A

Project Report

On

HOUSING PRICES PREDICTION USING REGRESSION ALGORITHM

Submitted in partial fulfillment for the award of degree of

Bachelor of Technology

In

Computer Science & Engineering

Submitted By
Rohitashva Singh Jhala
16EJCCS751

Project Guide
Ms. Sweety Singhal
Assistant Professor

Department of Computer Science & Engineering
Jaipur Engineering College & Research Centre
Jaipur, Rajasthan

2019-20

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this project entitled "**Housing Prices Prediction using Regression Algorithm"** in the partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science & Engineering at Jaipur Engineering College and Research Centre, Jaipur is an authentic work of my own.

I have not submitted the matter embodied in this project work anywhere for the award of degree of Bachelor of Technology in Computer Science & Engineering.

Rohitashva Singh Jhala

16EJCCS751

30th April'20

JECRC Foundation

BONAFIDE CERTIFICATE

This is to certify that the project entitled **"Housing prices prediction using regression algorithms"** is the bonafide work carried out by **Rohitashva Singh Jhala**, student of B.Tech.in Computer Science & Engineering at Jaipur Engineering College and Research Centre, during the year 2019-20 in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science & Engineering and the project has not formed the basis for the award previously of any degree, diploma, fellowship or any other similar title.

Ms. Sweety Singhal

Assistant Professor

JECRC Foundation

30th April'20

VISION OF CSE DEPARTMENT

To become renowned Centre of excellence in computer science and engineering and make competent engineers & professionals with high ethical values prepared for lifelong learning.

MISSION OF CSE DEPARTMENT

1. To impart outcome based education for emerging technologies in the field of computer science and engineering.
2. To provide opportunities for interaction between academia and industry.
3. To provide platform for lifelong learning by accepting the change in technologies
4. To develop aptitude of fulfilling social responsibilities.

PROGRAM OUTCOMES (POs)

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

The PEOs of the B.Tech (CSE) program are:

1. To produce graduates who are able to apply computer engineering knowledge to provide turn-key IT solutions to national and international organizations.
2. To produce graduates with the necessary background and technical skills to work professionally in one or more of the areas like – IT solution design development and implementation consisting of system design, network design, software design and development, system implementation and management etc. Graduates would be able to provide solutions through logical and analytical thinking.
3. To able graduates to design embedded systems for industrial applications.
4. To inculcate in graduates effective communication skills and team work skills to enable them to work in multidisciplinary environment.
5. To prepare graduates for personal and professional success with commitment to their ethical and social responsibilities.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- PSO1: Ability to interpret and analyze the real world data and predict the future outcomes from the initial dataset.

- PSO2: Ability to design and develop realistic data analyzing algorithms on the basis of training data provided to it.

COURSE OUTCOMES (COs)

On completion of project Graduates will be able to-

- CO1: Gather, organize, summarize and interpret technical literature with the purpose of formulating a project proposal.
- CO2: Design/Develop the solution using latest technologies and communicate via modern tools.
- CO3: Understand and develop the professional, social ethics, and team management principles.

MAPPING: CO's & PO's

Subject	**Code**	**L/T/P**	**CO**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**
Project Stage-II	8CSPR	P	Graduates will be able to gather, organize, summarize and interpret technical literature with the purpose of formulating a project proposal	H	H	M	H	M	M	--	--	M	M	M	M
		P	Graduates will be able to design/develop the solution using latest technologies and communicate via modern tools	H	H	H	H	H	M	L	--	M	H	M	H
		P	Graduates will be able to understand and develop the professional, social ethics, and team management principles	M	--	--	--	L	H	M	H	H	L	H	H

ACKNOWLEDGEMENT

I wish to express our deep sense of gratitude to our Project Guide **Ms. Sweety Singhal**, Jaipur Engineering College and Research Centre, Jaipur for guiding us from the inception till the completion of the project. We sincerely acknowledge her for giving her valuable guidance, support for literature survey, critical reviews and comments for our Project.

I would like to first of all express our thanks to Mr. Arpit Agrawal Director of JECRC, for providing us such a great infrastructure and environment for our overall development.

I express sincere thanks to Dr. V. K. Chandna, Principal of JECRC, for his kind cooperation and extendible support towards the completion of our project.

Words are inadequate in offering our thanks to Dr. Sanjay Gaur, HOD of CSE department, for consistent encouragement and support for shaping our project in the presentable form.

I also like to express our thanks to all supporting CSE faculty members who have been a constant source of encouragement for successful completion of the project.

Also our warm thanks to Jaipur Engineering College and Research Centre, who provided us this opportunity to carryout, this prestigious Project and enhance our learning in various technical fields.

Rohitashva Singh Jhala

16EJCCS751

Abstract

Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. For the selection of prediction methods we compare and explore various prediction methods. We utilize lasso regression as our model because of its adaptable and probabilistic methodology on model selection. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models. More over on other hand housing value indices, the advancement of a housing cost prediction that tend to the advancement of real estate policies schemes. This study utilizes machine learning algorithms as a research method that develops housing price prediction models. We create a housing cost prediction model In view of machine learning algorithm models for example, XGBoost, lasso regression and neural system on look at their order precision execution. We in that point recommend a housing cost prediction model to support a house vender or a real estate agent for better information based on the valuation of house. Those examinations exhibit that lasso regression algorithm, in view of accuracy, reliably outperforms alternate models in the execution of housing cost prediction.

