Assignment 3 实验报告

姓名: **高梓洋**

学号: **2001213232**

问题背景

近年来，各地房价都出现了不同程度的增长，为了应对高房价问题，政府频繁地针对房地产市场实施调控。二手房交易作为房地产市场的重要组成部分，是许多房产中介的经营重点，许多中介通过低买高卖等不良行为获取高额利润，哄抬价格，对于购买二手房的消费者而言，了解二手房价格显得尤为迫切，同时，二手房价格也是市场监管机构的关注重点，因此，对于二手房价格预测的研究极具现实意义。

本实验使用“链家-住房交易微观数据(2002-2018 年)”数据，考虑各种影响二手房价格的因素，运用多种模型对二手房价格进行预测。

数据来源

本实验所用数据的原始来源为链家官网（https://bj.lianjia.com/），分析数据由马克数据网-马克社区用户 "学渣爱计量" 提供，本数据涵盖了链家平台上 2002-2018 年在北京市内的二手房成交信息，原始数据需要使用 stata 统计软件打开，共 318852 个样本，26 个变量，变量包括：（1）地区信息：经纬度、街道、社区类别等；（2）成交信息：成交时间与价格等；（3）房产特征：如卧室数量、客厅数量、厨卫数量、修建年份、楼层数、是否有电梯等等。

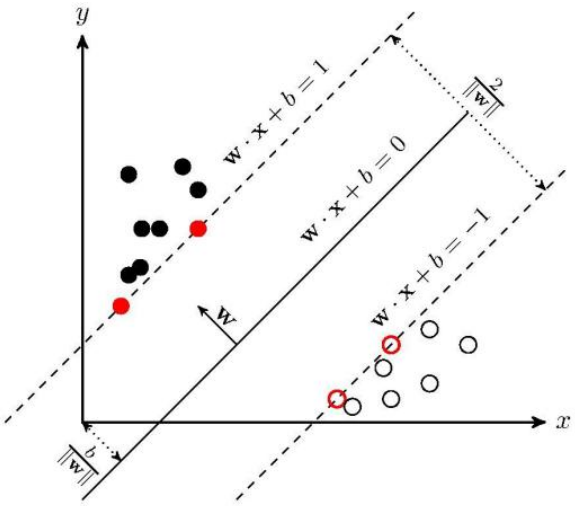
使用 stata 统计软件打开原始数据后，可做简单的数据清洗（代码见文末附注），考虑到数据体量，在 stata 软件中随机抽取 5%的样本用作后续实验。

程序说明

本实验采用的算法可以粗略分为支持向量机和集成学习两类。

1. 支持向量机

支持向量机算法是找到集合边缘上的若干数据（称为支持向量（Support Vector）），用这些点找出一个平面（称为决策面），使得支持向量到该平面的距离最大。本实验中分别使用线性核函数支持向量机和多项式核函数支持向量机。



算法示意图如上图所示，该算法也可以表述为：

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t.\ \ y_i\left(w \cdot x_i + b\right) \geq 1,\ i = 1, 2, \ldots, N$$

2. 集成学习

随机森林是一种经典的集成学习模型，是弱分类器为决策树时的 Bagging 算法。随机森林算法不仅对样本进行 Bootstrap 采样，对每个 Node 调用生成算法时都会随机挑选出一个可选特征空间的子空间作为该决策树的可选特征空间；同时，生成好个体决策树后不进行剪枝，而是保持原始的形式。除了和一般 Bagging 算法那样对样本进行随机采样以外，随机森林还对特征进行了某种意义上的随机采样。

极端随机树算法与随机森林算法十分相似，都是由许多决策树构成，但是该算法不对样本进行 Bootstrap 采样，而是使用所有的样本，只有特征是随机选取的。

梯度提升决策树(GBDT)也是一种基于集成思想的模型，不同于随机森林用到的 Bagging 算法，GBDT 使用 Boosting 算法，Boosting 会在迭代的每一步构建弱学习器弥补原有模型的不足。GBDT 每次迭代生成一颗新的决策树，计算损失函数在每个训练样本点的一阶导数和二阶导数，然后通过贪心策略生成新的决策树，并将新生成的决策树添加到模型中。GBDT 中的 Gradient Boost 就是通过每次迭代时构建一个沿梯度下降最快的方向的学习器，可以通过设置不同的损失函数来处理各类学习任务（多分类、回归等）。

实验结果

本次实验中按照 3:1 划分训练集与测试集，使用默认评估值(R square)、均方误差(Mean Squared Error, MSE)、绝对平均误差(Mean Absolute Error, MAE)来评估模型表现。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$$

实验结果如下表所示：

类别	模型名称	R square	MSE	MAE
支持向量机	线性核函数支持向量机	0.7544	12409.8291	70.0020
	多项式核函数支持向量机	0.7544	12409.8291	70.0020
集成学习	随机森林	0.9172	4182.1075	37.0579
	极端随机树	0.9172	4182.1075	37.0579
	梯度提升决策树	0.9172	4182.1075	37.0579

由上表可知，集成学习模型在本数据集上的表现明显好于支持向量机算法。值得注意的是，虽然三种集成学习算法在 R square、MSE、MAE 上的差异不大，但是三者的默认评估值不同（随机森林的默认评估值为：0.9172344989347025；极端随机树的默认评估值为：0.9263533530799376；梯度提升决策树的默认评估值为：0.9202579532221233），暗示着在本样本中极端随机树表现更好，由于本样本是从原有样本中随机抽取 5%得到的，样本量相对较小，在更大的样本（原始样本清洗后包含超过 29 万条数据）中哪种模型的表现更好还需要进一步探究。

附注：

打开原始数据并简单清理的 stata 软件代码如下：

```
use "E:\研二下学期\Python 大数据分析原理与应用\assignment_3_大作业\链家数据\链家数据.dta"

*****数据清洗，删除异常值
tab constructiontime
drop if constructiontime=="1"
drop if constructiontime=="0"
drop if constructiontime=="Î´Ö¤"
tab buildingtype
drop if buildingtype=="nan"

****随机抽取 5%的样本
sample 5
```