数据科学导论实验 6：ESI 深度学习排名预测

姓名：杨云天

学号：**10245501405**

一、 深度学习模型预测学科排名

1. 数据说明

学科总数：22 个
数据特征：
- Web of Science Documents
- Cites
- Top Papers
- 机构在该学科的全球排名

- 忽略 Cites/Paper 原因：篇均引用可以通过 Cites 与 Web of Science Documents 计算获得，相关性极高，无需选取。

数据形状：(row_size, 4)

2. 数据集划分

训练集：60%
验证集：20%
测试集：20%
注: 每个学科的数据在划分前进行随机打乱

3. 数据预处理

3.1. 数据归一化
- 排名：反向归一化到 [0,1]，排名越靠前（数字越小），归一化值越接近 1
- 特征：正向归一化到 [0,1]
- 使用 sklearn 的 MinMaxScaler
- 归一化原因：量纲差异会导致数值大的特征在计算中占据主导地位，例如引用量（万级）会掩盖顶级论文数（百级）。对于回归模型，量纲差异会影响距离计算或系数估计，导致结果偏向数值范围大的特征，无法客观反映各特征的真实重要性。

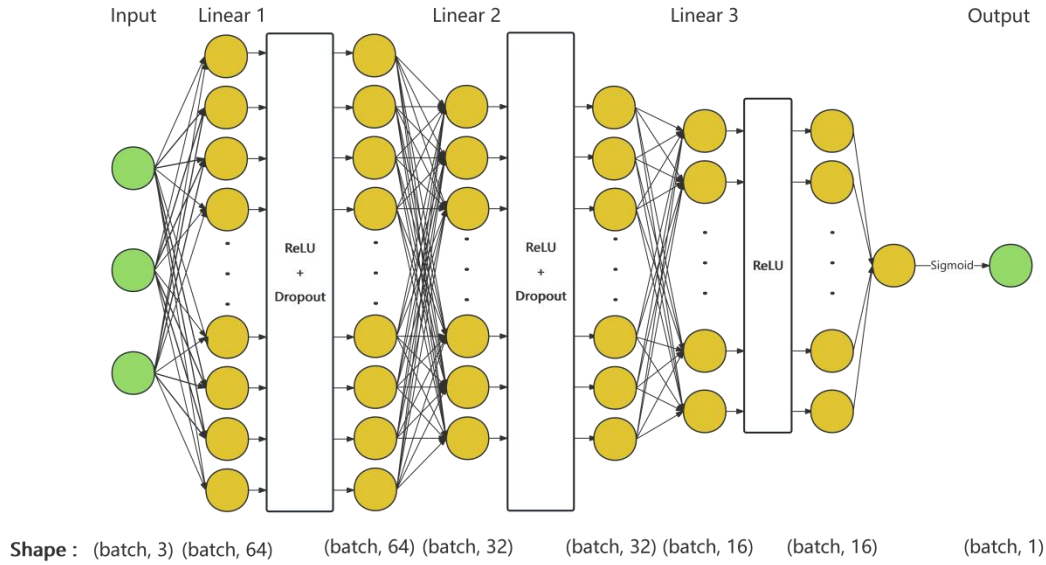
3.2. 反归一化:
预先保存所有学科的排名范围，便于后续将预测的归一化值转换回实际排名。

4. 深度学习模型说明

4.1. 模型架构

操作	输入形状	输出形状	说明
线性层 1	(batch, 3)	(batch, 64)	输入特征数 3→隐藏层大小 64
ReLU 激活	(batch, 64)	(batch, 64)	引入非线性特征
Dropout	(batch, 64)	(batch, 64)	dropout 率 0.3，防止过拟合
线性层 2	(batch, 64)	(batch, 32)	隐藏层大小 64→32
ReLU 激活	(batch, 32)	(batch, 32)	引入非线性特征
Dropout	(batch, 32)	(batch, 32)	dropout 率 0.3，防止过拟合
线性层 3	(batch, 32)	(batch, 16)	隐藏层大小 32→16
ReLU 激活	(batch, 16)	(batch, 16)	引入非线性特征
输出层	(batch, 16)	(batch, 1)	输出单个预测值
Sigmoid 激活	(batch, 1)	(batch, 1)	将输出映射到[0,1]范围