List of Figures

List of Tables

Table of Content

INTRODUCTION

What Is Learning? Rats Learning to Avoid Poisonous Baits: Rats normally stumble upon food items by its look and smell and start eating in small amounts and later depending on food and physiological effect the feeding of food goes on. If the rat notices the illness of the food, the rat will not touch that food. Similarly the machine learning mechanism plays a vital role same as animal usage of past experience for acquiring and expertise in detecting the food safety. By mistake if the knowledge with the food is negatively labeled, the prediction of the animal will also will be negatively affected and encountered in the future. With the inspiration of the previous example of successful learning we demonstrate a typical machine learning algorithm .Likewise we would want to program a machine that learns how to filter spam e-mails. A credulous result might be apparently comparable of the lifestyle of rats that, how to keep away from poisonous baits. The machine will basically remember the past e-mails that needed been named similarly as spam e-mails by the human user. When another email arrives, the machine will look for it in the past set about spam e-mails. Though it matches among them, it will be trashed. Otherwise, it will make moved of the user's inbox organizer. At the same time the first "learning by memorization" methodology may be useful, it fails to offer an important aspect known as learning systems – the capacity to mark unseen email messages. A fruitful learner ought to have the ability which will be the advancement from distinctive samples to more extensive generalization. This may be Likewise as inductive thinking or inductive induction. In the attraction nervousness exhibited previously, after the rats experience a sample of a certain sort about food; they apply their disposition at it once new, unseen illustrations from claiming nourishment of comparable emanation Also taste. Should attain generalization in the spam sifting task, those learner might examine the Awhile ago seen e-mails, and extricate An situated about expressions whose presence for a email message is characteristic of spam. Then, At another email arrives, those machine could weigh if a standout among the suspicious expressions gives the idea On it, and foresee its mark Appropriately. Such an arrangement might possibly have the ability effectively to foresee the name about unseen e-mails.

Responsibilities further than Human Capabilities: an additional totally crew about errands that profit starting with machine Taking in systems are identified with the Investigation for extremely substantial and intricate information sets: galactic data, turning restorative chronicles under restorative knowledge, climate prediction, and dissection of genomic data, Web search engines, Also electronic trade. With an ever increasing amount accessible digitally recorded data, it gets evident that there would treasures about serious majority of the data covered clinched alongside information chronicles that would best approach excessively little and also as well perplexing to people with bode well about. Taking in with recognize serious examples over substantial Also complex information sets may be a guaranteeing space for which the blending of projects that take for the Just about boundless memory limit and ever expanding transforming velocity about PCs opens up new horizons. Regulated versus Unsupervised, since taking in includes an association between those learner and the environment, you quit offering on that one might separate taking in assignments as stated by those nature for that connection.

The to start with qualification will note is the Contrast the middle of regulated and unsupervised Taking in. Likewise an illustrative example, think about that errand for Taking in will recognize spam email versus the undertaking about aberrance identification. For the spam identification task, we think about a setting to which the learner receives preparing e-mails to which the mark spam/not-spam may be Gave. On the support from claiming such preparation the learner ought to further bolstering to evaluate a tenet to labeling a recently arriving email message. For contrast, for those assignment about aberrance detection, every last one of learner gets Concerning illustration preparing is an extensive form of email messages (with no labels) and the learner's errand is on identify "unusual" messages.

Investment is a business activity on which most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks and property. In particular, property investment has increased significantly. Housing price trends are not only the concern of buyers and sellers, but it also indicates the current economic situation. There are many factors which has impact on house prices, such as numbers of bedrooms and bathrooms. Even the nearby location, a location with a great accessibility to highways, expressways, schools, shopping malls and local employment opportunities contributes to the rise in house price. Manual house predication becomes difficult, hence there are many systems developed for house price prediction. This Project has proposed an advanced house prediction system using linear regression. This system aim is to make a model which can give us a good house pricing prediction based on other variables. We are going to use Linear Regression for this dataset and hence it gives a good accuracy. This house price prediction project has two modules namely, Admin and User. Admin can add location and view the location. Admin has authority to add density on the basis of per unit area. User can view the location and see the predicted housing price for the particular location.

LITERATURE SURVEY

The latest worldwide financial crisis restored a sharp enthusiasm toward both academic and strategy circles on the part of asset costs and specifically lodging costs clinched alongside monetary movement. As Lamer (2007) notes those lodging showcase predicted eight of the ten post globe War ii recessions, acting Concerning illustration An heading woman for those true segment of the economy. Truth be told he dives Likewise significantly Concerning illustration with state that “Housing is those benefits of the business cycle”. Vargas and silva (2008) contend that lodging costs alterations assume a paramount part in the determination of the stage of the business cycle. When those economy booms, development and work in the lodging division expand quickly should react should overabundance demand, quickly pushing ostensible house costs upwards. Throughout those withdrawal phase, the drop in private money lessens aggravate interest Also ostensible house costs. By ostensible house costs normally fall sluggishly since householders would unwilling on bring down their costs. The majority of the conformity will be attained through declines clinched alongside bargains volume bringing about An drop in the development segment and the lodging built vocation. Moreover, Throughout withdrawal and subsidence true house costs fall quickly Likewise general inflationary patterns diminish true house costs much with sticky perceived costs.

Recently, a few writers scope to experimental discoveries that house costs can make instrumental molding to determining yield. (Forni etc, 2003; stock and Watson, 2003; Gupta Furthermore Das, 2010; das etc, 2009; 2010; 2011; Gupta and Hartley, 2013). Those lodging development divisions speaks to an expansive and only aggregate monetary action communicated in the GDP. Consequently, concerning illustration it reflects an extensive parcel of the general riches of the economy, house costs variances can make a pointer of the Development about GDP (Case etc, 2005). Concerning illustration it is those body of evidence with different assets, those development for house costs can make Additionally an pointer of the future course from claiming expansion (Gupta Also Kabundi, 2010). Overall, exact determining of the Development way from claiming house costs could make a suitable apparatus both on house business members and fiscal strategy powers. There is huge literature writing in regards to U.S. house prices. Rapach Furthermore strauss (2007) use an auto regressive dispersed slack (ARDL) model framework, holding 25 determinants with conjecture genuine lodging cost development to the unique states of the elected Reserve’s eighth region. They discover that ARDL models tend should beat a benchmark AR model. Rapach and strauss (2009) augment those same examination on the 20 biggest u. Encountered with urban decay because of de industrialization, innovation developed, government agent.

States dependent upon ARDL models looking at state, territorial and national level variables. When again, the creators scope comparative conclusions on the fact that joining together forecasts about models for different slack structure. Gogas and Pragidis (2011) utilize the hazard premium ascertained Likewise those Contrast the middle of Different long haul enthusiasm rates and the agents' desires over future fleeting rates as information variable to foreseeing what's to come heading for house costs. They infer that gurus Also investigators could use adequately those majority of the data given by those investment rate hazard premium today so as should evaluate the likelihood from claiming acquiring beneath pattern S&P CS-10 list three months ahead. Gupta and Das (2010) also forecast the recent downturn in real house price growth rates for the twenty largest U.S. states. The authors use Spatial Bayesian VARs (BVARs), based only on monthly real house price growth rates, to forecast their downturn over the period 2007:01 to 2008:01. They find that BVAR models are well-equipped in forecasting the future direction of real house prices, though they significantly underestimate the decline. They attribute this under-prediction of the BVAR models to the lack of any information on fundamentals in the estimation process. Rapach and Strauss (2009) expand the individuals same examination on the 20 most amazing . Encountered with urban rot due to de industrialization, advancement developed, administration agonizes. States reliant upon ARDL models taking a gander at state, regional Also national level variables. When again, those inventors degree similar finishes on the reality that joining together forecasts regarding models for separate slack structure.

