房价预测的深度学习方法

刘倡源

中国人民大学数学学院

2025 年 6 月 4 日

中國人民大學
RENMIN UNIVERSITY OF CHINA

预测变量概述

离散数据

- 相比于之前从文本信息中获得的分类变量，仅包含给定的分类变量。
- 缺失值另起一类。

连续数据

- 相比于之前从文本信息中获得的数值变量，仅包含给定的数值变量。
- 缺失值以均值填补。

文本数据

- 所有非离散数据、连续数据的均算作此。

整体处理思路

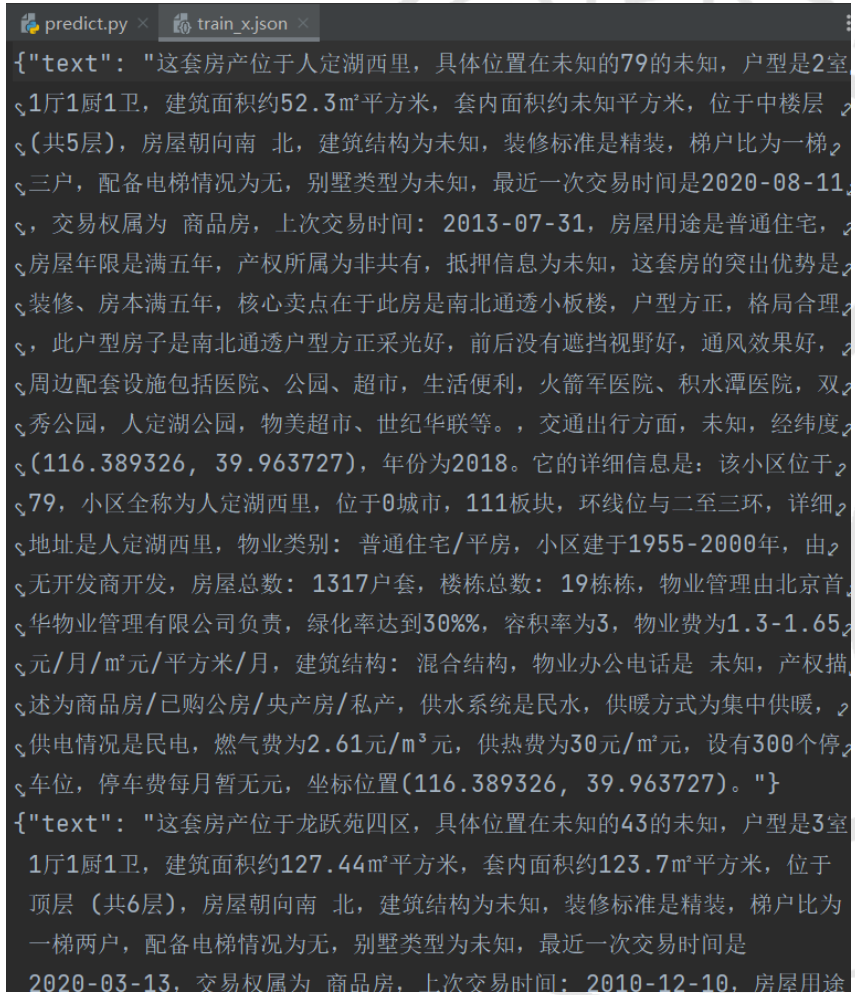
不同数据“分而治之”

- 离散数据过嵌入层（Embedding）。
- 连续数据过多层感知机（MLP）。
- 语言数据过 ~~预训练模型~~
  Embedding + 池化提取语义表示。

- 特征拼接后过多层 MLP 进行特征交互。
- 最后过一个 Linear 输出房价预测值。
- 进行多次循环，每循环十次后进行预测（Epoch=500）。

其它模型尝试

- 仍加入花哨的特征工程。
- 使用预训练模型 bert-base-chinese。
- 使用轻量级版本或冻结参数。
- 将所有信息总结成一段话，完全作为语言数据处理。



文本处理流程

分词与词频统计

- 使用 jieba 分词库对每个文本进行分词处理。
- 统计每个词在整个数据集中的出现频率。

构建词汇表

- <PAD>（索引 0）：填充标记，用于统一序列长度。
- <UNK>（索引 1）：未知词标记，处理未见过的词汇。
- 其余数据集中出现频率最高的 9998 个词。

深度学习算法

- 损失函数：均方误 (MSE)$L = \frac{1}{N}\sum(y_i - \bar{y}_i)^2$。
- 优化器：Adam 优化器，自适应调整各参数学习率。
- 正则化：两层 0.3 比率的 Dropout（随机失活）。

预测机制

- 每 10 个 epoch 进行完整评估：
- 训练集、验证集、6 折交叉验证评估（MAE，$RMSE$，R^2）。
- 保存每个评估点的模型权重和预测结果。

预测结果展示

表 1: Metrics Table(MAE)

Epoch	In Sample	Out of Sample	Cross-Validation	Datahub Score
200	214285	266786	257178	79.065
300	173752	200626	208869	83.919
400	157616	171939	176809	84.978
500	160415	177472	181869	84.663

表 2: Metrics Table(RMSE)

Epoch	In Sample	Out of Sample	Cross-Validation	Datahub Score
200	454763	519705	474975	79.065
300	430640	484190	473660	83.919
400	371939	433277	446786	84.978
500	400341	450029	468541	84.663