4.2. 模型架构示意图



5. 模型训练

5.1. 训练目标

针对每个学科单独训练一个 RankingPredictor 模型（全连接神经网络），以实现各个学科独立预测。

5.2. 数据处理

各学科数据独立划分训练集 (X_train, y_train)、验证集 (X_val, y_val) 和测试集 (X_test, y_test) 。所有特征与标签均转换为 PyTorch 张量。

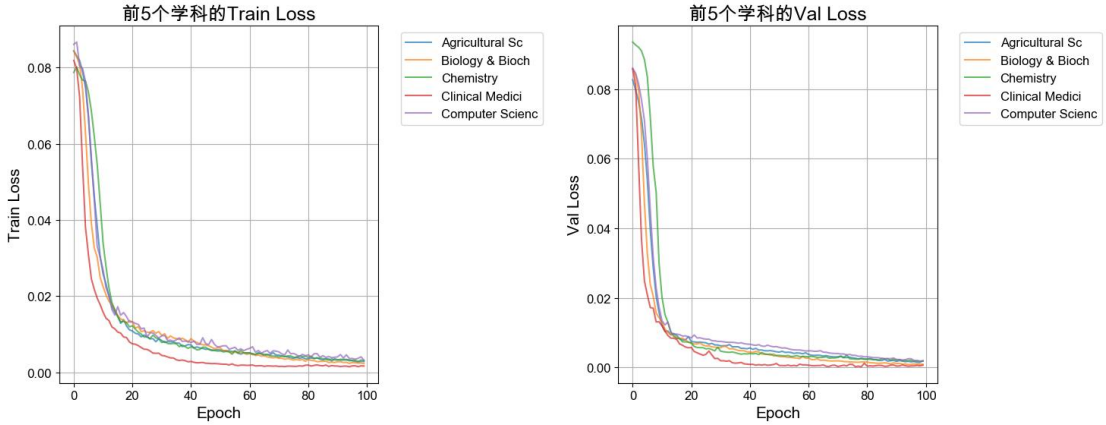
5.3. 训练过程

- 损失函数：均方误差损失（nn.MSELoss），适用于回归任务。

- 优化器：Adam 优化器，学习率 lr=0.001，自适应调整参数更新步长。
- 训练轮次（epochs）：100 轮，每 20 轮输出一次训练与验证损失。
- 批处理大小（batch_size）：32
- 训练流程

① 清空优化器梯度（optimizer.zero_grad()）
② 前向传播计算预测值
③ 计算损失并反向传播（loss.backward()）
④ 优化器更新参数（optimizer.step()）

- 验证阶段：关闭梯度计算（torch.no_grad()），计算验证集损失以监控过拟合风险。
- 记录数据：记录 train loss 和 val loss（下图，可认为 100 轮训练已经收敛）