Gogas and Pragidis (2011) use the danger premium determined similarly the individuals complexity those white collar for different whole deal energy rates and the agents' longings In future transient rates as data variable with foreseeing what's with turn heading to house expenses. They construe that masters also investigators Might utilization enough the individuals lion's share of the information provided for by the individuals financing rate danger premium today with the goal Similarly as ought further bolstering assess those probability from asserting securing underneath design S&P CS-10 rundown three months ahead.

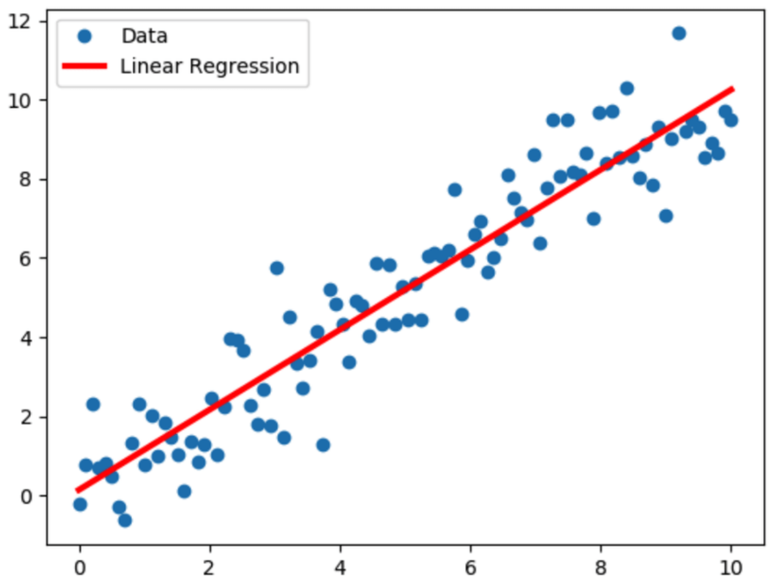
Design Approach

Linear Regression

linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.



Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Multiple Linear Regression

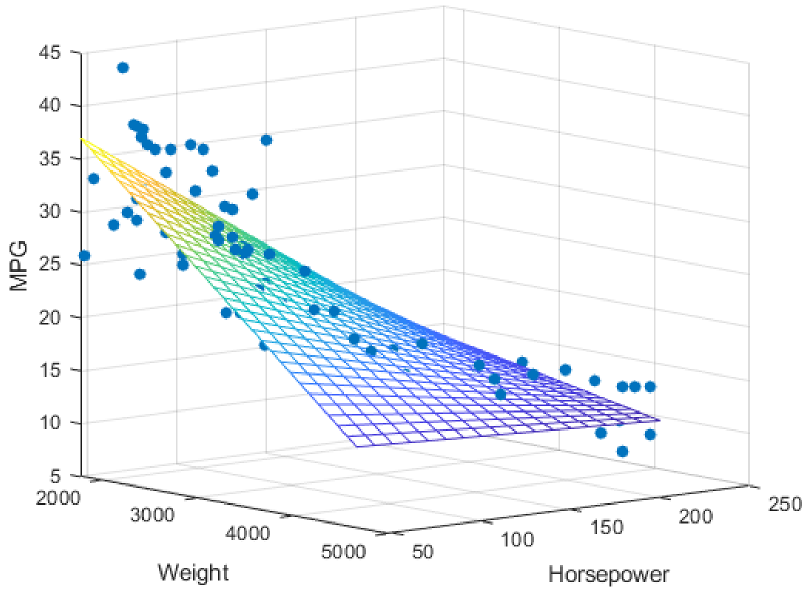
Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.

- yi observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.



The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

Multiple linear regression (MLR) is used to determine a mathematical relationship among a number of random variables. In other terms, MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.

Cost Function

It is a function that measures the performance of a Machine Learning model for given data. Cost Function quantifies the error between predicted values and expected values and presents it in the form of a single real number. Depending on the problem Cost Function can be formed in many different ways. The purpose of Cost Function is to be either:

- Minimized - then returned value is usually called cost, loss or error. The goal is to find the values of model parameters for which Cost Function return as small number as possible.
- Maximized - then the value it yields is named a reward. The goal is to find values of model parameters for which returned number is as large as possible.

For algorithms relying on Gradient Descent to optimize model parameters, every function has to be differentiable.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}^{(i)} - y^{(i)}|$$

here:

i - Index of sample,

ŷ - predicted value,

y - expected value,

m - Number of samples in dataset.

The MSE formula can be written like this:

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

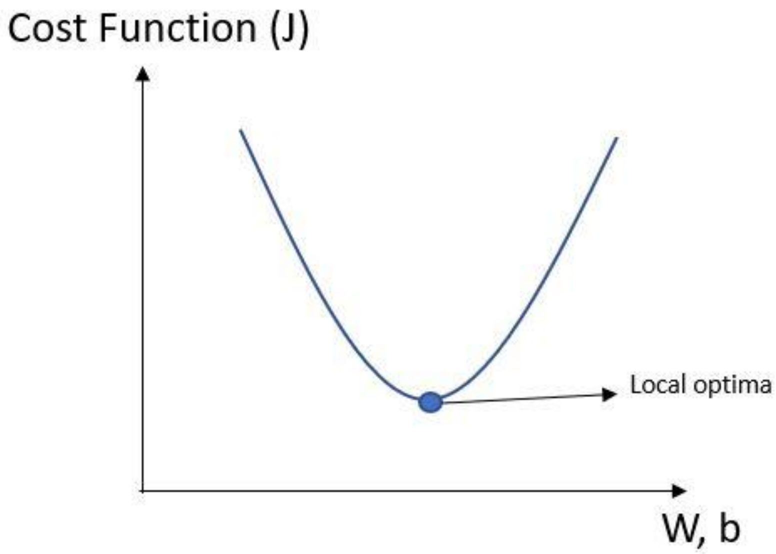
Where,

i - Index of sample,

ŷ - Predicted value,

y - Expected value,

m - Number of samples in dataset.



