

实验报告

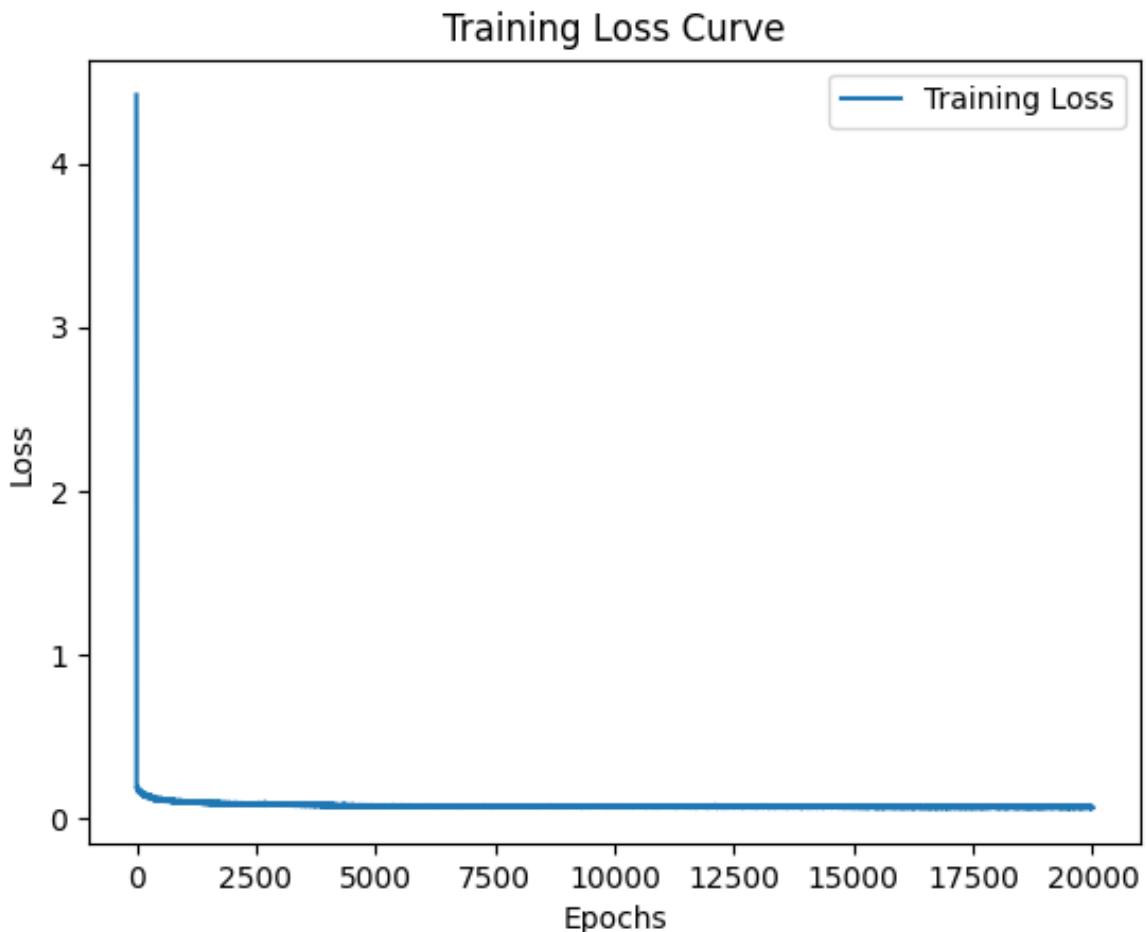
思路

设计了一个深度神经网络来预测房价，由于数据量其实比较少，只有几千条，所以神经网络不要太深，很容易训练不开。

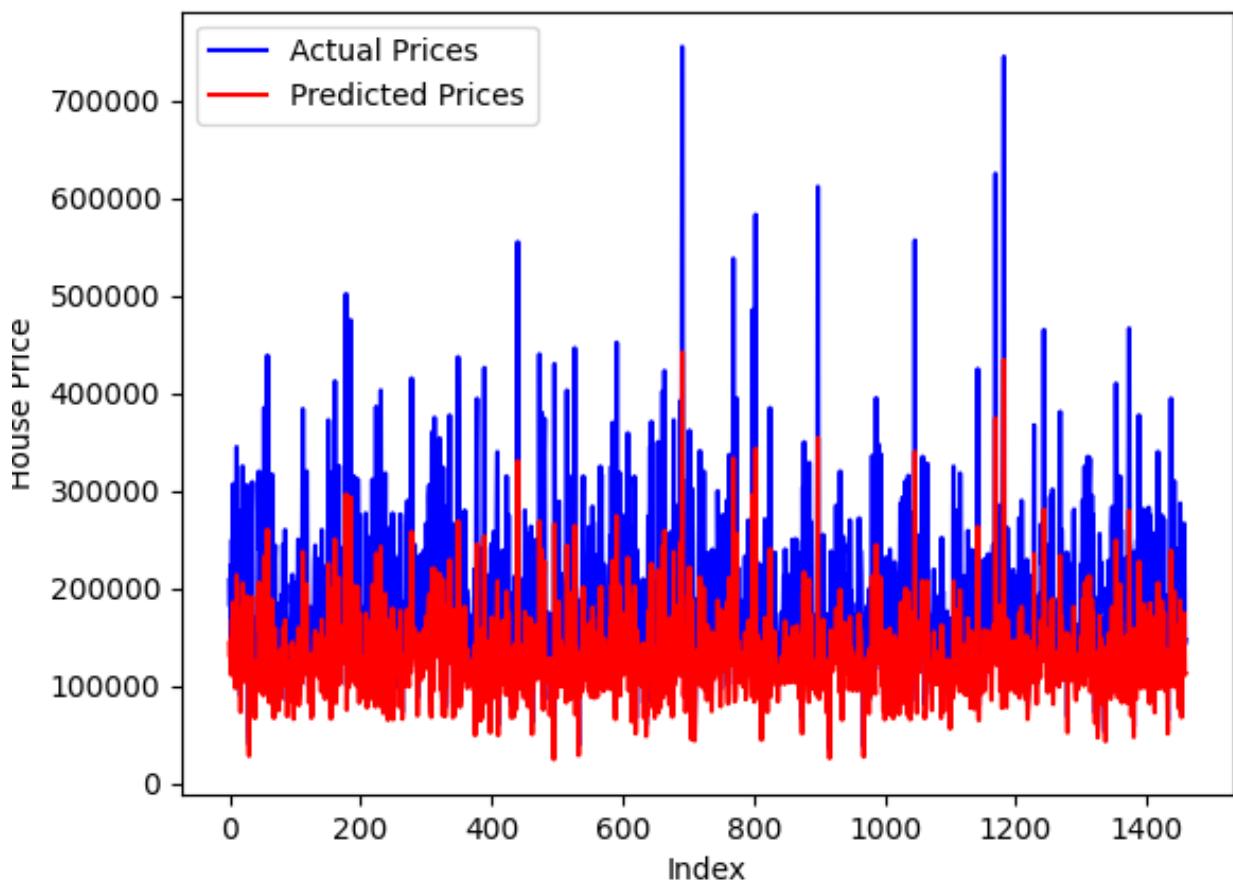
一开始使用 MSE 来作为损失函数，但 MSE 所产生的损失太大，非常容易导致训练不稳定、梯度爆炸等等问题。

后来使用 RMSE 作为损失函数，即标签价格和预测价格都先加一再取自然对数，然后相减再开根号作为损失值。这样做的目的在于可以控制损失值在小区间内，避免出现训练不稳定、梯度爆炸的问题。这么做也有缺点，由于自然对数函数会压缩高输入值，所以使得房价的预测值几乎都偏低于真实值。

使用 RMSE 函数训练了 2000 个 epoch，稳定下降。



最后的验证结果比对如下。



提交截图

The screenshot shows the Kaggle interface for the 'House Prices - Advanced Regression Techniques' competition. The left sidebar includes a navigation menu with 'kaggle' selected, 'Create', 'Home', 'Competitions' (which is highlighted), 'Datasets', 'Models', 'Benchmarks', 'Code', 'Discussions', 'Learn', and 'More'. Below this is 'Your Work' with links to 'VIEWED' (House Prices - Advanced Regression Techniques, Tyler, the Creator Data, Osic-Multiple-Quantile, Simple quant features ...), 'Gemma', and 'Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic.' At the bottom right are 'Learn more' and 'OK, Got it.'

The main content area shows the competition details: 'KAGGLE - GETTING STARTED PREDICTION COMPETITION - ONGOING', 'House Prices - Advanced Regression Techniques', 'Predict sales prices and practice feature engineering, RFs, and gradient boosting', and a 'Submit Prediction' button. Below this is a thumbnail image of houses. The 'Submissions' tab is selected, showing two successful submissions:

