Statistics and Application 统计学与应用, 2016, 5(3), 299-304
Published Online September 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.53030

Hans汉斯

Factor Analysis of Housing Price Based on Boosting Regression Tree

—Taking Boston as an Example

Jia Sheng, Dongdong Pan

School of Mathematics and Statistics, Yunnan University, Kunming Yunnan
Email: ddpan@ynu.edu.cn

Received: Sep. 6th, 2016; accepted: Sep. 22nd, 2016; published: Sep. 29th, 2016

Abstract

Housing price is a very important index which can reflect the economic and social development level and situation of a certain region or city. It is of great theoretical value and practical meaning to study important factors influencing housing price as well as their influence patterns and magnitude. Boosting regression tree has been recently developed as one of the most prevalent nonparametric modeling methods in the fields of machine learning, which has desirable properties such as high efficiency as well as easy-interpretation. In this paper, we take the housing price data in Boston as an example and try to analyze factors determining housing price based on Boosting Regression Tree method. We identify some relatively significant factors by comparing their relative importance in the model and also investigate their influence patterns. Results in this paper could be reasonably extended to housing price researches of some Chinese first-tire cities.

Keywords

Regression Tree, Boosting, Housing Price, Factor Analysis

基于增强回归树的房价影响因素分析

—以波士顿地区为例

盛　佳，潘东东

云南大学数学与统计学院，云南　昆明

文章引用: 盛佳, 潘东东. 基于增强回归树的房价影响因素分析[J]. 统计学与应用, 2016, 5(3): 299-304.
http://dx.doi.org/10.12677/sa.2016.53030

Email: ddpan@ynu.edu.cn

收稿日期：2016年9月6日；录用日期：2016年9月22日；发布日期：2016年9月29日

摘　要

房价是反映一个地区经济社会发展水平和状况的重要指标，对其影响因素以及影响的方式和程度进行探究具有理论价值和现实意义。增强回归树是近年来机器学习领域备受关注和推崇的一种非参数建模分析方法，具有建模效率高、模型结果易于解读等优势。本文以美国波士顿地区的历史房价数据为例，采用增强回归树方法来探寻该地区房价的主要影响因素，并比较不同因素在回归树中的相对影响强度。本文得出的结论可为我国某些中心城市的房价调控政策提供参考。

关键词

回归树，增强法，房价，因素分析

1. 引言

房价是各地区的重要经济指标，寻找影响房价的重要因素并分析这些因素对房价的影响方式和程度具有现实意义。由于影响房价的因素众多并且不同的因素对房价的影响各不相同，因此用传统的参数建模方法难以确定合适的模型来反映房价对众多影响因素的依赖关系。回归树算法由于其高效且易实现等优点近来成为机器学习和数据挖掘领域最流行的算法之一，它具有计算量小，模型易于解读且不需要复杂的数据准备等优势。而增强回归树通过将一列回归树进行组合极大的提高了回归树的预测准确度。本文用增强回归树的方法分析了波士顿地区的房价数据，找到了几个对该地区房价影响最大的因素并通过偏依赖曲线的方式分析了这几种因素对房价的影响方式。

本文的组织方式如下，第二节介绍了增强回归树方法的基本原理及优势；第三节简单介绍了本文要进行分析的波士顿房价数据；第四节是模型的参数选择及模型结果的分析；最后一节是总结。

2. 回归树与增强法

增强法的思想最初来自于 Freund 和 Schaprie (1997)提出的“AdaBoost.M1”分类算法[1]，用于提高分类树在二分类问题中的表现，Friedman (2000)将增强法的思想推广到了回归问题并将它与回归树的方法结合，提出了增强回归树的算法[2]。增强回归树的基本思想是对不断调整的训练数据重复应用回归树算法，得到一列回归树，然后将这一列回归树进行加权平均得到一个最终的回归树。理论研究与实际应用均表明增强回归树可以明显的提高回归树的预测准确性。

在理解增强回归树算法之前，首先来回顾一下回归树的基本概念。回归树是数据挖掘和机器学习领域应用最广的算法之一，它在拟合数据时，先将预测变量 X 的联合空间划分成互不重叠的 J 个小区域 R_j，称作树的终端节点(或叶子)；然后为每一个小区域拟合一个常数 γ_j 作为这个小区域内响应变量 y 的预测值：

$$x \in R_j \Rightarrow f(x) = \gamma_j$$

因此一个回归树可以形式的表示为：

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j), \quad \Theta = \{R_j, \gamma_j\}_1^J$$

