项目包括但不仅限于 MNIST, CIFAR10, GAN 以及一些传统机器学习的小项目，但总感觉缺了点什么。经过长时间的思考，我认为以下三个问题较为严重：

- 我并没有理解一些传统算法背后的数学本质。
- 我的调参经验并不过关，别人的 Loss 总是比我低。
- 我没有做过 kaggle 的项目。

为了理解一些算法的本质，我不得不开始接触一些算法背后的数学理论。刚开始学线性回归背后的数学理论时，说不吃力是不可能的，因为我只有微积分、线性代数和概率论的并不算非常坚固的基础，面对一些复杂的凸函数，显得有气无力。不过最后，我还是勉勉强强啃下来了。随后，我一鼓作气推完了线性回归、逻辑回归、GLM · · · 收获很大。

为了积攒调参经验，我做一些 NLP 的项目。照着视频和别人的博客，做了 Pytorch 的 NLP From Scratch 的第一个项目，做了 NER（命名实体识别）。此外，我也实现了 RNN, LSTM, GRU，但由于时间紧迫，没有来得及实现 Transformer。

最后，我在 kaggle 上做了一个 Advanced Regression Techniques。学完了不少东西之后，我自以为水平还可以，结果做这个项目上来就是当头一棒，发现自己啥也不是。我不得不参考了很多人的代码，加了不少数据预处理的步骤，同时学会了数据可视化的 matplotlib 库，最后优化了一下特征工程和模型参数，勉勉强强达到了 Top 6.3%。前前后后花了一个月时间，做完之后还是很有成就感的。虽然也很累。

我认为在这段时间内，我学了不少相当基础，同时也相当有用、重要的知识。虽然相比于业界标准我还是很菜，不过至少有一定的信心了。

同时，基于我所学的内容，我自然而然地开始思考 NOAI 的考试内容。

考试前的思考

Q1: 比赛是一定要有区分度的，对于这种 AI 的比赛，究竟能怎么区分呢?

像 kaggle 那样，看看谁的 Loss 更低？显然不合适，在临时的机器上（插件都得自己配，环境也不熟悉）限时完成项目，还有可能不给看 documentation、纯属搞笑。

像样题那样，从代码里面抠出一些空，让大家填，最后看看谁的代码和注释写的规范，谁能实现得更符合需求，谁的 Accuracy / Loss 更低？我一度以为是这样的，并且深信不疑。

但是我都错了。

Q2: 比赛环境?

编辑器？反正 vscode + jupyter notebook extension + vim extension 肯定是不给用了，伤心。到时候凑活着用个裸的 vscode 吧，应该还能接受。就算不是 vscode 也能忍一忍，但代码补全应该不可能没有吧（埋下伏笔）？

此外，比赛应该是不给上网的，documentation 应该不会不给吧？

以及，python 会预装哪些包？如果我需要 seaborn, nltk, transformers 这样的包，我从哪里下啊？

转头一想，这些问题真的让人很无奈。习惯于在自己用了好多年的机器上做项目的人，把他拉到一个几乎是裸机的机器上打比赛，还不给上网，真的是很难受的。

但又转头一想，也就难受几个小时，忍一忍其实也就过去了；并且这也毕竟是去大学报道前的最后一场比赛了，还是很像拿一个奖状再去的。

Q3: 考纲到底包含哪些内容?

考纲里写了很多东西。每个部分深挖下去都是很大一个坑，但问题在于我不知道要挖多深，以及哪些问题需要挖得更深。

不过最后发现，其实是我多虑了。

考试经历

不想再提。

考试后的思考

0 为什么成绩不予公开?

奥赛活动最基本的精神是：公平公正公开。但 NOAI 的机试和笔试分数，却全部不得公开。

不由得令人浮想联翩。

1 关于机试

1.1 关于规范性

首先，我们讨论机试规范性的问题。

第一，为什么没有代码补全，也不让使用其他编辑器?

没有配置 jupyter notebook 的代码补全其实也能理解，应该是主办方忘记或者懒得给机器装 nbextensions 了。

但真正让我相当困惑的地方在于，凭什么不允许使用其他的编辑器?

考试时监考员发现有人在用 vscode，直接对其进行警告。我不是说监考员按照规则警告不对，而是说这个规则本身就很离谱。我的评价是，vscode 怎么你了？逼着大家拿着一个裸的 jupyter notebook 写代码，jupyter notebook 是你爹吗？各人用各人喜欢的不行吗？

第二，主办方是否对所谓的"实验报告"有很大程度的误解?

