Machine Learning CSE 574

Project 2

Team

Sunil Umasankar (UBITName = suniluma, personNumber = 50249002)

Abhishek Subramaniam (UBITName = a45, personNumber = 50244979)

Prajna Bhandary (UBITName = prajnaga, personNumber = 50244304)

Overview

This project intends to understand the process of creating a linear regression machine learning model to predict the future outputs given new set of inputs.

The data given are synthesized data (10 features) from an unknown mathematical formula and LeToR data from real life observations (46 features). The output for these data inputs are given.

We use linear regression with Gaussian Radial Basis functions as our learning model. We partition the data set into a training, testing and validation set and train the model on the training set, then use the validation set to tune our hyper parameters. Then we calculated the RMS error on the test data to indicate the correctness of our model.

We implement two methods, namely closed form solution and stochastic gradient descent solution. These are explained below.

Implementation

Shuffle

We shuffle the input and the output dataset, and make the first 80% of the dataset our training set, the next 10% our validation set, and the last 10% as our testing set. Shuffling the data helps us achieve better results for the upcoming steps.

Design Matrix

We calculate the average k-means centers (using which we calculate spreads) over 30 runs of k-means, initialized with random centers, and the number of centers (equal to number of basis functions) starting from 2 and incrementing by 1. We use the sci kit learns in-built function to get the k-means results.

We map these results to build our design matrix (values of phi).

Since phi depends on the number of basis functions we choose, we perform the above steps repeatedly to find the number of basis functions for which our model performs best.

Training

Closed form Solution:

This is one of the method of finding the model parameter w. This method attempts to find the value of w for which the sum of squares error is the least.

Using the obtained design matrix and a fixed regularization constant lambda (which we have set to 0.1), we obtain the closed form solution for the problem using the formula.

$$\mathbf{w}^* = (\lambda \mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{t}$$

We vary the number of basis functions till we get the best value for w.

Stochastic Gradient Solution:

We try to obtain a similar result we got for our closed form solution using stochastic gradient descent. Here, we will run the gradient descent using learning rate = 0.01, and wait until the algorithm converges at the local minima. We also perform early stop that will stop the descent (before it reaches num_epochs = 10000) if the difference between errors on 2 consecutive runs of the loop is very low.

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta \mathbf{w}^{(\tau)}$$

Where $\Delta \mathbf{w}^{(\tau)} = -\eta^{(\tau)} \nabla E$

Validation

Every time we obtain a value of w for the hyper parameters we set, we calculated the validation error using sum of squares error. If the error is the lowest we have encountered so far, then the current number of basis functions fits the data better. We do this repeatedly until the minimum error doesn't vary for a set number of times. We assume the value of w we get after this to be the best fit model parameter.

Testing

Once we find the best model parameter w, we run our model on the test data and find the testing error. The error is we use here the Root Mean Square (RMS) error.

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^*)/N_{\mathrm{V}}}$$

Observations

For the assumed hyperparameters:

Lambda = 0.1

Learning rate = 0.01

Synthetic Data Observations	Closed Form	Stochastic Gradient Descent
No. of Basis Functions	5	2
Sum of Squares Error	572.183870918	618.133598868
Error RMS Validation	0.756428364697	0.786214728219
Error RMS Test	0.749081471125	0.775518209669

LeToR Data Observations	Closed Form	Stochastic Gradient Descent
No. of Basis Functions	30	29
Sum of Squares Error	1101.48889088	1104.84222131
Error RMS Validation	0.56659090069	0.56337571147
Error RMS Test	0.572647629022	0.572488501703

Note: All our observations are subject to change because of randomness in the k-means averaged centers and spreads and also the entire data is randomly shuffled each run, so the part of the data that is part of the training, validation and test set keep changing.

Inferences:

From the observations above, we see that the RMS error of the validation set is very close to the RMS error of the test set, for both methods. Also, we can see that our model gives us comparably good results for the LeToR data and the synthetic data. This proves the generality of our model.