回归树的两组基本参数是小区域 R_j 以及小区域上相应的常数 γ_j，将其统一记做 Θ，参数估计的标准是：

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_j \in R_j} L\left(y_i, \gamma_j\right)$$

其中 $L(\cdot)$ 是损失函数。在回归树中，最常用的损失函数是平方损失函数 $L\left(y, f\left(x\right)\right) = \left(y - f\left(x\right)\right)^2$，此时回归树的参数是使训练样本残差平方和最小的那一组。

同其它几种流行的数据挖掘算法相比，回归树有计算快、可解读性强(如果叶子数 J 比较小)、对预测变量的单调变换具有不变性等优点。同时树对异常值不敏感，且树在生成过程中可以自动进行变量选择。由于以上优点，树可以被称为“off the shelf”方法，即它可以直接用于数据处理而不需要进行耗时的数据预处理。但回归树的一个主要缺点是预测不够准确。我们知道在平方误差损失函数下，均方误差可以分解为：$MSE = Var + Bias^2$。回归树预测不够准确主要是因为它的方差比较大，而不是因为偏差。

增强法通过对回归树进行加权平均，显著降低了树的方差，从大幅提高了树的预测准确性。增强回归树是 M 个回归树通过加法模型组合在一起的，其一般形式为：

$$f_M\left(x\right) = \sum_{m=1}^{M} T\left(x; \Theta_m\right)$$

其中每一个树的参数估计标准为：

$$\hat{\Theta}_m = \arg\min_{\Theta} \sum_{i=1}^{N} L\left(y_i, f_{m-1}\left(x_i\right) + T\left(x_i; \Theta_m\right)\right)$$

平方损失函数下：$L\left(y_i, f_{m-1}\left(x_i\right) + T\left(x_i; \Theta_m\right)\right) = \left[\left(y_i - f_{m-1}\left(x_i\right)\right) - T\left(x_i; \Theta_m\right)\right]^2$，故此时 $T\left(x; \hat{\Theta}_m\right)$ 就是在平方损失下对上一步的残差拟合效果最好的回归树。

实际应用中，为进一步提高增强法的表现，往往对增强回归树进行进一步的调整和改进。一种方法是压缩每一个子树对最终回归树的贡献：

$$f_m\left(x\right) = f_{m-1}\left(x\right) + v\sum_{j=1}^{J} \gamma_{jm} I\left(x \in R_{jm}\right) = f_{m-1}\left(x\right) + vT\left(x; \Theta_m\right)$$

其中 $0 \le v \le 1$ 称为压缩系数，也称为学习效率。另一种改进方法是引入随机性，在每个子树的生成过程中，从训练数据中不放回的抽取样本的一部分 η 作为这棵子树的训练样本而不是将全部的训练数据用于生成新的树，通常取 $\eta = 1/2$。这种做法不仅能降低计算量，实际表明它还能大幅提高增强法的预测准确性。

3. 波士顿地区房价数据

本文以美国波士顿地区的房价数据[3]为例分析研究对房价产生重要影响的因素以及这些因素的影响方式。该数据集中包含 506 个样本，14 个变量，每一个变量的名称以及其意义见表 1。我们将自住房中位数房价(MEDV)作为我们关心的响应变量 y，其它 13 个变量作为预测变量。我们通过增强回归树来分析其它变量对房价的影响。通过分析，我们想要知道哪些因素对房价的影响最强，以及这些因素是通过什么方式影响房价的。

4. 建模及结果分析

本文利用 R 中的增强法软件包 gbm 来进行建模，gbm 提供了强大的增强回归树建模功能。通过上一节的讨论，可以发现增强回归树中共有 4 个参数需要在建模时加以选择。分别是每棵树的叶子数 J、树的规模(迭代次数) M，学习效率 v 以及每一步抽样比例 η。通常的做法是事先确定 J，v 和 η，然后将迭代次数 M 作为主要的调整参数。

Table 1. Boston area house price data variable
表 1. 波士顿地区房价数据变量

变量名	意义	变量名	意义
CRIM	人均犯罪率	DIS	到市中心加权距离
ZN	大面积土地的比例	RAD	到高速路的方便指数
INDUS	非商业面积比例	TAX	每$10,000 的税率
CHAS	是否接近 Charles 河	PTRATIO	学生教师比例
NOX	氮氧化物浓度	BLACK	黑人比例指数
RM	每房平均屋子数目	LSTAT	低阶层人的比例
AGE	1940 年前自住房的比例	MEDV	自住房中位数房价