首先，实验报告不是答题卡。

然而，比赛要求：选手需要把模型构建、训练日志、数学计算粘贴到 Word 中，填在对应的方框中。那 jupyter notebook 是干什么用的呢?

其次，实验报告也不是用来传附件的地方。然而，很多“附件”也被一股脑塞进了实验报告。甚至其中所有的代码从 Word 里粘到 jupyter notebook 里都没有缩进（注意语言是 Python，相当于 C / C++ 把缩进和大括号全删了给你）。于是，我不得不手动添加缩进。

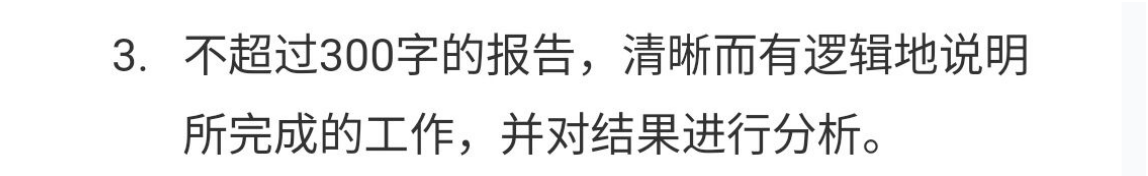
值得一提的是，实验报告也是 IOAI 比赛中的重要判分标准（详见 1.3.1）。我绝不相信 IOAI 的要求就是报告 = 答题卡 + 附件。

下图源自 链接。

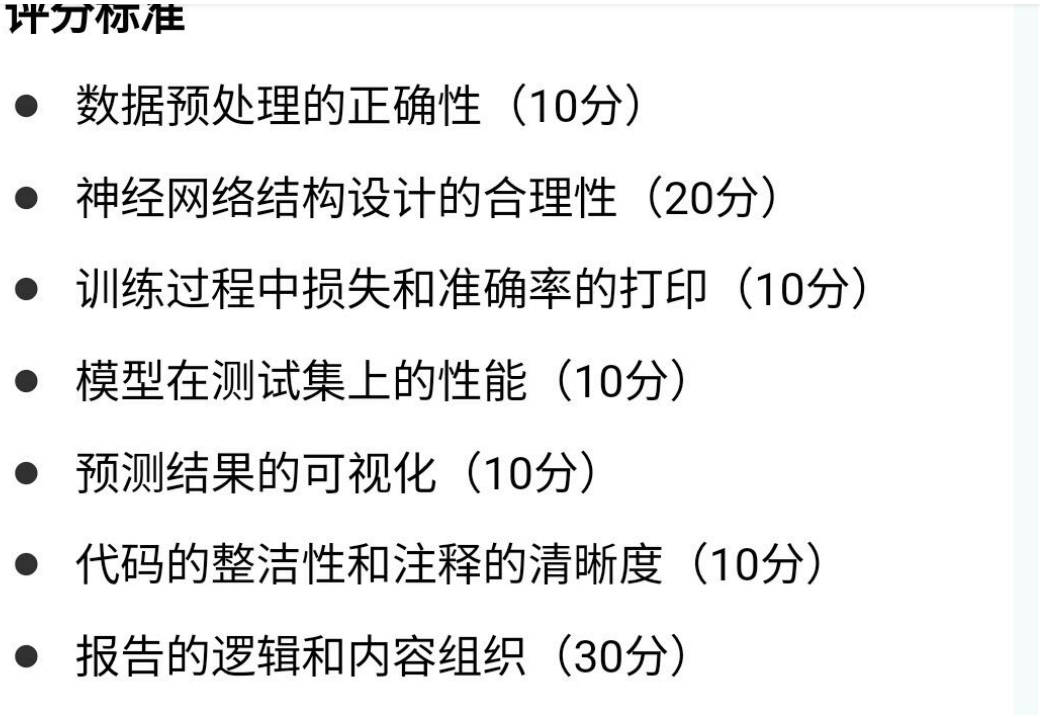
The deliverables for each problem will be clearly stated in the problem description and may include: a score measured on a specific data split, a short written answer or methodological report, a plot visualizing some statistics or results, and others. Each