5.4. 模型保存

模型保存为每个学科保存训练完成的模型，model/subject_name.csv

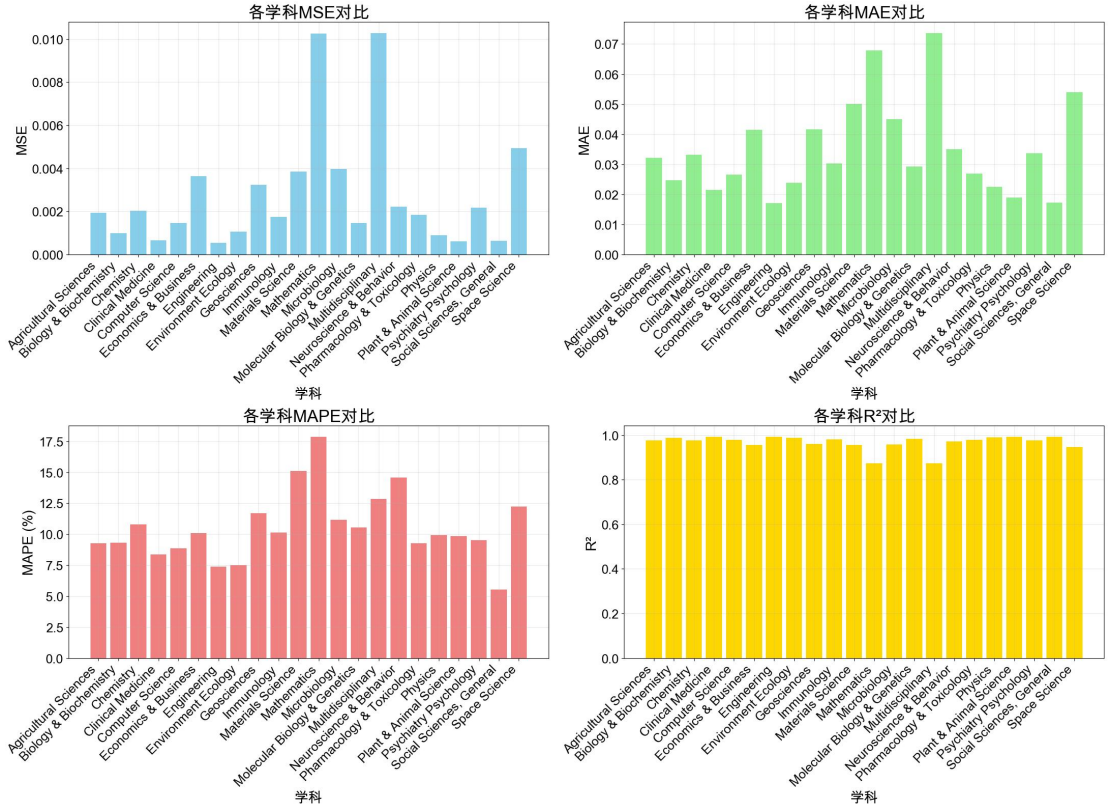
6. 评价指标

6.1. 各项评价指标含义

指标名称	含义	数值
MAE：平均绝对误差	预测排名与实际排名的平均差距	越小越好
MAE：平均绝对误差	预测值与实际值差值的绝对值的平均值，反映预测值与实际值的平均差距	越小越好
MAPE(%)：平均绝对百分比误差	预测值与实际值（反归一化后）差值的绝对值除以实际值后取平均值	为百分数形式，越小越好，超过 50%表示误差较大
R^2：决定系数	反映模型对数据变异的解释能力	取值范围为(-∞,1]，1 表示模型完全拟合数据，0 表示模型效果等同于均值预测，越接近 1 越好

6.2. 训练结果：各项指标分析

	学科名称	MSE	MAE	MAPE(%)	R^2
0	Agricultural Sciences	0.001928	0.032123	9.25	0.975797
1	Biology & Biochemistry	0.001000	0.024718	9.31	0.987788
2	Chemistry	0.002023	0.033146	10.80	0.977278
3	Clinical Medicine	0.000654	0.021424	8.36	0.992258
4	Computer Science	0.001473	0.026576	8.84	0.979503
5	Economics & Business	0.003649	0.041419	10.07	0.955865
6	Engineering	0.000535	0.017130	7.40	0.993360
7	Environment Ecology	0.001065	0.023814	7.49	0.987042
8	Geosciences	0.003248	0.041727	11.68	0.960250
9	Immunology	0.001754	0.030324	10.15	0.980018
10	Materials Science	0.003864	0.050029	15.10	0.954724
11	Mathematics	0.010250	0.067922	17.87	0.873684
12	Microbiology	0.003972	0.045085	11.15	0.957215
13	Molecular Biology & Genetics	0.001473	0.029310	10.56	0.982906
14	Multidisciplinary	0.010278	0.073623	12.84	0.873939
15	Neuroscience & Behavior	0.002225	0.034974	14.57	0.971303
16	Pharmacology & Toxicology	0.001838	0.026934	9.27	0.978258
17	Physics	0.000908	0.022469	9.91	0.989649
18	Plant & Animal Science	0.000606	0.019014	9.84	0.993206
19	Psychiatry Psychology	0.002168	0.033757	9.53	0.975591
20	Social Sciences, General	0.000631	0.017337	5.52	0.992632
21	Space Science	0.004931	0.053929	12.24	0.946475
22	平均值	0.002749	0.034854	10.53	0.967215