参数 J 控制了回归树中交叉项的阶数，有 J 个叶子节点的回归树中预测变量的交叉项阶不高于 $J-1$。若取 $J=2$ 意味着将树中预测变量 X 对响应变量 y 的作用限制为简单地加法关系，这种简单地关系在实际问题中很少见，但实际应用中 $J>10$ 的情况也几乎不可能，而且 J 越大增强法的计算量也越大。实际应用表明增强回归树的预测表现对 J 的选择不敏感，因此一般取 $J=6$ [4]。v 通常选为 $v \prec 0.1$，在这里我们取 $v=0.01$。先令树的规模足够大，取 $M=5000$，如图 1 所示，交叉核实(CV)的方法显示大约 3100 步迭代就足够了。

为了分析哪些因素对波士顿地区房价的影响最大，我们做出增强回归树中 13 个预测变量对响应变量的相对影响强度图(图 2)。图形表明低阶层人的比例对房价影响最大，更合理的解释是房价对低阶层群体的比例影响大，而不是相反。这是符合经验的，一般来说，一个社区的房价决定了该社区居民的收入水平，而该区居民的总体收入水平也可以用于推测房价。另一个对房价影响较大的因素是每房平均屋子数目，这也是合理的，因为高档住宅一般有更多的房间。从图中可以看出以上两个因素对房价的影响最大，几乎起决定性作用，在剩下的因素中，对房价影响较大的是到市中心的加权距离，人均犯罪率以及氮氧化物浓度。图 2 也反映出大面积土地比例、是否接近 Charles 河以及到高速公路的方便指数等几种因素对房价几乎没有影响。

为了发现不同因素对房价的影响方式，我们做出响应变量 y 对几个重要预测变量的偏依赖曲线。响应变量 y 对预测变量 X_S 的偏依赖函数的定义为：

$$f_S(X_S)=E_{X_C}f(X_S,X_C)$$

其中 $X=X_S\cup X_C$ 实际应用中我们用 f_S 的估计 $\overline{f}_S$ 来代替 f_S：

$$\overline{f}_S(X_S)=\frac{1}{N}\sum_{i=1}^{N}f(X_S,x_{ic})$$

从图 3 左上方的曲线中可以发现低阶层人的比例与该社区房价为负相关，这符合直观经验，一个家庭的住房标准往往与这个家庭的收入水平呈正比关系，低收入群体无法承担高额房价所以一个社区低收入群体数量越多该地区的房价也应该越低。

图 3 右上部分曲线反映出每房平均屋子数与房价成正相关关系，屋子数越多，往往意味着房屋越高档，房价也应当越高，这也是符合直观经验的。图中曲线还反映出当屋子数小于等于 5 时房价几乎不随屋子数增加而改变，可以推测在这个区间内的房子应该都属于低端住房，所以房价波动不大，而高端住房往往意味着每个房子拥有超过 5 间房间，因此房价只有在房间数超过 5 时才开始随房间数增多而升高。

我们再来分析到市中心的加权距离对房价的影响，离市中心距离越近意味着生活越便利，所以房价

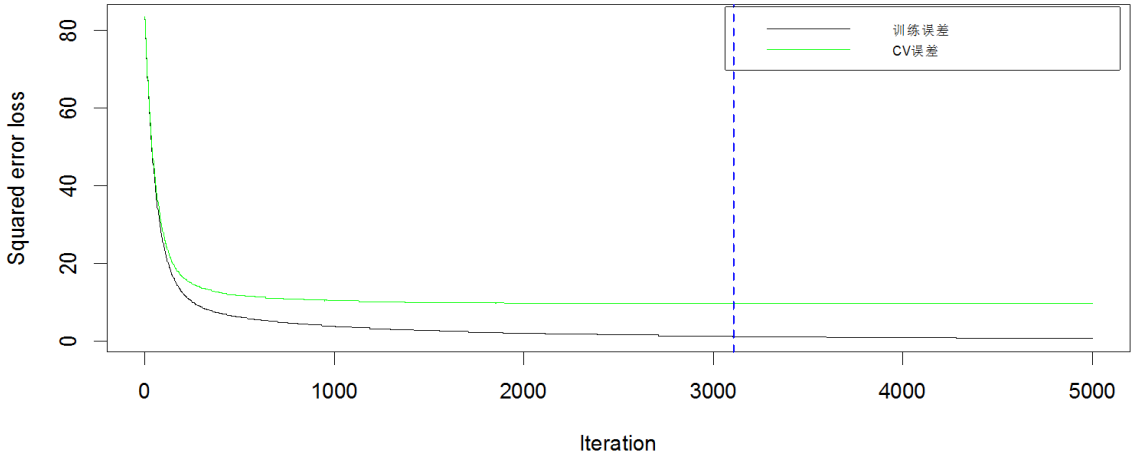

Figure 1. The error changes with the training times of the curve, the top of the CV error curve, the bottom for the training error curve
图 1. 误差随训练次数变化曲线，上方为 CV 误差曲线，下方为训练误差曲线

