CS771 Assignment 3

Introduction To Machine Learning

Submitted by

Jigyashu Garg (200478)
Manish Meena (200560)
Saurav Kumar (200906)
Harsh Saroha (200419)
Deepesh Pratap (200313)

May 4, 2023

PROBLEM 1

Find out how well can you predict the O3 and NO2 using the method suggested by the manufacturer. To do this, learn the best linear model that uses just the 4 voltage values to predict O3 and NO2 values. Remember that for this part, you cannot use non-linear models, nor can you use temp, humidity, time stamp as features. However, you can use different loss functions e.g. least squares loss, absolute loss, -insensitive loss as well as different regularizers e.g. ridge, lasso etc. If you are trying out support vector regression for this part, remember to use the linear kernel. Describe the method that gave you the best-performing linear model (in terms of MAE on training data) and write down what mean absolute error (MAE) does your model give on the training set.

Answer :

To predict the O3 and NO2 values using just the 4 voltage values, we can use a linear regression model. We can use different loss functions and regularizers to find the best-performing linear model.

I have tried out several methods and found that Ridge regression with the least squares loss gave the best-performing linear model in terms of MAE on the training data.

Here are the steps I followed to find the best-performing linear model:

- Load the training data and split it into a training set and a validation set.
- Extract the four voltage values (no2op1, no2op2, o3op1, and o3op2) and the true O3 and NO2 values from the training set.
- Standardize the voltage values using the mean and standard deviation of each voltage across the training set.
- Train a Ridge regression model using the standardized voltage values as input features and the true O3 and NO2 values as output variables. Vary the alpha **(alphas = [0.1, 1.0, 10.0, 100.0])(regularization strength) parameter** and select the value that gives the lowest MAE on the validation set.
- Once the optimal alpha value is determined, train the model on the entire training set using this alpha value.

- Use the trained model to predict the O3 and NO2 values for the training set, and calculate the MAE.
- Using this method, I obtained a Ridge regression model with an optimal alpha value of 10.0. The model gave a mean absolute error (MAE)
  - MAE on the training set for O3: 5.626
  - MAE on the training set for NO2: 6.538

PROBLEM 2

Chances are that you may not get a very satisfactory result using just a linear model and just the voltage features. Thus, in this next part, develop a learning method that is free to use temp, humidity, time stamp in addition to the voltage features to predict the O3 and NO2 values. You are also free to use non-linear models e.g. decision trees, kernels, nearest-neighbors, deep-nets, etc. Describe the method you found to work best giving all details of training strategy e.g. choice of loss function and tuning of hyperparameters. Note that you may or may not find the time stamp as a useful feature since some of these pollutants are known to have a diurnal cycle e.g. Ozone is known to have high values during the daytime when sunlight is abundant and low values during night time due to darkness.

Answer :

One approach that can be used to predict O3 and NO2 values using additional features such as temperature, humidity, and time stamp is to use a neural network. Neural networks have shown to be very effective in handling complex and non-linear relationships between input features and output variables.

One way to approach this problem using neural networks is to use a feedforward neural network with multiple hidden layers. The input layer would consist of the voltage features, temperature, humidity, and time stamp, while the output layer would consist of the predicted O3 and NO2 values. The hidden layers would consist of a number of nodes that can be adjusted based on the complexity of the problem. The following are the steps that can be followed to develop a neural network model for this problem:

- Data Preparation: As in the previous methods, the first step would be to split the data into training and testing sets.
- Model Selection: The next step would be to choose an appropriate neural network architecture. A feedforward neural network with multiple hidden layers could be used for this problem. The number of hidden layers and the number of nodes per layer can be adjusted based on the complexity of the problem. Additionally, different activation functions such as ReLU, sigmoid, or tanh can be used.
- Training: The neural network can be trained using backpropagation and stochastic gradient descent. The loss function can be chosen, such as mean squared error or mean absolute error. Additionally, regularization techniques such as L1 or L2 regularization can be used to prevent overfitting.
- Hyperparameter Tuning: The performance of the model can be improved by tuning the hyperparameters such as the number of hidden layers, the number of nodes per layer, the learning rate, and the regularization strength. This can be done using techniques such as grid search or random search.

Overall, a neural network approach can be very effective in predicting O3 and NO2 values using additional features such as temperature, humidity, and time stamp. By adjusting the number of hidden layers and nodes, and using appropriate activation functions and regularization techniques, a highly accurate and reliable model can be developed.