

# 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。

## 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。

```
predict.py  train_x.json
{"text": "这套房产位于人定湖西里，具体位置在未知的79的未知，户型是2室1厅1厨1卫，建筑面积约52.3㎡平方米，套内面积约未知平方米，位于中楼层（共5层），房屋朝向南 北，建筑结构为未知，装修标准是精装，梯户比为一梯三户，配备电梯情况为无，别墅类型为未知，最近一次交易时间是2020-08-11，交易权属为 商品房，上次交易时间：2013-07-31，房屋用途是普通住宅，房屋年限是满五年，产权所属为非共有，抵押信息为未知，这套房的优势是装修、房本满五年，核心卖点在于此房是南北通透小板楼，户型方正，格局合理，此户型房子是南北通透户型方正采光好，前后没有遮挡视野好，通风效果好，周边配套包括医院、公园、超市，生活便利，火箭军医院、积水潭医院，双秀公园，人定湖公园，物美超市、世纪华联等。，交通出行方面，未知，经纬度(116.389326, 39.963727)，年份为2018，它的详细信息是：该小区位于79，小区全称为人定湖西里，位于0城市，111板块，环线位与二至三环，详细地址是人定湖西里，物业类别：普通住宅/平房，小区建于1955-2000年，由无开发商开发，房屋总数：1317户套，楼栋总数：19栋栋，物业管理由北京首华物业管理有限公司负责，绿化率达到30%，容积率为3，物业费为1.3-1.65元/月/㎡元/平方米/月，建筑结构：混合结构，物业办公电话是 未知，产权描述为商品房/已购公房/央产房/私产，供水系统是民水，供暖方式为集中供暖，供电情况是民电，燃气费为2.61元/㎡元，供热费为30元/㎡元，设有300个停车位，停车费每月暂无元，坐标位置(116.389326, 39.963727)。"}
{"text": "这套房产位于龙跃苑四区，具体位置在未知的43的未知，户型是3室1厅1厨1卫，建筑面积约127.44㎡平方米，套内面积约123.7㎡平方米，位于顶层（共6层），房屋朝向南 北，建筑结构为未知，装修标准是精装，梯户比为一梯两户，配备电梯情况为无，别墅类型为未知，最近一次交易时间是2020-03-13，交易权属为 商品房，上次交易时间：2010-12-10，房屋用途
```

## 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。

- 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	Datashub 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	Datashub 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