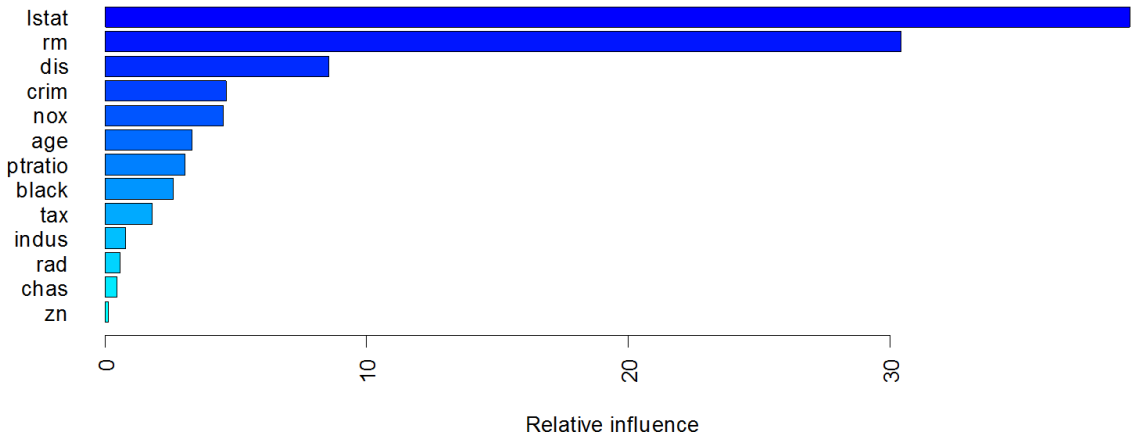

Figure 2. The relative influence of the different factors in the model
图 2. 模型中不同因素相对影响强度图

与到市中心的距离应该呈负相关关系。如图 3 中间左边曲线所示，图中的曲线大致表现出指数下降的趋势，在原点附近骤降，而在远离原点的区间内又呈现出缓慢下降趋势，最后又变为水平。这说明市中心的房价最高且比非市中心地区要高出许多几倍。而在市中心之外房价则在一定范围内与距离近似表现出负线性关系。而离市中心的距离超过一定标准后由于此时在距离上的一点差距并不会造成生活便利程度上的本质差别，所以这时距离已不再成为影响房价的因素。

图 3 中间右边曲线显示出犯罪率对房价的影响表现出奇怪的先增加而后下降的趋势，房价最高的地方犯罪率并不是最低的。这似乎与经验相违背，因为高档住宅区往往拥有更好的社会治安环境。但我们在上一步的分析中的出结论房价最高的地方在市中心，所以可以推测这里的犯罪率相对较高应该与市中心繁华商业区的存在有关。

氮氧化物的浓度对房价也有一定的影响，氮氧化物浓度反映的是一个地区的污染程度。图 3 最下方影响曲线大致表现出阶梯函数的形状且在跳跃之前有一段短暂的上升区。在非工业区，氮氧化物的浓度应该主要受汽车尾气量影响，房价最高处也不是氮氧化物浓度最低的地方，同分析犯罪率对房价的影响类似，这里较高的氮氧化物浓度应该与市中心较多的车辆有关，而曲线后半部分房价的突然下跌则应该与郊区的化工厂有关。

