Kernel regression

Lorenzo Moni
Andrea Marchetti
Luca Puglisi

UNIVERSITÀ DEGLI STUDI FIRENZE

13 ottobre 2022

Introduction

Given the model $Y = f(\boldsymbol{X}) + \epsilon$, we want to predict the outcome without considering any particular form of f.
In the **kernel regression Nadaraya–Watson estimator** the modelling of $Y = E(Y|\boldsymbol{x}) = g(\boldsymbol{x})$ is given by

$$\hat{g}(\boldsymbol{x}) = \frac{\sum_i y_i \kappa_\alpha(\boldsymbol{x}, \boldsymbol{x}_i)}{\sum_i \kappa_\alpha(\boldsymbol{x}, \boldsymbol{x}_i)} = \sum_i^n \omega(\boldsymbol{x}, \boldsymbol{x}_i) y_i \tag{1}$$

Where $\kappa_\alpha(\boldsymbol{x}, \boldsymbol{x}_i) = D\left(\frac{\|\boldsymbol{x}, \boldsymbol{x}_i\|^2}{\alpha}\right)$ is a kernel fuction and α is a bandwidth.

Descriptive statistics

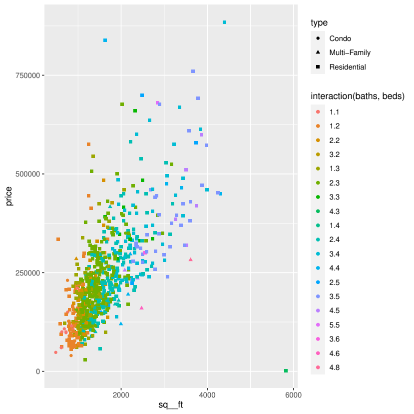
Using the dataset of 984 real estate transactions in Sacramento, California, we want to predict how house **prices** are affected by **size**, **house type**, and **number of beds** and/or **bathrooms**.

Some descriptive and graphic statistics

n. baths			
0	1	2	3+
108	180	544	153

n. beds				
0	1	2	3	4+
108	10	133	413	321

Type		
Condo	Multi-Family	Residential
54	13	917



Results

Using different kernels, we obtain the following results:

Model	Validation RMSE	Test RMSE
Normal Kernel	109 120.9	.
Epanechnikov Kernel	101 451.4	60 611.67
Uniform kernel	101 630.4	.
Normal kernel ord. 4	113 083.2	.
Epanechnikov kernel ord. 4	124 549.2	.

The prediction error is very large, we assume it is due to unobserved factors or underfitting of the model.

References

https://en.wikipedia.org/wiki/Kernel_regression

Cap. 6.1-6.4 The Elements of Statistical Learning; Trevor Hastie, Robert Tibshirani, Jerome Friedman