中山大学 MA7259《机器学习》期中作业

美国高校招生与毕业率统计数据的分析和预测

何瑞杰 25110801

摘　要

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

1. 研究背景与目的

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

2. 探索性数据分析

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

表 1　`college` 数据集的各数据域名称及含义

数据域名称	含义
`Private`	私立或公立大学
`Apps`	收到的申请数量
`Accept`	录取的申请数量
`Enroll`	入学新生数量
`Top10perc`	高中排名前 10% 的新生百分比
`Top25perc`	高中排名前 25% 的新生百分比
`F.Undergrad`	全日制本科学生人数
`P.Undergrad`	非全日制本科学生人数
`Outstate`	外州学生学费
`Room.Board`	食宿费用
`Books`	预估书本费用
`Personal`	预估个人开销
`PhD`	拥有博士学位的教师比例
`Terminal`	拥有最高学位的教师比例
`S.F.Ratio`	师生比例
`perc.alumni`	捐赠校友比例
`Expend`	生均教学支出
`Grad.Rate`	毕业率

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre

audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

下面对上述 `describe()` 的结果进行分析。

整体概述

这份数据描述了 777 所美国大学在 1995 年的各项指标。数据包含分类变量（如 Private）和数值变量。describe()函数为数值变量提供了丰富的统计信息，让我们能够快速了解数据的中心趋势、离散程度和分布形态。

2.1. 分变量详细分析

2.1.1. 学校基本属性

- Private
  - 分析： 这是一个分类变量，describe() 显示其唯一值（unique）为 2（Yes/No），其中 top 值为 True（即“Yes”），频数（freq）为 565。
  - 结论： 数据集中私立大学占绝大多数。565 所私立大学 / 777 所总数 ≈ 72.7% 的学校是私立的。

2.1.2. 招生情况

- Apps（申请数）， Accept（录取数）， Enroll（入学数）
  - 平均数： 平均每所大学收到 3001 份申请，录取 2018 人，最终有 779 人入学。
  - 统计行为：
    - 差异巨大： 三个变量的标准差（std）都非常大，几乎接近甚至超过其平均值（例如 Apps 的 std=3870 > mean=3001）。这表明不同大学的招生规模存在天壤之别。最大值（Apps max=48,094）和最小值（Apps min=81）也印证了这一点。
    - 录取与入学率： 我们可以粗略计算：
      - 平均录取率 = Accept / Apps ≈ 2018 / 3001 ≈ 67.3%
      - 平均入学率/报到率 = Enroll / Accept ≈ 779 / 2018 ≈ 38.6%
    - 分布形态： 中位数（50%）远小于平均数（均值）。例如，Apps 的中位数是 1558，但均值是 3001。这意味着有少量大学拥有极其庞大的申请量（极右偏分布），拉高了整体平均值。大部分大学的申请数集中在较低水平（一半的大学申请数少于 1558）。

2.1.3. 生源质量

- Top10perc（高中前 10%）， Top25perc（高中前 25%）
  - 平均数：平均而言，新生中有 27.6% 来自高中排名前 10% 的学生，55.8% 来自前 25% 的学生。
  - 统计行为： 分布相对均匀（标准差小于均值），Top25perc 的中位数（54%）和均值（55.8%）很接近，说明分布相对对称。而 Top10perc 的中位数（23%）低于均值（27.6%），表明存在一些顶尖生源高度集中的大学，使分布轻微右偏。

2.1.4. 学生与教师规模

- `F.Undergrad`（全日制本科）， `P.Undergrad`（非全日制本科）
  - 平均数： 平均全日制本科生为 3699 人，非全日制为 855 人。

- 统计行为：同样呈现出极大的差异（标准差很大）。全日制学生的规模分布极右偏（中位数 1707 << 均值 3699），说明少数大型大学主导了数据。非全日制学生的最大值（21,836）和标准差（1522）表明，不同大学在教学模式上差异显著。