第三，为什么机试的实际考题与样题相去甚远？

1. 样题明确要求了“不超过 300 字的报告，清晰而有逻辑地说明所完成的工作”、“模型在测试集的性能”。但在实际考试中，报告得分没有体现，并且模型的性能也明确要求忽略。



2. 样题的得分分布形如 $10+20+10+10+10+10+30=100$，实际考试却都是 $20+20+20+20+20=100$。同时，一个任务（例如第一题的第五小问，20 分）却对应样题的



上述原因导致样题很大程度上失去了其所拥有的参考价值，甚至具有相当的误导性。

这里还想着重讨论一下实验报告的问题：原来样题里面是有要求的，IOAI 自始至终也一直是有要求的，那为什么比赛的时候就突然变成答题卡了？

第四，为什么考前临时将上午机试的题目数量从两道题改为四道题，且不发邮件通知？

结合第三个问题，不难得出下面的推测：由于**比赛是临时准备的，时间非常紧迫**，实际情况难以全部与先前公开的信息吻合。于是，“聪明”的主办方想到了一个“办法”，即**直接修改这些信息**，使得让这些已公开的信息与实际情况“吻合”，而不是让实际情况与已公开的信息吻合。

这导致的两个直接的结果是：

- 先前已经仔细研究过考纲、学生细则等资料并认真准备的同学（包括我本人），都成了牺牲品。
- 题目中出现不少具有歧义的地方，明显是由检查不到位造成的。我在考试期间对监考老师进行了多次提问，其中题目本身出现问题至少占了一半。

1.2 关于区分度

1.2.1 究竟区分度是什么？

众所周知，比赛是一定需要有一个区分度的，否则大家都是一样的得分。那么，这场比赛的区分度是什么呢？

我也不完全清楚，可能是手速，也可能是对 pytorch, matplotlib 等常用包的熟练度，或者其他。但总之既不可能是代码规范，也不可能是解决实际问题的能力，更别谈调参经验。

原因如下：

- 主办方给出的代码都没有任何规范可言，而选手会写出怎样可怕的代码，是想都不敢想的。我也写出了不少屎山代码（当然是不得已的），毕竟写代码拼速度，不写屎山代码写什么？
- 解决实际问题本身就是个笑话。
- **试卷里明晃晃地写着让大家忽略 Loss 和 Accuracy。**过拟合，欠拟合，都没事，因为几乎不会对成绩造成影响。如果没记错的话，对这一点进行考察的只有一个小问，分值占比应该不超过整张卷子的 $\frac{1}{40}=2.5\%$。

如果说是因为考场机器的性能不一样，导致不能**直接**以 Loss 和 Accuracy 作为区分的话，其实也能理解。但是，对于所有搭建并训练模型的题目，将**过于糟糕**的 Loss 和 Accuracy，以及极端过拟合、极端欠拟合筛选掉，是否为更好的选择？

1.2.2. (※※※) 糟糕的核心区分度

下面，我们将谈到**一个比较深刻的问题**：糟糕的「核心区分度」。

1.2.2.1 OI 的核心区分度

首先，请允许我使用 oi 来引出「核心区分度」的问题。

先表明我的观点：oi 的「核心区分度」（低阶的比赛不作讨论，这里仅讨论 NOIP 及以上的比赛）是降低时间复杂度。

为什么呢？显然，能走到 NOIP、省队选拔、以及 NOI、甚至 CTT, CTSC, IOI 的选手谁不会写暴力，但是有的人能做到 $O(n^2)$，有的人能做到 $O(n\sqrt{n})$，有的人能做到 $O(n\log n)$，因此获得了不同的分数。

如果一个人能在这个「核心区分度」的方向上做到极致，结果会怎么样呢？说大一点，可能就 P = NP 了 ~~（虽然我更倾向于 P $\neq$ NP）~~。说小一点，能在一个 open problem 上获得进展，发一篇论文，纵使可能在几十年内仅有理论价值。总之，优化时间复杂度，这件事情本身是非常有意义的。

因此，oi 大体给大家指向了一个相当有价值的方向，这也在某种程度上体现出了 oi 的意义所在。

1.2.2.2 Kaggle 的核心区分度

通常来讲，kaggle Competition 的评估指标会非常明确地在主页上被提到。例如 advanced-regression-techniques 中的取对数后 RMSE，再比如 ai-mathematical-olympiad 中的 Accuracy，etc.