- test_predictions.csv (Complete - 2m ago) with a Public Score of 0.34743
- test_predictions.csv (Complete - 3h ago) with a Public Score of 0.35409

截屏显示了 Kaggle 网站上的“House Prices - Advanced Regression Techniques”竞赛的 Leaderboard 页面。

左侧导航栏包含以下链接：

- Home
- Competitions (当前选中)
- Datasets
- Models
- Benchmarks
- Code
- Discussions
- Learn
- More

右侧顶部有搜索栏、提交预测按钮和更多选项按钮。

Leaderboard 标签页显示了以下竞赛结果：

Rank	User	Score	Submissions	Time
4261	Pranav Dhangra	0.34451	1	1mo
4262	桃生蝶梦	0.34467	1	23d
4263	ttvvv7	0.34480	1	1mo
4264	Dmitriy Mineev	0.34516	1	1mo
4265	Yashrajsinh Vala	0.34526	2	1d
4266	NO EYE DEER	0.34709	1	16h
4267	wolfgangPauli	0.34743	2	14s
4268	Bishal Adhikari	0.34836	1	11d
4269	Cheng Methanol	0.34852	2	1mo
4270	QuintinCovington2003	0.34952	1	21d
4271	Jiwon Lim	0.34959	2	2mo

中间部分显示了“Your Best Entry!”通知，内容为：“Your most recent submission scored 0.34743, which is an improvement over your previous score of 0.35409. Great job!”，并有一个“Tweet this”按钮。

底部状态栏显示：“Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic.”，带有“Learn more”和“OK, Got it.”按钮。

收获与反思

1. 应根据不同的数据分布使用不同的损失函数
1. 将字符串类型的数据转换为分类型数据
2. 模型不应太大，因为数据量有限，容易训练不开