Final Project Report
Basic Probability: Programming

Mrinalini Luthra ILLC mrinalini.luthra@gmail.com
Gian Carlo Milanese ILLC giancarlo.milanese@gmail.com
Julia Turska ILLC julia.turska1@gmail.com

Abstract

The goal of our inquiry was to predict the median price of houses in the suburbs of Boston in the 1980s. We used the method of multivariate linear regression to this purpose. The details of the technique are put forth in the first section. We propose a baseline model and three improvements which allow for obtaining more accurate predictions. Finally, we discuss the collected results and justify our methodological approach.

1 Introduction

This report is based on an analysis, the goal of which was to predict the median price of houses in the suburbs of Boston in the 1980s. In order to accomplish this, we used a supervised learning method, i.e. one in which we are given the "right answer" (Ng, 2018) for each of the data points, known as *multivariate linear regression*.

This technique allows for a predictive inference from a dataset describing features of certain objects, or *training examples*, which culminates in an estimation of the correlated dependent variable, i.e. the *target variable* (Johnson and Wichern, 2002). This approach is used to model the relationship between a dependent variable and one or more independent variables. In our analysis, we use the following notation:

- n = the number of predictors
- m = number of training examples
- $\mathcal{X} = \{x^{(1)}, \ldots, x^{(m)}\}$ is the set of input variables, where each training example $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$ is a vector of feature components
- $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(m)}\}$ is the set of dependent, i.e. target variables.

The core ingredient of the analysis involves putting forth a function, or a *hypothesis*, given by the set of parameters $\theta_0, \ldots, \theta_n$, which accurately describes the relationship between the features and the estimated variable:

$$h_\theta(x^{(i)}) = \sum_{j=0}^{n} \theta_j x_j^{(i)}$$

where $x_0^{(i)}$ is set to 1 for all $1 \leq i \leq m$. The values of the parameters are determined through a minimization of the *cost function*, or the *squared error cost function*:

$$J(\theta_0, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2,$$

which indicates what $\theta_0, \ldots, \theta_n$ should be chosen so that the value of the hypothesis function is as close to the target variable for the exemplary data as possible, i.e. so that $h_\theta(x) \approx y$.

As defined above, our task is to minimize the cost function, that is $\min_{\theta_0,\ldots,\theta_n} J(\theta_0, \ldots, \theta_n)$ in order to arrive at a "good" estimator or hypothesis function. Our aim is thus to find the values of the parameters $\theta_0, \ldots, \theta_n$ that will minimize the cost function. The algorithm we will use in order to minimize our cost function is called *Gradient Descent* (henceforth GD). The intuitive idea behind GD is to start with some initial values of the parameters and update the parameters in a way such that the the cost function reduces and hopefully converges to a minimum[1] (Ng, 2018). More formally, the algorithm is:

[1] In the case of linear regression, the cost function is convex, thus has a global minimum and is thus guaranteed to converge for an appropriate learning rate.

Repeat until convergence[2]:

$$\{\theta_j := \theta_j - \alpha \frac{\partial J(\theta_0, \ldots, \theta_n)}{\partial \theta_j} (\text{for } j = 0, 1, \ldots, n)\}$$

The partial derivative

$$\frac{\partial J(\theta_0, \ldots, \theta_n)}{\partial \theta_j}$$

is the slope of the curve $J(\theta_j)$ which controls the direction in which we update the parameter θ_j. α is the learning rate which controls the amount by which we update the parameters (also known as step-size). If α is too small, GD can be slow in convergence while if α is too large GD can overshoot the minimum and thus it may fail to converge or diverge.

We further used a method called *feature scaling* in order to improve the rate of convergence. It is a technique which standardizes the range of features present in the dataset according to the following computation (Ng, 2018):

$$\text{new } x_j^{(i)} = (x_j^{(i)} - \text{mean}(x^{(i)}))/\text{max}(x^{(i)}).$$

2 Improvements

In order to achieve the best estimation, we used a number of techniques of various kinds which allow for an improvement of the model and thus a more accurate prediction.

The most basic or trivial baseline model would be one that considers only the dependent variable and thus the prediction is the average value of the dependent variable. We decided to choose a more informative baseline which compares the dependent variable with an independent variable/feature with a high (positive or negative) correlation. We then tried to improve on this model, by taking into account more features and eventually all features. Finally, we tried to improve on this baseline by adding new predictors that are combinations of some correlated features (using Pearson's coefficient).

3 Experiments

The dataset under consideration contains a description of 506 houses from the Boston Mass area collected during the 1970's by multiple governmental agencies, including U.S Census Service, FBI and others. Each description contains information about 13 features of each house. The features are presented in Table 1.

