Table of Contents

Abstract. This assignment represents the final project of SPM.

1 Introduction

With the development of technologies, a number of data rise day by day.
With the large various data nowadays, tasks
The remaining of the assignment is constructed: the Section 2 describes about the reprocessing of the data sets. The implementation of the assignments is represented in the Section 3, while the two last Section 4 and 5 are about the result and conclusion.

2 Reprocessing

In this part of the report, the reprocessing of data sets (Monk and Competition) are described in two subsections.

2.1 Monk Dataset

2.2 Competition Dataset

In the competition data set including two sets: training and testing, both of these sets need to be reprocessed before applying the neural network. Firstly, the description and the first column (id attribute) in these sets are removed. After that two last columns (the label attributes) and the remaining attributes in the training set are split into two sets: labels and training features.

3 Implementation

The implementation is explained in this part of the report with details of general neural network approach and the way applying in two data sets.

3.1 Neural Network

In general, the back-propagation, momentum and regulization are applied to build this approach. In details, the approach uses only one hidden layer in the NN. However, the number of units in this layer is tried with variety of number.
Grid search technique is also applied with different parameters such as: learning rate, hidden unit (the number of hidden unit in the hidden layer)

3.2 Monk Dataset

3.3 Competition Dataset

For the evaluation of parameters and the accuracy of the model, the double cross-validation (CV) technique is applied.
With this technique, the training competition data set is split into “k” training data parts and testing data (this is the first parameter in CV). Each of “k” part data is applied the approach from the Section 3.1 to generate the Least Mean Square Error (LMS). From these LMS, the best hyper parameters are chosen and apply on the whole data set. At this time, the second CV is applied with a different “k” folds parameter. It is similar with the first CV, but in this case the hyper parameter from the first CV is apply to whole data set and generate the accuracy. Finally, the blind data set uses this model to archive the label or target attributes.
With this technique, the training competition data set is split into “k1” couple training and testing parts. From each couple training and testing part, the training part is applied with another CV and a “k2” parameter to generate “k2” couple training and validation parts. After training, the best “k1” hyper parameters from the “k1” * “k2” are found. From these “k1” hyper parameters, the “k1” accuracy of couple training and testing parts are generated.
Then the best hyper parameter is chosen to apply the whole data set before predicting the separated blind test set.

4 Result

4.1 Monk Dataset

4.2 Competition Dataset

5 Conclusion

The assignment exploits WebGraph framework to solve some statistics tasks and experiments from the paper [**?**]. With these works, the framework shows their useful techniques and tools which simply exploits large graphs.

6 Ackowledgements

I would like to show my gratitude to Prof. Marco Danelutto about lessons. Finally, I would also like to extend my thank to LaTeX and Springer for this format.