The LeToR data produces a large value of the number of basis functions, this may be because the output depends on a large number of features.

The closed form solution produces a more accurate result for model but takes longer to settle on the correct value of m compared to the stochastic gradient descent approach.

We rounded up the predicted output to the nearest decimal to compare it with the test set output. We observed that 5166 out of 6963, and 5170 out of 6963 predicted output matched the test set output for the LeToR closed form and SGD solutions respectively.

Output

```
CLOSED FORM FOR SYNTHETIC DATA

Optimal M: 5

Minimum error found in Validation set:  572.183870918

Minimum RMS error found in Validation set:  0.756428364697

RMS in test data is:  0.749081471125

Min w:

[  0.84652415   7.46373505   0.09928049   5.70342219
12.50735065]

Predicted Y:

[ 0.86231362  0.86283305  0.90453179 ...,  0.84868824
0.84685651

  0.93322784]

Original Y:

[[ 0.  0.  1. ...,  1.  2.  0.]]


Number of records matching in test data after rounding to the
nearest decimal: 819 out of 2000


SGD FOR SYNTHETIC DATA

Optimal M: 2

Minimum error found in Validation set is:  618.133598868

Minimum RMS error found in Validation set is:  0.786214728219
```

RMS in test data is: 0.775518209669

Min w:

[0.95249978 0.64098207]

Predicted Y:

[0.95286878 0.95467568 0.96090089 ..., 0.95409414 0.95344297

  0.9632604]

Original Y:

[[0. 0. 1. ..., 1. 2. 0.]]

Number of records matching in test data after rounding to the nearest decimal: 779 out of 2000

CLOSED FORM FOR LETOR DATA

Optimal M: 30

Minimum error found in Validation set: 1101.48889088

Minimum RMS error found in Validation set: 0.56659090069

RMS in test data is: 0.572647629022

Min w:

[0.31947137 -0.11134069 1.65757655 0.20138838 0.80200646 -0.74597096

  1.13157199 -1.1047003   0.5767907 -1.30597851 0.03425102 -1.16709647

  2.07965175 1.0883203   2.57702922 -2.93705112 -0.98335937 -2.14703266

-0.15667077 0.01899197 -2.26037514 -1.3050102   4.01757476 0.40857519

-0.51586423 0.40070451 1.41123005 -0.38863431 0.08596718 0.19616162]

Predicted Y:

[-0.03033255 0.31918996 0.31947134 ..., 0.31947137
0.31935173

  0.31946923]

Original Y:

[[0. 1. 0. ..., 0. 1. 0.]]

Number of records matching in test data after rounding to the nearest decimal: 5166 out of 6963

SGD FOR LETOR DATA

Optimal M: 29

Minimum error found in Validation set is: 1104.84222131

Minimum RMS error found in Validation set is: 0.56337571147

RMS in test data is: 0.572488501703

Min w:

[3.05092146e-01   3.92423958e-05   4.63950491e-05 -
3.34881858e-06

   2.21294514e-06   7.11550626e-07   1.51318114e-04
2.91721368e-06

  -7.66023627e-07 -7.65359463e-05   6.78934084e-06
1.27875812e-04

   1.29385821e-04   1.24814758e-05   3.56202506e-06 -
2.46711750e-03

   1.21191533e-06   8.30146135e-06   2.13334677e-04
1.95831514e-04

   1.51319611e-05 -2.93378890e-06 -7.10142303e-05
3.41574930e-07

   7.45867253e-05   1.43188969e-05   2.16093466e-06 -
7.30418552e-06

   2.25662687e-05]

Predicted Y:

```
[ 0.30509215  0.30509215  0.30509215 ...,  0.30509215
0.30509227   0.30509215]

Original Y:

[[ 1.  0.  0. ...,  0.  0.  0.]]


Number of records matching in test data after rounding to the
nearest decimal: 5150 out of 6963
```