Gradient Descent

Gradient descent is a first order iterative optimization algorithm for finding a local minimum of a differentiable function. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. But if we instead take steps proportional to the positive of the gradient, we approach a local maximum of that function; the procedure is then known as gradient ascent. Gradient descent is also known as steepest descent; but gradient descent should not be confused with the method of steepest descent for approximating integrals.

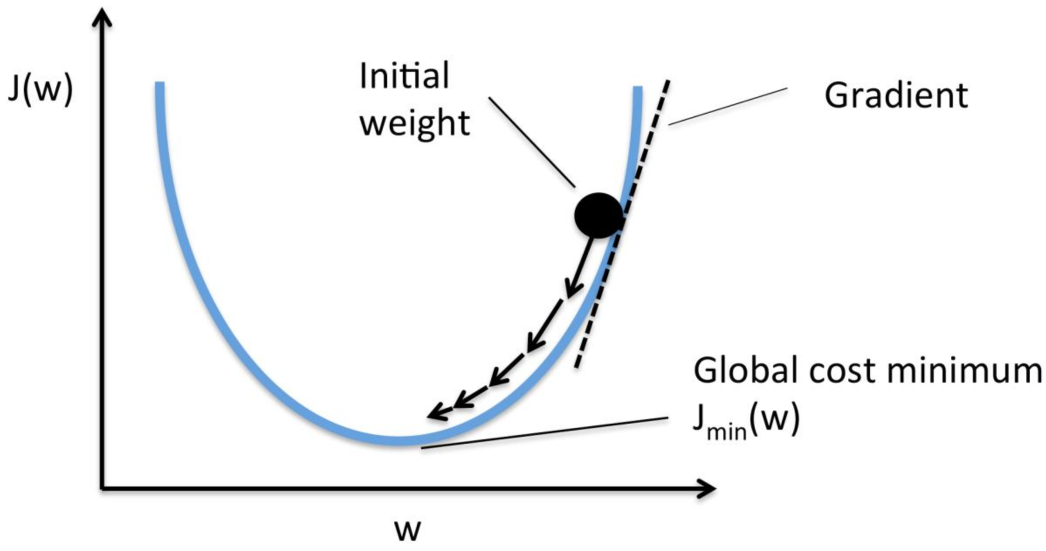
Gradient descent works in spaces of any number of dimensions, even in infinite-dimensional ones. In the latter case, the search space is typically a function space, and one calculates the Fréchet derivative of the functional to be minimized to determine the descent direction.

The gradient descent can take many iterations to compute a local minimum with a required accuracy, if the curvature in different directions is very different for the given function. For such functions, preconditioning, which changes the geometry of the space to shape the function level sets like concentric circles, cures the slow convergence. Constructing and applying preconditioning can be computationally expensive, however.

The gradient descent can be combined with a line search, finding the locally optimal step size on every iteration. Performing the line search can be time-consuming. Conversely, using a fixed small can yield poor convergence.

Methods based on Newton's method and inversion of the Hessian using conjugate gradient techniques can be better alternatives. Generally, such methods converge in fewer iterations, but the cost of each iteration is higher. An example is the BFGS method which consists in calculating on every step a matrix by which the gradient vector is multiplied to go into a "better" direction, combined with a more sophisticated line search algorithm, to find the "best" value of For extremely large problems, where the computer-memory issues dominate, a limited-memory method such as L-BFGS should be used instead of BFGS or the steepest descent.

Gradient descent can be viewed as applying Euler's method for solving ordinary differential equations to a gradient flow.Gradient descent can be extended to handle constraints by including a projection onto the set of constraints. This method is only feasible when the projection is efficiently computable on a computer. Under suitable assumptions, this method converges. This method is a specific case of the forward-backward algorithm for monotone inclusions (which includes convex programming and variation).



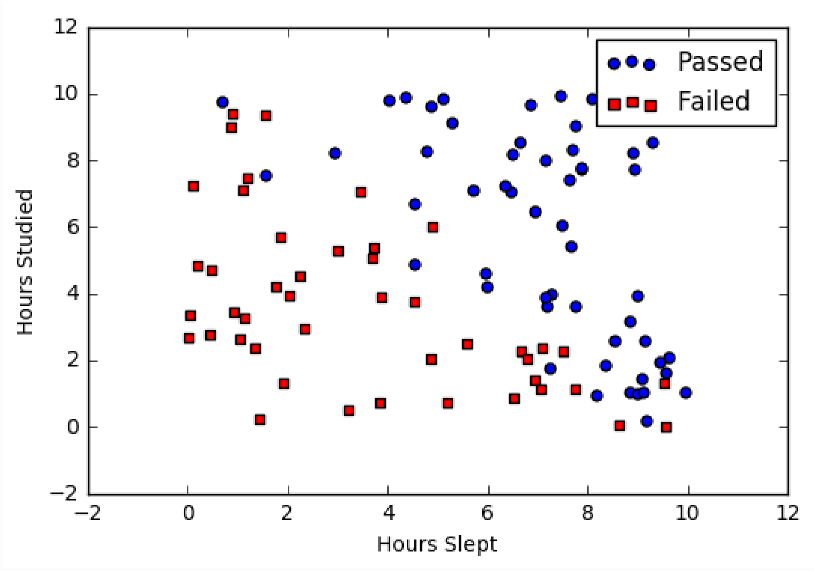
Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Logistic Regression could help use predict whether the student passed or failed. Logistic regression predictions are discrete (only specific values or categories are allowed). We can also view probability scores underlying the model's classifications.



