Machine Learning

Name: Mohamed Nabil Aly

ID: 40-7226

Assignment 1

Objective

The main objective of the assignment is to learn what affects the price of houses using data given in form of csv and accurately predict the price of houses given a set of properties.

Filtering the data

The data stored in the csv consist of many different aspects that are related to the house some of which have great impact on the price and others have less impact, these values are identified using correlation technique by finding the correlation of each field in the csv with the price of each house using the corr() method. Then by visually analyzing the output of the correlation the decision is made to drop all fields that show small impact on the price (have a correlation factor of 0.3). The rest of the values are then stored in a variable “X” where the price is also removed as it is the value intended to be predicted, the price is also stored in another variable “Y” to enable us to train the system.

Adding degree

In this step the data chosen are again filtered to find the fields that show the most impact and adding a degree to them. The data is divided into 4 categories, starting from the 2nd degree which contains fields that showed a correlation factor of 0.4, moving to the 3rd degree which holds fields that had a correlation factor of 0.5 and the 4th degree which is decided by a correlation factor of 0.6 and finally the 5th degree which contained fields that had correlation factor of 0.7. These fields are then added again to “X” as a new field, this is done to infasize on the fields that are mast affective.

First sampling technique

Dividing the sample into 60 20 20

This sampling technique is done to divide the set into the set into 3 parts, the first part is 60% of the total sample set is used to train the machine “X_train”, the second set is 20% of the total sample set is used for cross validation “X_cross” and the final set is also 20% which is used as a test sample test the production of the machine. This is done by finding the total samplesize and multiplying it by 0.6 to find the X_train and then by 0.2 to find the X_cross and by 0.2 to find the X_test, these steps are done on both the “X” and the “Y” as shown below.

```
X_train = X[0:(math.ceil(len(X)*0.6)),:]

Y_train = Y[0:(math.ceil(len(X)*0.6)),:]

X_cross = X[(math.ceil(len(X)*0.6)):(math.ceil(len(X)*0.8)),:]

Y_cross = Y[(math.ceil(len(X)*0.6)):(math.ceil(len(X)*0.8)),:]

X_test = X[(math.ceil(len(X)*0.8)):(math.ceil(len(X))),:]

Y_test = Y[(math.ceil(len(X)*0.8)):(math.ceil(len(X))),:]
```

Training the machine

The X_train array is then inputted into the gradientDescentMulti() method which trains the machine using initial thata which is initialized at the beginning as an array of zeros of the same size as the X_train and out puts a the final thata after iteration and the cost J of each iteration.

After training the machine computeCostMulti() function is used to calculate the effectiveness of the train data set.

This function first takes the X_cross and computes the cost for linear regression.

Then this function is used again to compute the cost for linear regression using the test set X_test.

Comparing the results

To ensure that these steps are effective n training, these steps are run on firstly the whole data set, then the data set after filtering without degree, then the data set after filtering and with degree.

- This showed that removing some of the field that are uncorrelated with the price had less effect on the cost J.
- This also showed that adding degree on the most correlated fields had a great effect on the cost J

K-fold sampling technique

The k-fold method is another sampling method that divides the data set into K sets, in my case K is initialize as 4, then a combination of K-1 of the resulting sets are combined and used as the train set and one set is set to be the test set.

Then a regularization techniques is tried in attempt to improve the cost then after applying lambdas for regularization and testing, there is no improvement found in the results, which mean that there is no over fitting present in the model as there is no need to apply regularization methods.

Assignment 2

Objective

Given a table to predict the value of the third column given the first and the second column.

Changing the data

Firstly, the fields are checked for correlated, the first field is found to be slightly more correlated to the third field more than the second field, this caused the introduction of degree to the first field, sense both fields show a good correlation with the second field there is no dropping of fields.

sampling technique

Dividing the sample into 60 20 20

The sample set is then divided into 3 sets with the train set is 60% of the sample size and the cross variance is 20% and finally the test is the last 20% of the sample set.

The cost function is based on logistic regression, after computing the cost from the cost function scipy library is used to optimize any given function and the goal is to minimize the cost therefore scipy is used to minimize the cost function. Then the optimum cost function is given to the prediction function to predict the output.

Results

The best predictions are found when no change in the data is implemented and the functions are left to predict on the original data. The results of the prediction are accurate by 90% which is done by comparing the true data with the predicted data and finding the true negative and true positive and dividing them with the total test sample size.