Report on Homework 3

CS420, Machine Learning, Shikui Tu, Summer 2018

Zelin Ye 515030910468

1 SVM vs. Neural Networks

1.1 Introduction

In machine learning, SVM (Support Vector Machine) is a commonly used classfication method due to its high efficiency and accuracy. Recent years, the neural network has been attracting more and more attention, and also used to solve classification problems. In this homework, I would investigate the performances of SVM and neural network (e.g. MLP) on some classification datasets under different experimental settings (e.g. pass).

1.2 Methodology

In this section, I would introduce the datasets and models in my experiments.

1.2.1 Datasets

In my experiments, I use two datasets that are from **LIBSVM Data** [1]. One is **madelon**, a binary classification dataset with 500 features, 2000 training samples and 600 testing samples. Another is called **satimage**, which is for multi-class classification and has 36 features, 6 classes, 3104 training samples and 2000 testing samples. More details about the two datasets can refer to Appendix A.1.

1.2.2 Models

For neural network, considering the complexity of features and scale of samples, I choose MLP (Multilayer Perceptron) instead of popular DNN or CNN.

In experiments, I would investigate the performances of MLP under different architectures or parameter settings (e.g. number of hidden layers or hidden neurons).

1.3 Experiments and Results

1.3.1 Preprocess

Most datasets are likely to have missing data, and those in LIBSVM Data are no exception. Therefore, I first make up for the omission in the datasets, replacing empty data with corresponding mean values. Afterwards, I convert the labels from numbers to one-hot vectors for the calculation of loss.

A Appendix

A.1 Details of Datasets in Experiments

A.1.1 Madelon [2]

Madelon is an artificial dataset, which was one of five datasets used in the NIPS 2003 feature selection challenge. This is a two-class classification problem with continuous input variables.

The difficulty is that the problem is multivariate and highly non-linear.

A.1.2 Satimage [3]

Satimage dataset is also called **satalog** dataset. It contains multi-spectral values of pixels in 3×3 neighbourhoods in satellite images, and the classification associated with the central pixel in each neighbourhood.

As a classification dataset, the aim of satimage is to predict the classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number.

References

[1] "Libsvm data: Classification, regression, and multi-label." `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

[2] "Madelon data set." `http://archive.ics.uci.edu/ml/datasets/Madelon`.

[3] "Statlog (landsat satellite) data set." `https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)`.