Types of logistic regression

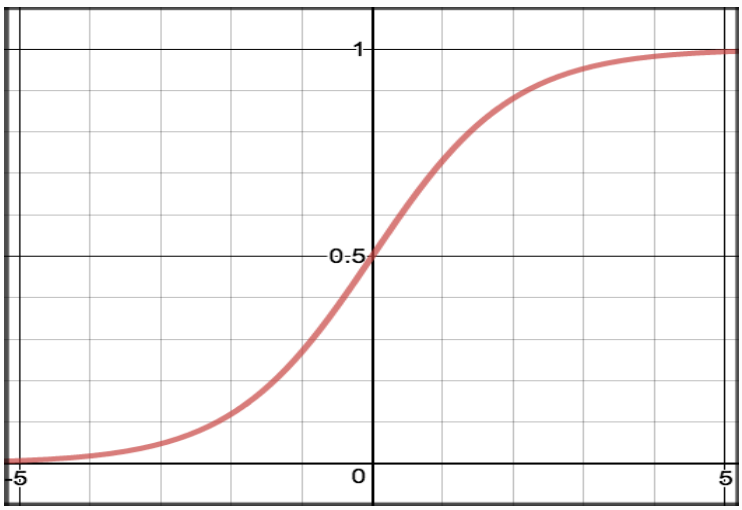
- Binary (Pass/Fail)
- Multi (Cats, Dogs, Sheep)
- Ordinal (Low, Medium, High)

Sigmoid activation

In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

Math

S(z)=11+e−z



Decision boundary

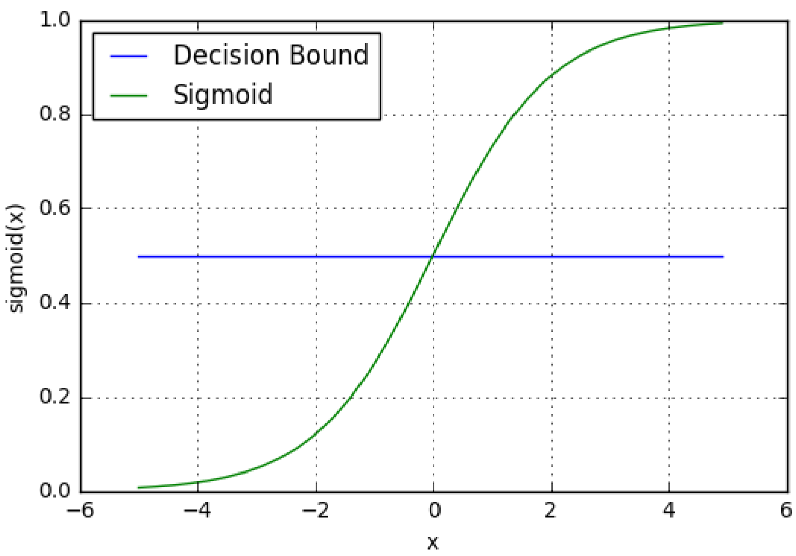
Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (true/false, cat/dog), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

$$p \geq 0.5, class=1$$

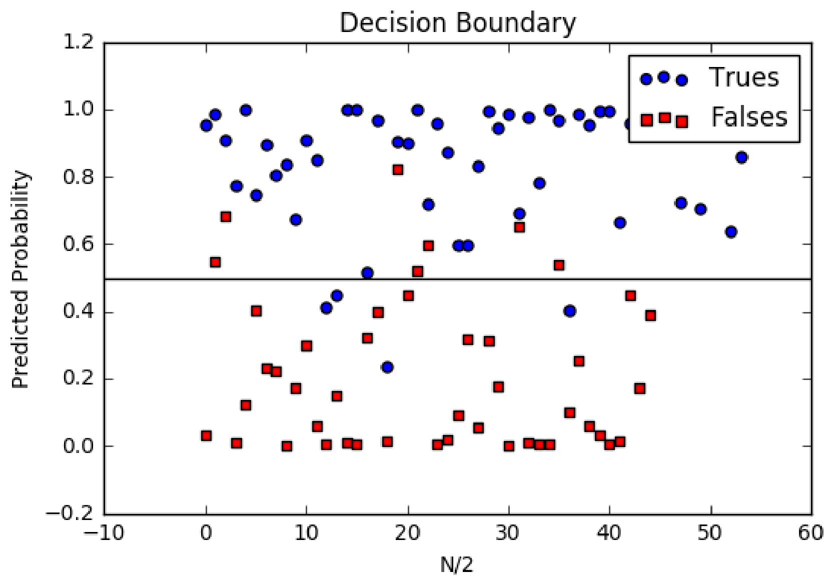
$$p < 0.5, class=0$$

For example, if our threshold was .5 and our prediction function returned .7, we would classify this observation as positive. If our prediction was .2 we would classify the observation as negative. For logistic regression with multiple classes we could select the class with the highest predicted probability.

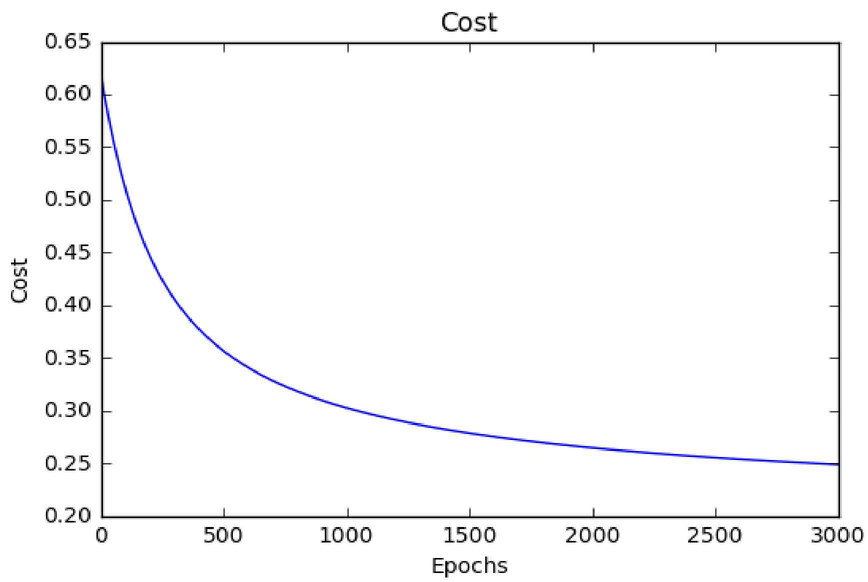
Using our knowledge of sigmoid functions and decision boundaries, we can now write a prediction function. A prediction function in logistic regression returns the probability of our observation being positive, True, or “Yes”. We call this class 1 and its notation is P(class=1)P(class=1). As the probability gets closer to 1, our model is more confident that the observation is in class 1.