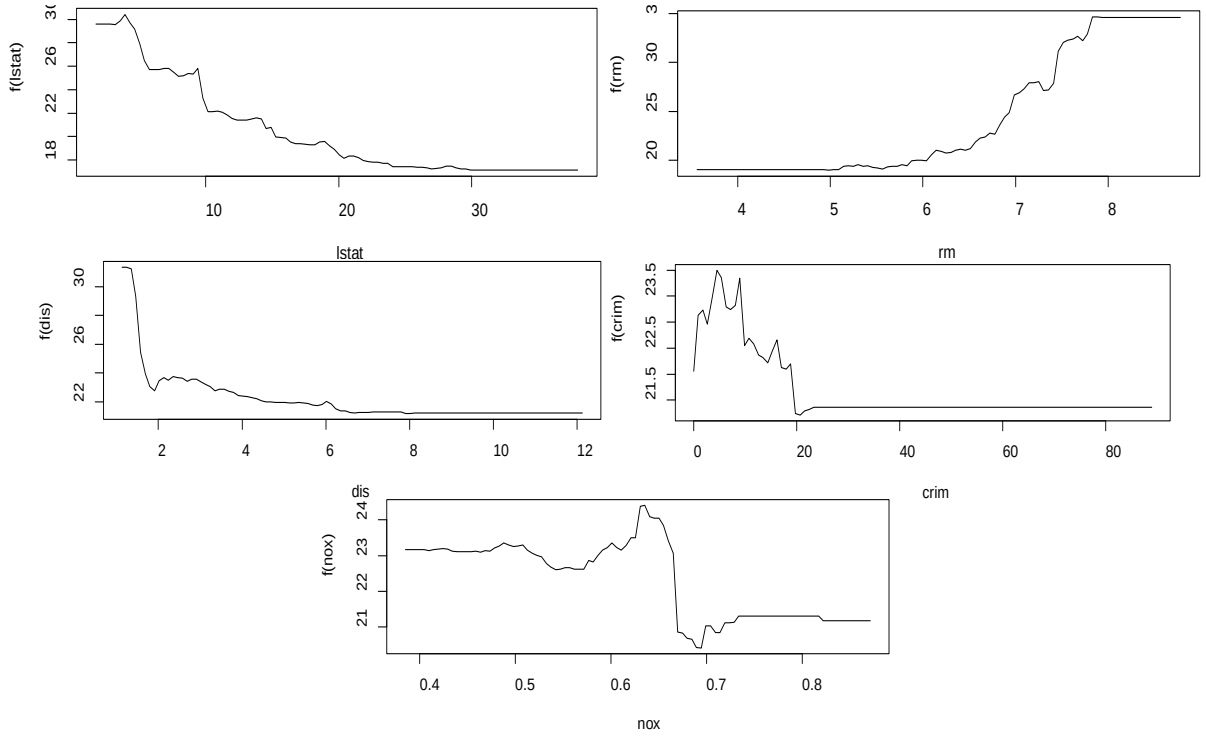

Figure 3. Partial dependence curves of several important factors of the housing prices, from left to right from the top down factors that are: low level, the proportion of people per room to room number, the average weighted distance, the center of crime rate, the concentration of nitrogen oxides

图 3. 房价对几个重要因素的偏依赖曲线，自左向右自上向下所表示的因素依次是：低阶层人的比例、每房平均屋子数、到市中心的加权距离、犯罪率、氮氧化物浓度

图 3 中间右边曲线与最下方曲线均在一定程度上偏离了经验，具体表现是房价最高处既不是犯罪率最低的地方也不是污染物浓度最低的地方。而这种反常意味着这两种因素对房价的影响较小，从相对影响强度可以看出，住房到市中心的距离这个因素对房价的影响超过了犯罪率与污染物浓度对房价的影响，而市中心较高的氮氧化物浓度与相对较高的犯罪率应该是后两条曲线表现出反常趋势的原因。

5. 总结

本文用增强回归树的方法以波士顿地区房价为例研究了对房价产生重要影响的因素以及这些因素对房价的影响方式。结果表明社区的低收入人群比例是与房价相关性最强的变量，其次是每个房子的房间数以及距离市中心的距离对房价影响较大。犯罪率以及氮氧化物浓度对房价也有一定的影响，但影响强度均不如前面的三个变量。本文研究的这些因素都具有一般性，因此它们对波士顿地区房价的作用方式及影响强度也具有一般性，对于研究其他地区的房价也据有一定的参考意义。

参考文献 (References)

[1] Freund, Y. and Schapire, R. (1997) A Decision-Theoretic Generalization of Online Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. http://dx.doi.org/10.1006/jcss.1997.1504

[2] Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting (with Discussion). *Annals of Statistics*, **28**, 337-307. http://dx.doi.org/10.1214/aos/1016218223

[3] Harrison, D. and Rubinfeld, D.L. (1978) Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, **5**, 81-102. http://dx.doi.org/10.1016/0095-0696(78)90006-2

[4] Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York. http://dx.doi.org/10.1007/978-0-387-21606-5