而这些评估指标，无一例外都是为了检验你的预测是否足够准确，从而反应出模型结构、超参数等的优劣程度。

1.2.2.3 NOAI 的核心区分度

反观 NOAI，其「核心区分度」又是什么呢？

就像 1.2.1 所说的那样，关于核心区分度，我依然不完全清楚——可能是手速，也可能是对 pytorch, matplotlib 等常用包的熟练度，etc。

不管是哪一种，在这种「核心区分度」上做到极致，会怎样呢？对不起，没有任何用处。至少我不明白把包调得滚瓜烂熟，以及在一个完全没用过的机器上拼手速有什么真正的用处。

如果你觉得有用的话，那当我没说好了。

1.2.2.4 更进一步的思考

在我看来，比赛的**本原价值**从来不是为了比赛而比赛，而是找出一群真正热爱一个领域的人，把他们聚到一起，交流讨论，相互竞争；同时吸引一批又一批的人，**引领他们向着某个大致正确的方向努力**。典型的例子就是 OI。

而在 NOAI 中，由于糟糕的「核心区分度」，导致其指向了一个错误的方向。既然方向是错的，那比赛本身便毫无意义了。

在我看来，这是 NOAI 的致命伤。先前提到的比赛规范之类的都可以归结为首次举办并无太多经验——但核心区分度的问题，便无可辩驳了。

那么，究其根本，为什么会出现一个错误的「核心区分度」呢？

1.3.1 能否真正与 IOAI 接轨?

如果说单单是为了选拔国家队去参加 IOAI 的话，这场比赛的区分度也做得有些奇怪。IOAI 分为两个部分，在 IOAI 规则 中给出了详尽的解释。我这里做一个概括：

- **科学轮**：在这一轮中，参赛团队将**模拟现实科学研究**、识别和解决现有方法局限性的过程。**团队**将**提前三周**收到**基于 AI 研究前沿**的问题。在现场，团队将收到一组新的题目，在先前收到的题目中稍作改动。评分基于数据划分、简短答案、方法报告、可视化结果 · · · · · ·
- **实践轮**：选手需要观察、分析并解释针对 Chatgpt, Dalle-2 等广泛使用的 AI 软件的行为的问题。

首先，“模拟现实科学研究”这一点，在 NOAI 比赛里就是纯搞笑。

其次，NOAI 的报告都是粘图片、粘代码，几乎不需要写什么文字内容来连接。没有语言组织能力、不知道怎么写报告的选手，或许也能拿高分。至于可视化的部分，都是规定好的——如果不规定的话，我害怕会不会真的有人不知道要可视化什么。

还有就是老生常谈的英语能力、交流沟通能力，和团队精神，虽然在最后征召 CN team 的时候应该是有所考察（在征召的 8 个人里面选 6 个进入 CN team），但单论 NOAI 而言，这一点完全没有，或者说没有办法进行考察。

2 关于笔试

再谈谈笔试。

2.1 代码补全

笔试最大头的部分在于补全代码，这种题型占了至少一半的分值。

我的疑问是，既然都有机试了，这种题型难道不是多此一举吗？把代码放到纸上让选手补全，难道不应该是迫不得已（机位不够用，参考 CSP 初赛）的事情吗？更加黑色幽默的一点在于，笔试和机试在同一个考场，在纸上补全代码的时候，电脑就在跟前，但不让用，滑天下之大稽。

而且补全代码的时候不提供 documentation。考完路过走廊，大家都在讨论矩阵转置是 .t 还是 .T 还是 .t() 还是 .T()。

至于代码本身，依然非常不规范。代码里多次使用 `np.dot` 做矩阵乘法，迷惑性非常强，不知道的还以为是 element-wise 的。下面的图来自 `np.dot` 的官方文档（注意这个 `preferred`）：

> - If either *a* or *b* is 0-D (scalar), it is equivalent to `multiply` and using `numpy.multiply(a, b)` or `a * b` is preferred.