Table 1: Features (Harrison Jr and Rubinfeld, 1978)

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned
INDUS	proportion of non-retail business acres
CHAS	Charles River dummy variable
NOX	nitric oxides concentration
RM	average number of rooms per dwelling
AGE	proportion of units built prior to 1940
DIS	distances to employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	the proportion of blacks by town
LSTAT	% lower status of the population

In order to evaluate whether the improvements we have implemented for the baseline model have in fact made our prediction more accurate, we have computed the coefficient of determination R^2 after applying each of the improvements (Table 2). The computation of R^2 is a standard method for comparing regression models and contains information about the distances between the predictions to the actual values of the target variable, which goes in the following way:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(y_i - h(x^{(i)})\right)^2}{\sum_{i=1}^{m} \left(y_i - \frac{1}{m} \sum_{j=1}^{m} y_j\right)^2}$$

Table 2: Evaluation of the models

Model	R^2
MEDV vs RM	0.48
MEDV vs LSTAT	0.54
ALL	0.74
LOGDIS	0.80
INTER	0.844

Most Basic Baseline/Trivial

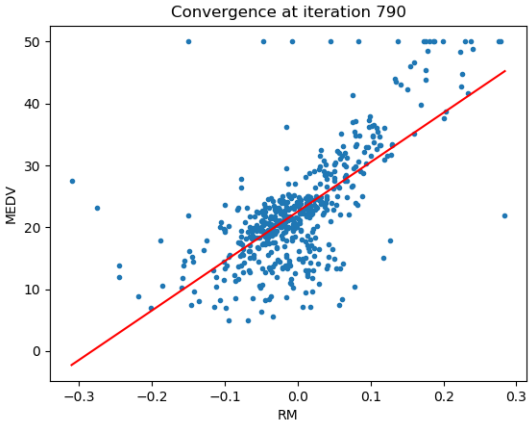
In this case, we calculate the average of the MEDV values. Thus, it is just a constant and does not depend on the features at all. Furthermore, we have noticed that θ_0 approximates the average value of

[2]It is crucial that the parameters are updated simultaneously.

MEDV. Clearly, in this case, the R^2 value is 0, making it the most trivial and highly uninformative prediction.

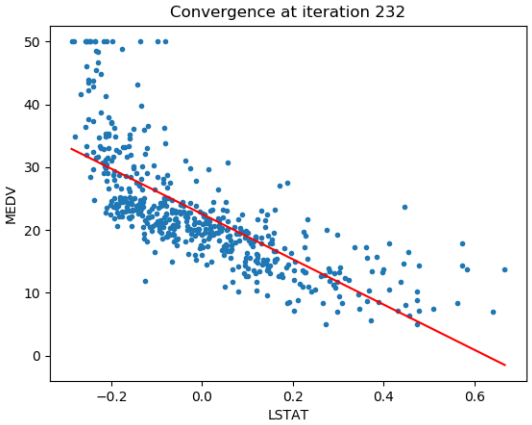
MEDV vs RM

We noticed that there is a strong positive correlation between RM and MEDV. Moreover, this matches the intuition that the higher the number of rooms in a house, the higher would be the price of the house. Thus, taking our hypothesis function to be a function of the RM feature as a sole predictor we obtained an R^2 value of 0.48.



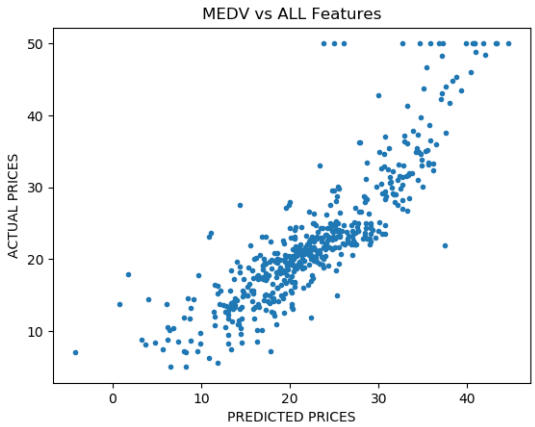
MEDV vs LSTAT

As was the case with RM, we noticed that there is a strong negative correlation between LSTAT and MEDV values, which also corresponds to the intuitive assessment of what could affect the price. We obtained an R^2 value of 0.54 taking LSTAT as our only predictor of house prices.



