Classification of Active Sites using Neural Network

Table of Contents

template by: AhmetSacan. [[Write your name here]]

In this assignment, you are going to predict catalytic residues in proteins using sequence and structural information. The dataset (courtesy of Natalia Petrova) is a subset of the data used in "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties", Natalia Petrova and Cathy Wu, 2006. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-312

Please review that publication to learn more about this dataset and the catalytic residue prediction problem.

Load the data

Get the data using the loadcatsite() function. No changes needed for this section.

General Guidelines

- Do not show the Network GUI. You may show the GUI when you are developing and testing your code; your final code and report should not contain the GUI.
- Do not use any validation set when training the network.
- You may use any number of hidden units in your final code. You may try different numbers of hidden units to decide how many hidden units you want to use, but you do not need to include code to that effect.

Train a network for classification using all features

Use only one of the training & test subsets of the cross-validation sets when reporting training & test accuracies in this section.

- Setup the cross-validation indices
- Use only the first cross-validation set to divide the dataset into training & testing
- Create the network, set its parameters, and train on the training set.
- Test the network on the test data.

- Report the test accuracy.
- No repetitions needed for this section.

```
testaccuracy =

    0.5635
```

Train a network for classification using a subset of the features

Create a subset of the dataset that includes only the following features: amino acid P, amino acid V, nearest_cleft_distance, and ScoreConsScore.

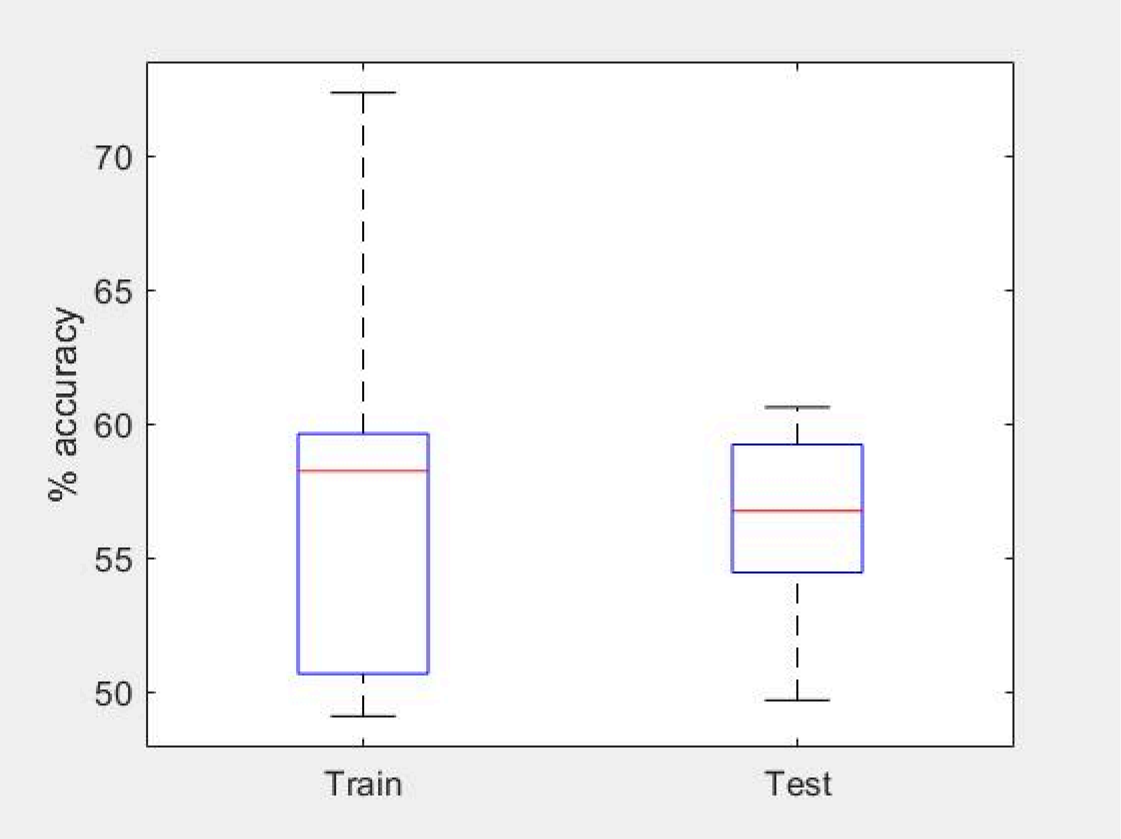
Report cross-validated training and test accuracies on this new dataset. You must report average accuracies across multiple repetitions and across all cross-validation sets.

- foreach repetition:
- ...Setup the cross-validation indices
- ...foreach cross-validation set:
- ......Train & test the network
- ...Calculate and store the training and test accuracies for this repetition.

After you complete all repetitions,

- Report the average training accuracy across all repetitions.
- Report the average test accuracy across all repetitions.
- Show distribution of training and test accuracies across all repetitions and cross-folds in a boxplot.

```
Average training accuracy: 0.57%
Average test accuracy: 0.56%
```



Published with MATLAB® R2018b