- `S.F.Ratio`（师生比）
  - 平均数：平均师生比为 14.09（即平均每 14 名学生对应 1 名教师）。
  - 统计行为：分布相对集中（标准差 3.96），中位数（13.6）和均值（14.09）接近，大部分学校的师生比在 11.5 到 16.5 之间（25%-75% 分位数）。

2.1.5. 费用与开支

- `Outstate`（外州学费），`Room.Board`（食宿费），`Books`（书本费），`Personal`（个人开销），`Expend`（生均支出）
  - 平均数：外州学费平均为 10,441，食宿费为 4,358，书本费为 549，个人开销为 1,341，生均教学支出为 $9,660。
  - 统计行为：
    - 学费和支出差异显著：Outstate 和 Expend 的标准差非常大，表明大学的收费水平和资源投入相差悬殊。Expend 的最大值（56,233）是均值（9,660）的 5 倍多，再次印证了资源的高度不平等。
    - 固定费用相对稳定：Room.Board 和 Books 的分布相对集中，说明这些基础生活成本在不同大学间差异较小。
    - 分布形态：Outstate 和 Expend 的分布明显右偏（中位数 < 均值），说明有一小部分高学费、高支出的精英大学。

2.1.6. 师资力量

- PhD（拥有博士学位教师比例），Terminal（拥有终极学位教师比例）
  - 平均数：平均 72.7% 的教师拥有博士学位，79.7% 拥有终极学位（通常指本领域的最高学位，如博士、艺术硕士 MFA 等）。
  - 统计行为：分布较为集中，大部分大学的教师博士学位比例在 62% 到 85% 之间（25%-75% 分位数），师资队伍整体素质较高且在不同大学间相对均衡。Terminal 的比例普遍高于 PhD，这是合理的。

2.1.7. 学校成果与声誉

- perc.alumni（捐赠校友比例）
  - 平均数：平均 22.7% 的校友会捐款。
  - 统计行为：分布较为分散，不同大学的校友捐赠文化和忠诚度差异很大。

- Grad.Rate（毕业率）
  - 平均数：平均毕业率为 65.5%。
  - 统计行为：
    - 异常值：最大值 118% 是一个明显的异常值，因为毕业率不可能超过 100%。这可能是数据录入错误，或者计算方法特殊（例如包含了超期毕业的学生），需要进一步核查。
    - 分布：剔除异常值影响，毕业率的分布相对正常，中位数（65%）与均值（65.5%）基本一致，表明分布大致对称。但毕业率本身在不同大学间差异不小（标准差 17.2%）。

2.1.8. 总结

1. 数据构成：数据集以私立大学为主（72.7%）。
2. 极度不均衡：大学在规模（申请数、学生人数）和资源（学费、支出）上表现出极端的差异，存在明显的“头部效应”。大部分统计量的平均值都被少数大型/富裕的大学拉高，中位数通常能更好地代表“典型”大学的情况。
3. 招生漏斗：从申请到录取再到入学，数量大幅减少，平均入学率仅为 38.6%。
4. 潜在数据问题：Grad.Rate 存在超过 100% 的异常值，需要在后续分析中处理。
5. 相对稳定的指标：师生比（S.F.Ratio）、师资博士比例（PhD, Terminal）、基础生活成本（Room.Board, Books）等指标在不同大学间的分布相对集中。

3. 方法与模型

3.1. 数据预处理

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

3.2. 线性回归

线性回归可以用直线拟合、向数据矩阵 $\boldsymbol{X}$ 的列空间做正交投影、极大似然估计等视角进行理解和解释。记数据集为 $D=\left\{\left(\boldsymbol{x}_{\{n\}}, y_n\right)\right\}_{n=1}^{N}$，定义对标签的预测函数为下面的线性形式

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D = \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix} \tag{1}$$

其中 $\boldsymbol{x} \in \mathbb{R}^D$，$\boldsymbol{w} \in \mathbb{R}^{D+1}$. 这是线性回归的基本形式。若将特征和对应的标签堆叠起来，即