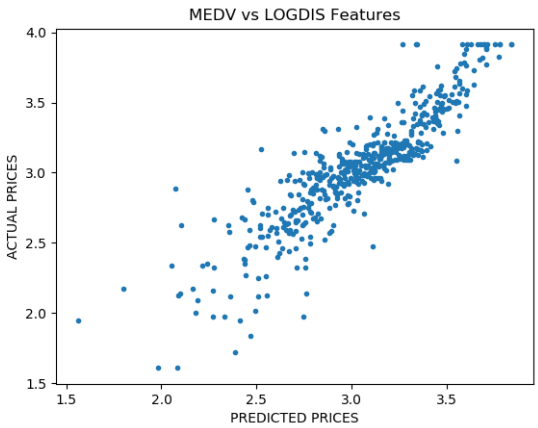
MEDV vs All Features (ALL)

Furthermore, we explored the idea of using all available features as predictors in the model which yielded an an R^2 value of 0.74[3].



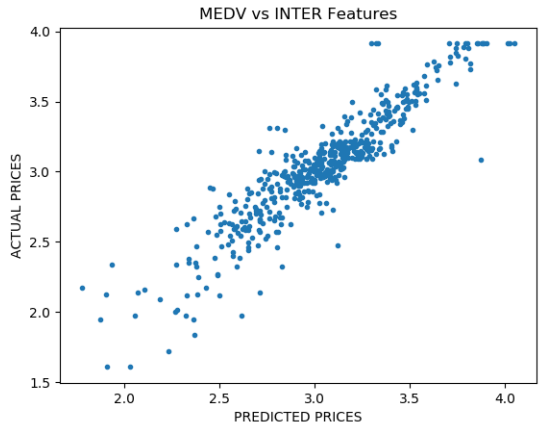
MEDV vs All Features and log-distances (LOGDIS)

This improvement adds an additional predictor that takes the logarithm of the feature "DIS" (thus 14 predictors in total). This improved our R^2 value to 0.8.



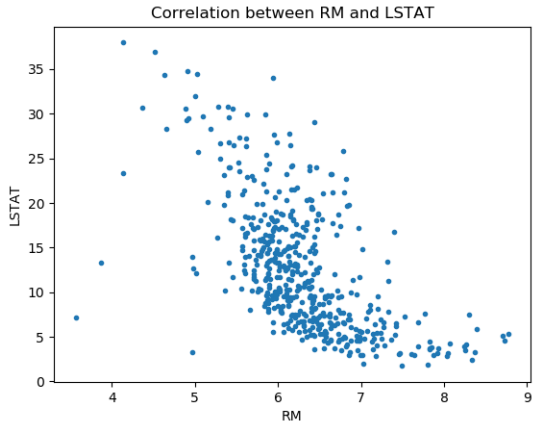
[3]Our sincerest apologies for not including a 14-dimensional plot. This task proved to be beyond our skill sets and given the task of improving R^2 also unnecessary. Thus, we provide a plot depicting the relationship between predicted prices and actual prices.

Interactions (INTER)



Pearson's correlation coefficient or Pearson's r is a method which allows for an assessment of the linear correlation between two variables of the model. The value of r is in the range $[-1, 1]$ where $r = 1$ implies total positive linear correlation, $r = -1$ implies total negative linear correlation and $r = 0$ implies zero correlation. The Pearson correlation between two variables is the covariance of the variables divided by the product of their standard deviations. In our linear regression model we combined variables with high Pearson's correlation r to obtain new predictors.

We looked at features with correlation $|r| > 0.5$ and added 13 new predictors ($n = 26$) by combining those features which fit that criterion. For instance, we noticed that there is a strong negative correlation between RM and LSTAT, as can be seen in the figure below, and thus we included their combination into our set of predictors.



4 Conclusions

We believe that our results are probably not reliable. This is because of the fact that we have a small dataset ($m = 506$ examples) and a large number of features ($n = 13$). Moreover, we have only focused on improving the value of R^2. Thus, there is a high chance of over-fitting. We learnt that one way of drumming up a more reliable model would be to split the training and test sets and perform cross-validation.

Given that each of our improvements has led to an increase in the value of R^2 by at least four percent, we find that our results are satisfactory.

What we have come to understand about data analysis is that there is no standard way of performing it correctly. Assessing what can be done to improve the predictions of the model involves a good comprehension of the particular dataset (as, for instance, which features correlate with each other) as well as fluency in employing a tool set of mathematical methods which enable achieving better predictions (e.g. log-distances method).

References

David Harrison Jr and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

Richard A Johnson and Dean Wichern. 2002. *Multivariate analysis*. Wiley Online Library.

Andrew Ng. 2018. Coursera: Machine learning. `https://www.coursera.org/learn/machine-learning`. Accessed: 2018-05-31.