分析：

- MSE：平均 MSE 仅为 0.002749，表明模型预测值与真实排名之间的平方误差极小。在 22 个学科中，大多数学科的 MSE 低于 0.004，显示模型在这些领域具有极高的预测精度。仅 Mathematics 和 Multidisciplinary 两个学科的 MSE 略高，但仍在可控范围内，整体而言

MSE 指标反映了优秀的模型拟合能力。

- MAE：平均 MAE 为 0.034，意味着模型预测排名与实际排名的平均绝对偏差很小。Social Sciences、Engineering 等学科的 MAE 低于 0.02，接近完美预测水平。表现相对较差的为 Mathematics 学科。这一指标证明了模型在不同学科间均能保持稳定的预测准确性。

- MAPE：平均 MAPE 为 10.47%，属于优秀预测水平。12 个学科的 MAPE 低于 10%，其中 Social Sciences 达到 5.52%。仅 2 个学科 MAPE 超过 15%。MAPE 指标直观显示了模型在实际应用中的可靠性和实用价值。

- R^2：平均 R^2达 0.969，接近完美拟合水平（接近 1）。15 个学科的 R^2超过 0.97，其中 3 个学科甚至超过 0.99，表明模型能解释绝大部分排名方差。即使最低的 Mathematics 学科 R^2也达 0.874，仍具较强解释力。这一指标充分证明了模型特征选择和结构设计的合理性。

- 总结

该排名预测模型在四个核心评估指标上均展现出较好性能，形成了一个相互印证的高质量评估体系：极低的 MSE 和 MAE 反映了模型精确的预测能力，较低的 MAPE 显示了实际应用价值，接近 1 的 R^2则证明了模型设计的科学性。模型在不同学科间表现稳定，仅在少数复杂交叉领域略有波动，具备投入实际应用的要求。

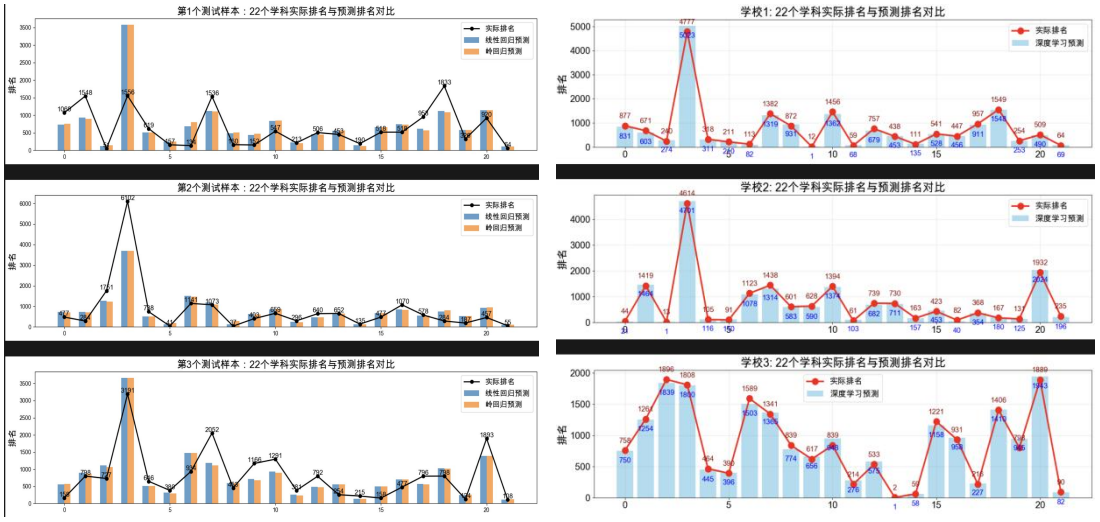
7. 预测

7.1. 加载模型进行预测

从测试集中随机选择三个样本进行 22 个学科的排名预测

7.2. 分析

对比线性回归/岭回归模型（左图），深度学习模型（右图）展现出了强大的优势。（注：线性回归/岭回归 与 深度学习模型 预测样本均为随机选择，不是相同的三个样本）



原因：

神经网络在线性模型基础上引入了非线性变换能力，通过三层 ReLU 激活函数能够捕捉排名预测中复杂的特征交互效应，如论文数量与引用量之间的非线性协同作用。双重

Dropout 层提供了强有力的正则化效果，有效防止在有限数据上的过拟合，提升了模型泛化能力。输出层的 Sigmoid 函数将预测约束在合理范围内，更符合排名得分的物理意义。这些特性共同使神经网络能够超越线性模型的表达能力限制，在复杂排名任务中实现更精准的预测。

