Essay

Your Name

1 Data Cleaning

我们的主要模型将任务处理为了一个二分类问题，而原题中的数据集有很多特征表示并不是可读取的类型，或者并不适用于二分类任务，所以我们首先更改了数据的格式，将 p1_score, p2_score, winner_shot_type 进行了转换，将网球特殊的计分规则转化成 0,1,2,3,4 这样的一般性计分规则，将正手和反手等特殊的击球方式转化成了 0,1,2 这样的分类特征。同时，为了获取特征重要性，我们在构造主要模型之前，使用了机器学习的方法，希望能够知道哪些特征对于比赛结果的影响最大，进而对这些特征进行增强。

1.1 Feature Engineering

1.1.1 Lasso Regression Analysis

一开始我们使用了 Lasso Regression 的方法，因为 Lasso Regression 的损失函数中加入了 L1 正则项，有助于让模型忽略一些不相关的特征，使不重要的数据重要性分数趋于 0。

1.1.2 Over Sampling

但是 lasso 回归并没有展现出很好的效果，我们分析这是因为不平衡的数据导致的。毕竟顶尖的网球选手出现双发失误等情况的概率较小，很多变量的分布并不平均。

(附图：一些特征的分布)

为了处理不平衡的特征，我们使用了 SMOTE(Synthetic Minority Oversampling Technique) 方法合成了少数类样本。通过 SMOTE 过采样，可以使得模型更好地学习到少数类别的特征，从而提高模型的泛化能力和准确

性。此外，SMOTE 过采样方法还可以减少模型的过拟合倾向，提高模型的鲁棒性。SMOTE 过采样方法基于样本的特征空间，通过对少数类样本进行插值来生成合成样本。其主要步骤如下：

1. 对于每一个少数类样本，计算其与所有其他少数类样本之间的距离，并找到其 K 个最近邻居。

2. 从这 K 个最近邻居中随机选择一个样本，并计算该样本与当前样本的差异。

3. 根据差异比例，生成一个新的合成样本，该样本位于两个样本之间的连线上。

4. 重复上述步骤，生成指定数量的合成样本。

1.1.3 Random Forest Feature Importance

最终，我们使用了 Random Forest 的方法，通过对特征的重要性进行分析，我们最终选取 ranfom forest 的原因是它作为一种 Ensemble Learning Method，拥有比较好的抗噪能力，同时不会出现过拟合的情况。

(可能可以加入 Random Forest 的原理介绍？)

在结果层面上，我们发现了一些特征对于比赛结果的影响较大，比如选手的 Elo score，选手的体力、谁的发球局等等。

表 1: Distribution of Some Features.

Feature	0	1	2	Total
Game Victor	6096	616	572	7284
Winner shot type	5619	1163	502	7284
$\mathrm{P}_i Ace$	13902	666	-	14568
$\mathrm{P}_i Winner$	12135	2433	-	14568
$\mathrm{P}_i DoubleFault$	14314	254	-	14568
$\mathrm{P}_i UnforcedError$	12658	1910	-	14568
$\mathrm{P}_i NetPoint$	12872	1696	-	14568
$\mathrm{P}_i NetPointWon$	13386	1182	-	14568
$\mathrm{P}_i BreakPoint$	14064	504	-	14568
$\mathrm{P}_i BreakPointWon$	14391	177	-	14568
$\mathrm{P}_i BreakPointMissed$	14241	327	-	14568

2 Extend Our Model

2.1 Generalizing in Table Tennis

为了考察模型的泛化能力，我们使用了 2021 年东京奥运会的乒乓球男单决赛作为检验模型泛化能力的乒乓球数据集。用乒乓球来检验模型的泛化能力是因为乒乓球和网球有很多相似之处。但是，网络上并没有和题目所给的数据集相似的乒乓球数据集，于是我们使用了人工标注的方式，

3 Experiment Results

The table below demonstrates the sensitivity and comparison of different models. The candidate pool column indicates whether the candidate pool is used or not. The quality column indicates whether the quality feature is used or not. The accuracy columns show the average loss and the accuracy of the models for the Tennis, New Tennis, and Pingpong datasets. The Elo score column indicates whether the Elo score is used or not. The results show that the Lasso $+\lambda$GRU model with the candidate pool and quality features has the best performance for the Tennis and New Tennis datasets.

3.1 Analysis

（这一部分是表格的描述）

在我们的实验中，我们对于两个模型均选择了交叉熵损失函数作为模型的损失函数，在原始数据集上裁剪出了几场比赛作为测试集，为了检查我们模型的泛化能力，在对完整的数据集训练结束后，我们又只对训练集进行了训练，以测试模型对于新网球比赛的预测能力。相似的，我们选取了 2021 年东京奥运会的乒乓球男单决赛作为检验模型泛化能力的乒乓球数据集。这张表格反馈了不同的模型在对温网比赛的学习能力以及对其它比赛的泛化能力。从结果上我们可以看到，在我们的实验中，λ 取 0.5 的情况下，模型又可以学习到更多的数据，有更强的准确性，同时还继承了 Lasso Regression 的泛化能力，找到了一个平衡点。

表 2: Sensitivity and Comparison of Different Models.

	Candidate pool		Accuracy			
Model	Elo score	Quality	Avg loss	Tennis	New Tennis	Pingpong
Lasso Regression	✗	✗	0.42	0.85	0.84	0.75
	✗	✓	0.41	0.85	0.84	0.75
	✓	✗	0.40	0.86	0.84	-
	✓	✓	0.41	0.85	0.84	0.79
GRU	✗	✗	0.45	0.88	0.79	0.51
	✗	✓	0.40	0.90	0.79	0.54
	✓	✗	0.40	0.90	0.79	-
	✓	✓	0.40	0.91	0.80	0.56
Lasso +λGRU	✗	✗	-	0.88	0.82	0.68
	✗	✓	-	0.89	0.82	0.69
	✓	✗	-	0.90	0.82	-
	✓	✓	-	0.91	0.83	0.72