Another helpful technique is to plot the decision boundary on top of our predictions to see how our labels compare to the actual labels. This involves plotting our predicted probabilities and coloring them with their true labels.



If our model is working, we should see our cost decrease after every iteration and after a point of time the variation in the value of cost function will vary very less as compare to the number of iterations. At that time we have to stop the algorithm and the function then produced fits the best.



Feature scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling, Consider the two most important ones:

- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\textbf{Standard Deviation}}$$

Implementations

```
%% Initialization
clear ; close all; clc

%% ==================== Part 1: Basic Function ==================
% Complete warmUpExercise.m
fprintf('Running warmUpExercise ... \n');
fprintf('5x5 Identity Matrix: \n');
warmUpExercise()

fprintf('Program paused. Press enter to continue.\n');
pause;



%% ==================== Part 2: Plotting =======================
fprintf('Plotting Data ...\n')
data = load('ex1data1.txt');
X = data(:, 1); y = data(:, 2);
m = length(y); % number of training examples

% Plot Data
% Note: You have to complete the code in plotData.m
plotData(X, y);

fprintf('Program paused. Press enter to continue.\n');
pause;
```

```
%% ================= Part 3: Cost and Gradient descent ============

X = [ones(m, 1), data(:,1)]; % Add a column of ones to x
theta = zeros(2, 1); % initialize fitting parameters

% Some gradient descent settings
iterations = 1500;
alpha = 0.01;

fprintf('\nTesting the cost function ...\n')
% compute and display initial cost
J = computeCost(X, y, theta);
fprintf('With theta = [0 ; 0]\nCost computed = %f\n', J);
fprintf('Expected cost value (approx) 32.07\n');

% further testing of the cost function
J = computeCost(X, y, [-1 ; 2]);
fprintf('\nWith theta = [-1 ; 2]\nCost computed = %f\n', J);
fprintf('Expected cost value (approx) 54.24\n');

fprintf('Program paused. Press enter to continue.\n');
pause;

fprintf('\nRunning Gradient Descent ...\n')
% run gradient descent
theta = gradientDescent(X, y, theta, alpha, iterations);
```

```
% print theta to screen
fprintf('Theta found by gradient descent:\n');
fprintf('%f\n', theta);
fprintf('Expected theta values (approx)\n');
fprintf(' -3.6303\n  1.1664\n\n');

% Plot the linear fit
hold on; % keep previous plot visible
plot(X(:,2), X*theta, '-')
legend('Training data', 'Linear regression')
hold off % don't overlay any more plots on this figure

% Predict values for population sizes of 35,000 and 70,000
predict1 = [1, 3.5] *theta;
fprintf('For population = 35,000, we predict a profit of %f\n',...
    predict1*10000);
predict2 = [1, 7] * theta;
fprintf('For population = 70,000, we predict a profit of %f\n',...
    predict2*10000);

fprintf('Program paused. Press enter to continue.\n');
pause;

%% ============ Part 4: Visualizing J(theta_0, theta_1) ============
fprintf('Visualizing J(theta_0, theta_1) ...\n')
```

```
% Grid over which we will calculate J

theta0_vals = linspace(-10, 10, 100);

theta1_vals = linspace(-1, 4, 100);


% initialize J_vals to a matrix of 0's

J_vals = zeros(length(theta0_vals), length(theta1_vals));


% Fill out J_vals

for i = 1:length(theta0_vals)

    for j = 1:length(theta1_vals)

              t = [theta0_vals(i); theta1_vals(j)];

              J_vals(i,j) = computeCost(X, y, t);

    end

end




% Because of the way meshgrids work in the surf command, we need to

% transpose J_vals before calling surf, or else the axes will be flipped

J_vals = J_vals';

% Surface plot

figure;

surf(theta0_vals, theta1_vals, J_vals)

xlabel('\theta_0'); ylabel('\theta_1');


% Contour plot

figure;

% Plot J_vals as 15 contours spaced logarithmically between 0.01 and 100
```

```
contour(theta0_vals, theta1_vals, J_vals, logspace(-2, 3, 20))

xlabel('\theta_0'); ylabel('\theta_1');

hold on;

plot(theta(1), theta(2), 'rx', 'MarkerSize', 10, 'LineWidth', 2);
```

Data Set:

E1 | fx | 18.945,22.665

	A	B	C	D	E	F
1	6.1101,17.592		18.945,22.638		18.945,22.665	
2	5.5277,9.1302		18.945,22.639		18.945,22.666	
3	8.5186,13.662		18.945,22.640		18.945,22.667	
4	7.0032,11.854		18.945,22.641		18.945,22.668	
5	5.8598,6.8233		18.945,22.642		18.945,22.669	
6	8.3829,11.886		18.945,22.643		18.945,22.670	
7	7.4764,4.3483		18.945,22.644		18.945,22.671	
8	8.5781,12		18.945,22.645		18.945,22.672	
9	6.4862,6.5987		18.945,22.646		18.945,22.673	
10	5.0546,3.8166		18.945,22.647		18.945,22.674	
11	5.7107,3.2522		18.945,22.648		18.945,22.675	
12	14.164,15.505		18.945,22.649		18.945,22.676	
13	5.734,3.1551		18.945,22.650		18.945,22.677	
14	8.4084,7.2258		18.945,22.651		18.945,22.678	
15	5.6407,0.71618		18.945,22.652		18.945,22.679	
16	5.3794,3.5129		18.945,22.653		18.945,22.680	
17	6.3654,5.3048		18.945,22.654		18.945,22.681	
18	5.1301,0.56077		18.945,22.655		18.945,22.682	
19	6.4296,3.6518		18.945,22.656		18.945,22.683	
20	7.0708,5.3893		18.945,22.657		18.945,22.684	
21	6.1891,3.1386		18.945,22.658		18.945,22.685	
22	20.27,21.767		18.945,22.659		18.945,22.686	
23	5.4901,4.263		18.945,22.660		18.945,22.687	
24	6.3261,5.1875		18.945,22.661		18.945,22.688	
25	5.5649,3.0825		18.945,22.662		18.945,22.689	
26	18.945,22.638		18.945,22.663		18.945,22.690	
27	18.945,22.639		18.945,22.664		18.945,22.691	