二、 基于多学科排名的全球高校聚类分析（同上次报告）

1. 构建排名特征张量

从 22 个 ESI 学科中提取每所高校的排名（依据引用数 Cites）。其中，入榜高校使用其实际排名；未入榜高校引入惩罚机制，填入该学科的最大排名*2（即该学科总机构数*2）。

没有惩罚机制时，未入榜学科被视为"缺失值"，聚类算法可能产生偏差，综合性大学与专业院校难以公平比较。添加惩罚机制后，所有高校在所有学科都有排名，聚类基于完整数据矩阵，能更好识别“全面发展型”vs“特色突出型”高校。

- 构建特征张量 df_features：9990 个机数 × 22 个学科（下图为部分）

	Agricultural Sciences	Biology & Biochemistry	Chemistry	Clinical Medicine	Computer Science	Economics & Business
CHINESE ACADEMY OF SCIENCES	1	3	1	188	1	60
CHINESE ACADEMY OF AGRICULTURAL SCIENCES	2	175	357	4360	826	1086
UNITED STATES DEPARTMENT OF AGRICULTURE (USDA)	3	179	756	1476	738	1086
CHINA AGRICULTURAL UNIVERSITY	4	215	398	3470	273	1086
INRAE	5	54	348	696	591	320

2. 使用 k-means 算法聚类，k = 5

2.1. 数据标准化

- 使用库与方法：scikit-learn 的 StandardScaler 类，实现 Z-score 标准化
- 目的与原理：
  - 将各特征转换为均值为 0、标准差为 1 的标准正态分布
  - 计算公式：

$$z = \frac{x - \mu}{\sigma}$$

- 必要性：避免某些数值较大的特征主导聚类过程

2.2. K-means 算法

- 利用 scikit-learn 的 KMeans 类，基于划分的无监督学习算法
- 参数：
    - n_clusters=5：聚类数 k = 5
    - random_state=42：设置随机种子，确保结果可重现
    - n_init=20：使用不同质心种子运行算法的次数，避免局部最优

3. 聚类结果分析

3.1. 簇规模统计

```
各簇高校数量:
  簇 0: 5100 所高校
  簇 1: 417 所高校
  簇 2: 792 所高校
  簇 3: 3198 所高校
  簇 4: 483 所高校
```

3.2. 华东师范大学定位

```
EAST CHINA NORMAL UNIVERSITY:
所属簇: 1
```

4. PCA 降维实现聚类可视化

4.1. PCA（主成分分析）原理

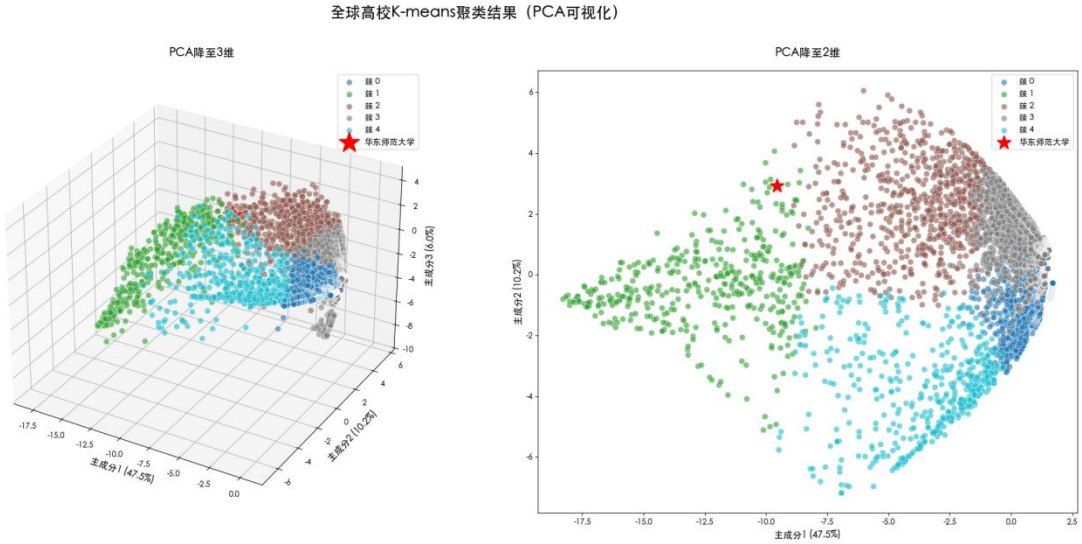
PCA（主成分分析）是一种无监督的线性降维方法。其核心思想是通过线性变换，将高维数据映射到低维空间，同时尽可能保留数据中的大部分变异信息。