还有一点值得吐槽。Transformer 并不在考试大纲内，而考的时候只介绍了寥寥几行就让大家补全代码，且主要内容是从 Attention Is All You Need 中抽出来然后翻译的；很多片段缺少上下文，难以理解。

主办方不会真的觉得，一个没有学过 Transformer 的人，只看这篇论文中的寥寥几段，就能明白每一段的意思了吧。我好歹是看过一遍论文的，但只看给的片段还是一脸懵。

同时，`np.dot` 做矩阵乘法和不给转置 API 的来源也是这个题，我严重怀疑出这道题目的人并没有实现过 Transformer，而是从网上某个犄角旮旯那里抄过来的代码，搬到了试卷上。

选择题占的分数是第二多的，着重考察了大家的知识面（覆盖了大纲里的很多内容）。这是机试所不具备的，应该可以说是整个比赛中唯一的亮点了吧。

我对自己的定位一直是：理论水平不咋地，但折腾得很多，捣鼓过很多项目。结果最后笔试获得了“BEST”的成绩，而机试的分数却较低，也算是相当幽默了。

充满挑战性和竞争性。

然而，尽管我降低了预期，整个比赛还是让我大跌眼镜。整个比赛纯粹是为了区分而区分。应试教育的问题我不想再谈。但 AI 的浪潮扑面而来，世界将被彻底改变，某些人却毫不犹豫地旧事物的方法套在了新事物上，就像给新事物套了一层枷锁。

我不知道有多少初次上机的选手，因为这场比赛，对 AI 的第一印象成为了“调包”、“拼速度”、“补全代码”。对于这些原本对 AI 世界充满着好奇心的他们来说，这种比赛形式只会给他们带来一种片面的、甚至是误导性的认知。他们或许会认为 AI 竞赛只是在拼凑代码片段、快速实现功能，而忽略了 AI 领域背后深厚的理论基础、创新性思维以及解决实际问题的能力。

如果这些选手因为这一场糟糕的比赛，失去了他们对 AI 的珍贵好奇心，那将是一个巨大的遗憾。初次接触 AI 的他们，本该被充满技巧的模型训练、美妙的模型架构所吸引，被 AI 的潜力和其无限可能性所激励。而现在，这场比赛或许正在无形中抹杀他们的热情，让他们误以为 AI 不过是冰冷的代码和枯燥的速度比拼。

以上加粗的两段由 AI 扩写生成。由此可见，无论是碳基生物还是硅基生物，似乎都表达了强烈的不满。

叠甲

以上仅代表个人观点，仅针对比赛本身。

考前几天的培训很棒，给了我和大佬面对面交流的机会，我也在课后向老师请教了不少问题（比如 AI + Science 的数据获取问题，GAN generator 和 discriminator 模型复杂度间的 trade-off，label smoothing，etc），学到了很多知识。至于课堂内容，印象最深的是在 NLP 的课上，老师把 RNN, LSTM, GRU, Transformer 串起来讲，让我在各个模型的细节之上，第一次以一个宏观的视角，去同时看待这几个模型的优劣。

此外，考试时监考老师的态度也非常好，我多次举手提问，老师非常耐心地回答了我的问题。

在这样的一套评价体系下，我技不如人，心服口服，甘拜下风。

作者：ducati 创建时间：2024-06-21 16:06:46

 9  54  不推荐

评论区

发表评论

发表一条友善的评论吧！

请先登录

12 条评论 默认排序

 NaCly_Fish 绒布球 回复于 8 个月前

__szh_DNCB__ 回复于 8 个月前

不公开……

hhoppitree Madeline 回复于 8 个月前

在这样的一套评价体系下，我技不如人，心服口服，甘拜下风。

yzy1 Ed<βDream 回复于 8 个月前

这什么逆天比赛.

yaoxi 回复于 8 个月前

NO 开头的竞赛都终将被取缔!

chenxia25 回复于 8 个月前

ya 开头的人都终将会死!

Eason_cyx 回复于 7 个月前

可能因为第一届吧，，不过确实抽象，玩原神玩的

ducati 寄 回复于 7 个月前

唉，其实规范性都还算小问题，正常来说以后只会一届比一届更规范。

但问题在于区分的方向是错的，以后再怎么改，估计也都是换汤不换药

rachel2021 回复于 7 个月前

比赛组织机构就很迷，明显是个商业机构。训练营应该收了不少钱
不知道为啥不找计算机学会、人工智能学会这样的机构来组织。

听取T声一片 回复于 7 个月前

好石依托，依托好石啊

wrongaaa 回复于 7 个月前

都不公开那有什么意义，实属逆天

bcyq 回复于 5 个月前

加载更多