Assignment ML

Piyush Bhatore

April 16, 2017

Code Explanation

Given topics explains about the files submitted a s code

0.1 train_net

- Load th data into a pandas dataframe to process it
- remove the id column (because we don't need it here)
- declaring the varibles like dimension of the matrixs which are useful incode
- doing the data featuring by normalising the data which is in value
- removing the '?' by replacing it with the mode of particular column
- also emumerating the objects field (it alloows the data to get train)
- initialised the weights array
- in the for loop first the forward propagatioin is done and then the backward propagation followed by updataing the weights
- converting the hidden and output wweights to a single array and saved it to weights.txt

0.2 test_net

- Load th data into a pandas dataframe
- doing the data featuring by normalising the data which is in value
- removing the '?' by replacing it with the mode of particular column
- also emumerating the objects field
- loaded the test data and the weighted array
- convert it to correct datset
- added bias and done forward propagations
- printing output to predictions.txt

0.3 train_net

- Load th data into a pandas dataframe
- doing the data featuring by normalising the data which is in value
- removing the '?' by replacing it with the mode of particular column
- also emumerating the objects field
- loaded the test data

- convert it to correct datset
- added bias and done forward propagations for three methods described in next page
- printing output to predictions.txt for all three prediction

Observation

the train data which was given to us was highly unbalanced so I also tried it to balance by takeing all the ones(salary) and 1.3*ones zeros() salary this balanced the data . This is done in lines 8-11 so please comment if you use new training data set

SVM is better than other predictors (when I decreased the number of zeroes in train data) though it takes a lot of time to get trained

Score are as follows(zeros removed) Logistic Regression - 0.79339 (kaggle score) takes very less time
Gaussian Naive Bias - 0.72614 (kaggle score) takes very less time
SVC - 0.81417 (kaggle score) takes a little time
Neural net - 0.80797 (kaggle score) takes a little time

Without filtering the data neural network works the better that others which means it does not over fit the data

the train data which was given to us was highly unbalanced so I also tried it to balance by takeing all the ones(salary) and 1.3*ones zeros() salary this balanced the data . This is done in lines 8-11 so please comment if you use new training data set