Application of High Dimensional Techniques to House Price Prediction

Pala, Alessandro
alessandro.pala@studenti.unipd.it

Tremaggi, Domenico
domenico.tremaggi@studenti.unipd.it

Abstract

This experiment explores two different sets of grouped variables through a Group Lasso regularized regression model and compares them to basic regression models, structural regression models, and a deep Multi-Layer Perceptron in the context of predicting house prices. The results... TO COMPLETE

1 Introduction

In order to approach the dataset, we first think about which models and variables can predict real estate prices in the real world, with the constraint of the set of variables given by the dataset. Since we handle a price prediction regression, a fair comparison of linear and non-linear basic models - be them regularized or not - would be multivariate. Hence, we first carry a synthetic application of the following models: multivariate linear model, multivariate GLM, multivariate GLM with Elastic Net, structural multivariate GLM with elastic Net. The term "structural" is used to indicate a model in which we use new variables we aggregated via functions and are included in the model's regression (much like in the field of Economics).

We then train a deep Multi-Layer Perceptron for performance comparison and we regularize it with a simple Lasso. Finally, we utilise a Group Lasso regression on two sets of groups of variables that we think may be good predictors of house prices in the real world. The first set of groups uses variables that are usually clustered together by real estate professionals when they do house price estimation. The second set of groups uses clusters of highly statistically correlated (Pearson's r) variables.

All selection for regularized models is carried through with cross-validation.

2 Dataset

The dataset contains real estate property listings, each described by a variety of attributes. Temporal attributes are: the listing date, the year the property was built, and the year of its last renovation. Physical characteristics are: the number of bedrooms and bathrooms, living area size, total lot size, number of floors, basement area, and living area above ground. Scenic qualities are represented by a waterfront view indicator, a view quality rating, an overall condition rating, and an overall grade rating. Locational data instead include zip code, latitude, and longitude. There is also some additional spatial context given by the average living area and lot sizes of the 15 nearest properties.

2.1 Pre-processing

In order to pre-process the data, we first convert the date column into a proper date object by extracting the first 8 characters (in "YYYYMMDD" format). We then replace 0 values in the "yr_renovated" column with NA, treating them as missing data. Duplicate rows are removed based on the id column. We add new variables in the dataset: total_sqft calculates total area by summing living and basement space, bath_per_bed computes the bathroom-to-bedroom ratio (defaulting to 0 if bedrooms are 0), and total_rooms sums bedrooms and bathrooms. sqft_diff_15 measures the difference between the property's living area and that of its 15 nearest neighbors, while age_since_reno calculates the property's age since construction, defaulting to 0 if the year is invalid. The first, second, fifteenth and sixteenth columns are dropped since they are analytically irrelevant (date, id, yr_built - substituted by "age" - and yr_renovated - substituted by a structural variable). We then scale only the features of the data frame. The price and the non-structural features are extracted as columns 3 to 18 and scaled in the new matrix X, which will be used for the Group Lasso. The price, or target, is stored in a separate matrix Y.

3 Base Models

3.1 Multivariate linear model

The multivariate linear regression with all original features proved a very sparse interpretation from a p-value standpoint, with only one coefficient - the number of bedrooms being a solid price predictor. We try with some more complex models that also use Elastic Net to check whether the sparsity is justified.

3.2 Multivariate GLM with Elastic Net

We proceed with a Poisson regression,

3.3 Structural multivariate GLM with Elastic Net

4 Neural Network

5 Group Lasso

5.1 The Group Lasso

The Group Lasso is an extension of the classical Lasso method applied for regularization and variable selection. It is designed to work with grouped variables, and such a situation is helpful when given variables are naturally organized in predefined groups, and we want to choose or discard entire groups of variables simultaneously rather than choosing/dropping individual variables.

The main goal of the Group Lasso is to do regression by penalizing the sum of the norms of the coefficients in each group. That will induce sparsity at the level of groups, that is, all coefficients in a group will either be shrunk towards zero, or not. Therefore, entire groups of variables are either selected or not.

5.1.1 Mathematical Formulation

For a given linear regression problem, the model can be represented as:

$$y = X\beta + \epsilon \tag{1}$$

where: y is the $n \times 1$ vector of observed responses, X is the $n \times p$ matrix of predictors, β is the $p \times 1$ vector of regression coefficient and ϵ is the error term, assumed to be normally distributed.

In ordinary linear regression, we try to minimize the residual squared error:

$$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 \right) \tag{2}$$

In the case of Group Lasso, we add a regularization term to the objective function that penalizes the coefficients in groups. If the variables are divided into G groups, and group g contains p_g variables, the Group Lasso penalty is given by:

$$\lambda \sum_{g=1}^{G} \|\beta_g\|_2 \tag{3}$$

The idea would be: where: β_g represents the coefficient corresponding to the g-th group of variables, $\|\beta_g\|_2$ is the euclidean norm of the coefficient vector, λ is a regularization parameter that regulates its strength.

Thus, the optimization problem is:

$$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \right) \tag{4}$$

5.2 Experiments

6 Conclusion