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{2}$$

则线性回归的预测结果为 $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\omega}$，我们要求预测结果 $\hat{\boldsymbol{y}}$ 距离真实目标 $\boldsymbol{y}$ 越近越好，这就得到了线性回归的目标函数：

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 = \frac{1}{2}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2 = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\|_2^2 \tag{3}$$

这是一个拥有光滑凸目标函数的优化问题，若 $\boldsymbol{X}^\top\boldsymbol{X}$ 可逆，令 $J(\boldsymbol{\omega})$ 的梯度为零，我们能得到其解析解

$$\boldsymbol{\omega}^* = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{4}$$

实际操作中，我们会在目标 $J(\boldsymbol{\omega})$ 中加入正则项 $R(\boldsymbol{\omega})$，通过对权重向量进行软限制的方式防止过拟合。修改过后的目标函数为

$$J(\boldsymbol{\omega}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\|_2^2 + \lambda R(\boldsymbol{\omega}), \tag{5}$$

其中 $\lambda \geqslant 0$ 是需要人为给定的权重超参数。常用的正则项可以是权重向量的 L2 范数（这将得到岭回归）或 L1 范数（这将得到 LASSO 回归）。若选取的 L2 范数，问题总是存在整洁的解析解：

$$\boldsymbol{\omega}^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{6}$$

从计算上看，这会使得括号中的矩阵可逆，此时该模型总是有解析解。

线性回归还有结合了核函数的变体形式。具体而言，考虑一列核函数 $\{\varphi_i\}_{i=1}^M$ 作用于每个样本 $\boldsymbol{x}_n$，得到新的特征矩阵

$$\boldsymbol{X} = \begin{bmatrix} \varphi_1(\boldsymbol{x}_1) \\ \vdots \\ \varphi_M(\boldsymbol{x}_N) \end{bmatrix} \tag{7}$$

这在数据科学中被称为特征工程。在得到了新的特征 $\boldsymbol{X}$ 后，接下来的线性回归操作和上面相同。

3.3. CART 决策树

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

4. 实验与结论

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

参考文献

[1] QUVA-Lab, 《E(2)-Equivariant CNNs Library for Pytorch》. 见于: 2025 年 10 月 15 日. [在线]. 载于: https://github.com/QUVA-Lab/e2cnn

[2] 《PySeagull Documentation》. 见于: 2025 年 10 月 15 日. [在线]. 载于: https://pyseagull.readthedocs.io/

[3] M. Weiler 和 G. Cesa, 《General E(2)-Equivariant Steerable CNNs》, CoRR, 2019, [在线]. 载于: http://arxiv.org/abs/1911.08251

[4] G. Cesa, L. Lang, 和 M. Weiler,《A Program to Build E(N)-Equivariant Steerable CNNs 》, 收入 International Conference on Learning Representations (ICLR), 2022.

[5] M. Weiler 和 G. Cesa,《General E(2)-Equivariant Steerable CNNs》, 收入 Conference on Neural Information Processing Systems (NeurIPS), 2019.

[6] J. Ho, A. Jain, 和 P. Abbeel, 《Denoising Diffusion Probabilistic Models》, CoRR, 2020, [在线]. 载于: https://arxiv.org/abs/2006.11239

[7] Y. Song, S. Garg, J. Shi, 和 S. Ermon, 《Sliced Score Matching: A Scalable Approach to Density and Score Estimation》, CoRR, 2019, [在线]. 载于: http://arxiv.org/abs/1905.07088

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, 和 B. Poole, 《Score-Based Generative Modeling through Stochastic Differential Equations》, CoRR, 2020, [在线]. 载于: https://arxiv.org/abs/2011.13456

[9] Y. Song 等,《Kuramoto Orientation Diffusion Models》, 2025. [在线]. 载于: https://api.semanticscholar.org/CorpusID:281411420

[10] Z. Geng, M. Deng, X. Bai, J. Z. Kolter, 和 K. He, 《Mean Flows for One-step Generative Modeling》, CoRR, 2025, doi: 10.48550/ARXIV.2505.13447.

5. 附录