4.2. PCA 降维的作用

- 数据简化：高维数据往往包含冗余信息，PCA 降维可以减少数据的维度，使数据更易于处理和可视化，就像图中把高校相关的高维特征数据降维到 3 维和 2 维后，能在 3D 空间和 2D 平面中直观展示聚类结果。
- 去噪：在降维过程中，那些方差较小的成分会被舍弃，从而在一定程度上去除数据中的噪声，保留主要信息。
- 特征提取：从原始的众多特征中提取出最具代表性的主成分，这些主成分能够概括原始数据的大部分特征，有助于后续的数据分析，让分析过程更高效且更聚焦于关键信息。

4.3. PCA 降维效果

4.3.1. 可视化



4.3.2. 坐标含义：

- 左图（PCA 降至 3 维）：
  - 横轴：主成分 1：47.5%表示该主成分能够解释原始数据总方差的 47.5%，它是从原始高维数据中提取出的、包含信息最多的一个维度。
  - 纵轴：主成分 2，10.2%表示其解释原始数据总方差的比例为 10.2%，是仅次于主成分 1、能解释较多剩余信息的维度。
  - 竖轴：主成分 3，6.0%表示解释原始数据总方差的比例为 6.0%，是三个主成分中解释方差比例相对较小的，但仍能补充部分信息。
- 右图（PCA 降至 2 维）同理

4.3.3. PCA 降维可视化结果分析

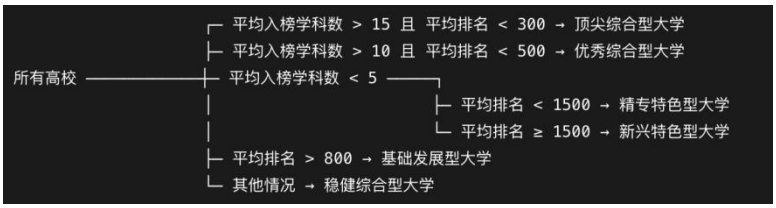
从图中可以看到，华东师范大学在 3 维空间和 2 维空间中，都处于绿色簇（簇 1）的分布区域内。这表明，基于 K - means 聚类以及 PCA 降维后的特征，华东师范大学与绿色簇内的其他高校在相关特征上具有较高的相似性，所以被归为同一类簇。

5. 各簇特征分析

5.1. 总体情况统计

簇	高校数	平均入榜学科数	入榜学科平均排名	最佳排名
0	5100	1.5	3032	1
1	417	19.3	373	1
2	792	9.7	985	4
3	3198	2.1	1368	3
4	483	8.8	771	5

5.2. 高校类型决策树



5.3. 各簇类型推断

```
簇 0: 新兴特色型大学
簇 1: 优秀综合型大学
簇 2: 基础发展型大学
簇 3: 精专特色型大学
簇 4: 稳健综合型大学
```

6. 与华东师范大学相似的高校

6.1. 思路

计算与华东师范大学欧氏距离最相近的 5 所高校

欧氏距离公式：

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

6.2. 结果

```
与华东师范大学最相似的 5 所高校（簇 1 - 优秀综合型大学）：
1. UNIVERSITY OF BASQUE COUNTRY
相似度距离: 2.596
2. UNIVERSITY OF ELECTRONIC SCIENCE & TECHNOLOGY OF CHINA
相似度距离: 3.077
3. SOUTHWEST UNIVERSITY - CHINA
相似度距离: 3.635
4. UNIVERSITY OF MILANO-BICOCCA
相似度距离: 4.250
5. SOUTHEAST UNIVERSITY - CHINA
相似度距离: 4.348
```