ex1data1

Select destination and press ENTER or choose Paste

Execution and Outputs

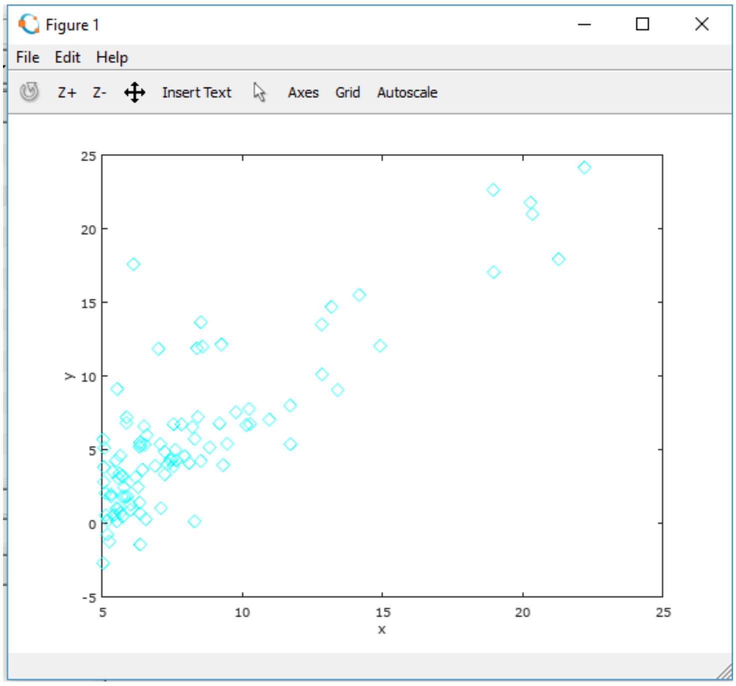
When the code gets executed first we get outputs plots and then prediction takes place. These plots help us to understand the correlation between target variable (price) and different predictor variables. This plot gives us the relation between area of the house (in feet^2) with respect to the cost of house (in 1000 $).

Command Window

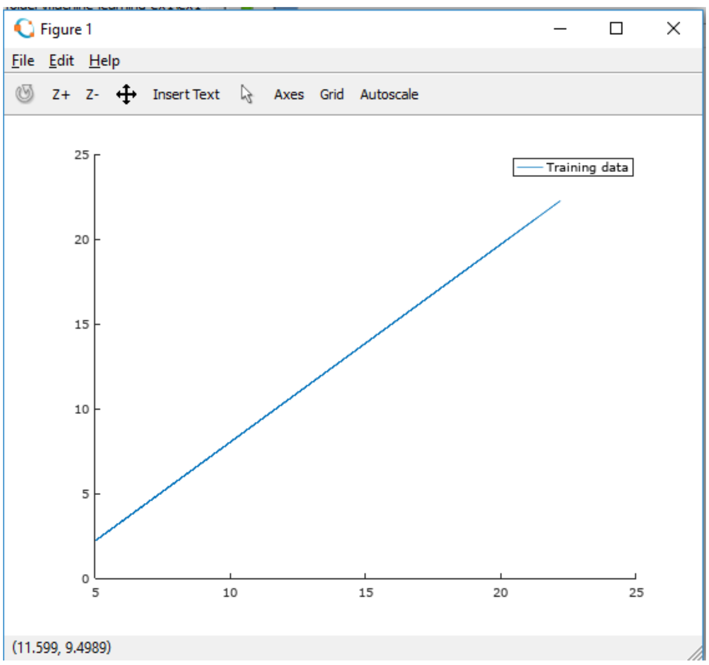
```
   1    0    0    0    0
   0    1    0    0    0
   0    0    1    0    0
   0    0    0    1    0
   0    0    0    0    1

Program paused. Press enter to continue.
Plotting Data ...
Program paused. Press enter to continue.

Testing the cost function ...
With theta = [0 ; 0]
Cost computed = 32.072734
Expected cost value (approx) 32.07

With theta = [-1 ; 2]
Cost computed = 54.242455
Expected cost value (approx) 54.24
Program paused. Press enter to continue.

Running Gradient Descent ...
Theta found by gradient descent:
-3.630291
1.166362
Expected theta values (approx)
 -3.6303
  1.1664

warning: legend: ignoring extra labels
warning: called from
    legend at line 428 column 9
    ex1 at line 88 column 1
For population = 35,000, we predict a profit of 4519.767868
For population = 70,000, we predict a profit of 45342.450129
Program paused. Press enter to continue.
Visualizing J(theta_0, theta_1) ...
>>
```

Command Window | Editor

- This is the command window of the execution. Initially we plot the data sets onto the graph and see the relation and variation between them. The data points are represented using the blue markers in the figure.

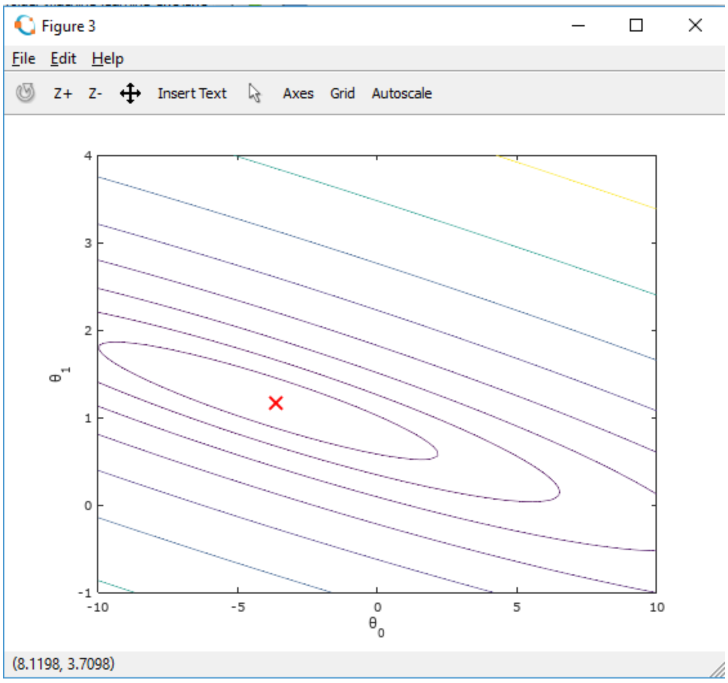


- Now, we begin training the theta and calculating the cost function for theta values [0; 0] and [-1; 2].

