Ames, Iowa 地区房价的分析预测

单铭铭

摘要

如果你有一套房子想出售，给它定价就是一件很重要的事情。但是感性地定价可能会错误地估计房子的价格，或高或低。也许你会参考一些类似的房子，但是完全类似的房子可能并不存在。于是，我们就会利用现有的数据，来尽可能最好的预测房子的价格。本文利用 Ames，Iowa 地区个人住宅房的买卖记录，来建立模型，给定房子的各种属性，预测房子的价格。

数据描述

原数据集来自 the Ames City Assessor's Office，经过 Dean De Cock 的整理，将一些只有专家才会关心的变量去掉了，留下了普通人会关心的一些变量，才得到我们在这里要使用的数据集：2006-2010 年里的 1460 个买卖记录，每个记录有 81 个字段，并且交易都发生在 Ames。还将同一套房子在几年里多次交易的记录，选择只保留最近一次的交易记录，防止在这个房子上的权重增加 (同一套房子的买卖价格总是类似的)。

我们关心的是房子的价格，即 **SalePrice** 字段。其他 80 个字段，分为

- 23 个**无序名义**变量（nominal）
- 23 个**有序名义**变量（ordinal）
- 14 个**离散数值**变量（discrete）
- 20 个**连续数值**变量（continuous）

无序名义变量例如 **Street**，表示通向房屋的路的类型，有碎石路（gravel）和平坦的公路（paved）两种类型，不能排序。有序名义变量例如 **OveralQual**，表示房屋整体质量评分，从 Very Poor 到 Very Excellent 共 10 个评级，可以比较排序。离散数值变量例如 **KitchenAbvGr**，表示合格的厨房

的个数，是一个大于等于 0 的离散数值。连续数值变量例如 **LotArea**，表示占地面积，可以是大于 0 的数值，连续。

这和传统的回归分析不一样，涉及到了大量的自变量，因此变量的筛选变得异常重要。