Assignment report

Introduction

Machine learning is the field of artificial intelligence that is concerned with the automatic detection of patterns where the machine is able to learn and predict data based on experience rather than following a specific algorithm. Supervised learning is the class of machine learning algorithms that is concerned with predicting the output of a specific input based on a training process where the predicted output is compared to the actual output of the input. The cost of the predicting function, also known as the hypothesis, is calculated as the average sum of differences between the hypothesis and the actual output and this cost function is differentiated with respect to the parameters of the hypothesis function to calculate the gradient descent in order to be able to update the parameters, theta, of the hypothesis function. The degree of the polynomial of the hypothesis function can change to suit the output pattern and create better efficiency of the algorithm. The purpose of the assignment of this report is to find the polynomial equation that creates the best output while avoiding the problems of overfitting and underfitting. In this report I will discuss the methods used to avoid overfitting and underfitting while also providing an acceptable error rate.

Problem definition

The problem in this assignment was to find the suitable polynomial equation for the hypothesis function and also test it for the generalization degree of the model so it can be applicable to other data. Furthermore, using regularization technique to add a penalty term to different features in order to reduce the parameters of some features.

Methodology

Model selection

The proposed idea of model selection is to divide the given dataset into three different datasets, 60% for training the model, 20% for the validation dataset and 20% for the testing the dataset. The dataset needs to be large so the training dataset will be large enough to train the model on 60% of the data, the training dataset. The training dataset will be used to train the model with different hypothesis equations according to theta and the features of the given dataset. The cost calculations and gradient descent steps are recorded and repeated for a given number of iterations where the parameters of the hypothesis function are updated after each iteration. After all the iterations, the recorded costs of each iteration are plotted against the number of iterations. After the training process, the different hypotheses are tested using the validation dataset for cross-validation to know which hypothesis function secures the best error rate. The final step is to test the chosen hypothesis equation using the testing dataset to ensure the generalization of the model and the model is not suffering from overfitting or underfitting

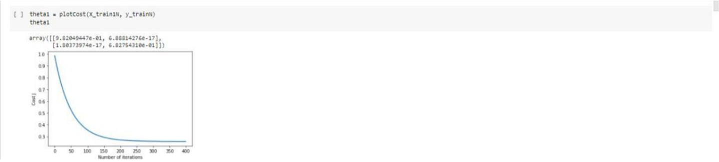
Regularization

Increasing the number of features used in training will result in better fitting, however, it may increase the risk of overfitting of the model. Regularisation was used to add a “penalty term” to create different weights to the effect of each parameter of the hypothesis function while simultaneously increasing the complexity of the function.

Regularization is performed by adding a regularization term to the original cost function of the model which represents the average sum of the model parameters squared parameters of the hypothesis function multiplied by the regularization factor, lambda. Regularization will make us able to increase the complexity of the model without increasing the risk of the model suffering from overfitting.

Results

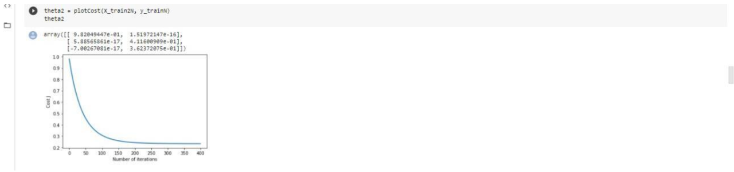
Plotting cost function using one hypothesis parameter



```
[ ] theta1 = plotCost(X_train1N, y_trainN)
    theta1

    array([[9.82049447e-01, 6.88814276e-17],
           [1.80373974e-17, 6.82754310e-01]])
```

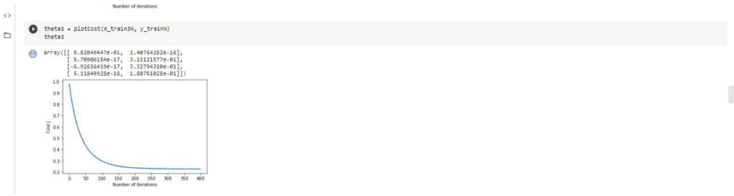
Plotting cost function using two hypothesis parameters



```
theta2 = plotCost(X_train2N, y_trainN)
theta2
```

```
array([[ 9.82049447e-01,  1.51972147e-16],
       [ 5.88565861e-17,  4.11600909e-01],
       [-7.00267081e-17,  3.62372075e-01]])
```

Plotting cost function using three hypothesis parameters



```
theta3 = plotCost(X_train3N, y_trainN)
theta3
```

```
array([[ 9.82049447e-01,  1.40764282e-16],
       [ 5.70906154e-17,  3.13121577e-01],
       [-6.92636419e-17,  3.32794310e-01],
       [ 5.11849928e-18,  1.80751028e-01]])
```

Conclusion

In this report I discussed the use of model selection techniques and regularization for model model diagnostics and evaluation of errors. The use of model selection technique and dividing the given dataset for model training, validation and testing are used for tuning the model while the regularization technique was used to increase the complexity of the model while containing the increase in overfitting.