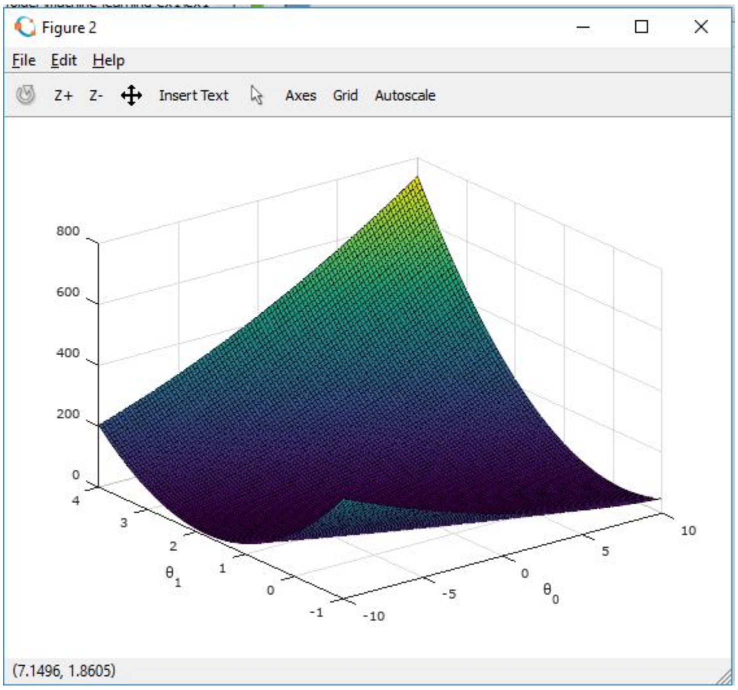


- Now the function is trained and is represented with a linear equation of a single line representing a relation between house area and its respective prices.

- Now by running the Gradient descent, we determine the values of theta as [-3.630291; 1.166362]. For these values of theta we now have calculated the output as 4519.767868 and 45342.450129 for the values 35000 and 70000 respectively.



- Now, By visualizing the cost function for the theta values of theta_0 and theta_1.



Conclusion

We have managed out how to prepare a model that gives users for a novel best approach with take a gander at future lodging value predictions. A few relapse strategies have been investigated Furthermore compared, when arriving during a prediction strategy In light of XG support. Straight former imply works bring been utilized within our model, something like that that future value predictions will have a tendency towards All the more sensible values. We concocted an approach with use similarly as considerably information as time permits for our prediction system, by adopting those ideas from claiming gradient boosting. In spite of Hosting generated all the attempting provision that met our introductory requirements, there are Different upgrades that could be produced later on. These incorporate upgrades we didn't settle on because of constrained duration of the time.

A real worry for the prediction framework may be the stacking period. Moreover, our data set takes more than one day should prepare. As opposed performing the computations sequentially, we might utilize various processors and parallel the computations involved, which might possibly decrease the preparation time Furthermore prediction period. Include All the more functionalities under the model, we can give choices for client with select a district alternately locale should produce those high temperature maps, as opposed to entering in the list.

References

[1]. https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471

[2]. https://www.sas.com/en_us/insights/analytics/machinelearning.html#machine-learning-importance

[3]. http://www.wired.co.uk/article/machine-learning-ai-explained

[4]. https://deeplearning4j.org/ai-machinelearning-deeplearning

[5]. David E. Rapach , Jack K. Strauss " Forecasting real housing price growth in the Eighth District states" , Rangan Gupta*, Periklis♦

[6]. Vasilios Plakandaras+ and Theophilos Gogas "Forecasting the U.S. Real House Price Index"

[7]. Gupta and Das (2010) Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals

[8]. Rangan Gupta "Forecasting US real house price returns over 1831– 2013: evidence from copula models".

[9]. Patel, Chirag, Dipti Shah, and Atul Patel (2013) " Forecasting US real house price returns: A Survey ", International Journal of Computer Applications,69(9): 21-33.

[10]. Faqheri, Wizam Al, and Mashohor S Yamish (2009) "A Real Time Malaysian data using hybrid fuzzy" in the proceedings of International Journal of Computer Science and Network Security. 9(2): 333-340.

[11]. Azad Reza, Davami Fatemeh ,Jeo Jung, and Hamid Reza Shayegh.(2014) "New framework based on complementary methods for efficiency and accuracy" in the proceedings of 6th Conference on Information and Knowledge Technology (IKT): 171– 176.

[12]. Ashtari, A. H., M. J. Nordin and Seyed Mostafa Mousavi Kahaki (2011) "A new reliable approach for colour images" in the proceedings of the 2011 International Conference on Electrical Engineering and Informatics: 1 – 5.

[13]. B. Pechiammal, J., and RenjithArokia (2017) "An efficient approach for Machine learning system" in the proceeding of Third International Conference on Science Technology Engineering & Management : 121-129.

[14]. Praphananurak, Kitipong, Sanghitkul Winai, ChovichienVoraniti and Watana chaturaporn Pakorn (2017) "A framework for origin-destination estimation using regression method" in the proceedings of 9th International Conference on Information Technology and Electrical Engineering : 1 – 5.

[15]. Yao, Donghui, Zhu Wenxing, and Chen Yanjun. (2017) "Chinese character recognition based on convolution neural network" in the proceedings of Chinese Automation Congress : 1547- 1552.

[16]. Prabhu, B. Sachin, Kalambur Subramaniam, and Sitaram Dinkar (2017) "Recognition from live stream videos" in the proceedings of 2017 International Conference on Advances in Computing, Communications and Informatics: 2359-2365.

[17]. Björklund Tomas, Fiandrotti Attilio, and Annarumma Mauro (2017) “With convolutional neural networks trained on synthetic data” in the proceedings of 19th International Workshop on Multimedia Signal Processing : 1 – 6.

[18]. Sérgio Montazzolli Sliva, and Clausio Rosito Jung (2017) “Real-Time Brazilian Using Real House Price Index DeepConvolutional Neural Networks” in the proceedings of 30th SIBGRAPI Conference on Graphics, Patterns and Images : 55- 62.