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] to investigate the performances of SVM and neural network. One is **splice**, a binary classification dataset with 60 features, 1000 training samples and 2175 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.

Additionally, I choose a dataset called **ConvexNonConvex** from LISA [2] to conduct comparison between SVM and deep learning algorithm benchmarks. It is a binary classification dataset that contains 784 features, 8000 training samples and 50000 testing samples.

More details about these 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.

1.4 SVM

1.4.1 Binary Classification

1.4.2 Multi-class Classification

1.5 Neural Network

1.5.1 Binary Classification

1.5.2 Multi-class Classification

1.6 Comparison Between SVM and Deep Learning Algorithm Benchmark

A Appendix

A.1 Details of Datasets in Experiments

A.1.1 Splice [3]

Basically, according to the biological knowledge, splice junctions are points on a DNA sequence at which ‘superfluous’ DNA is removed during the process of protein creation in higher organisms.

Splice dataset aims to recognize two classes of splice junctions, given a DNA sequence, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out).

A.1.2 Satimage [4]

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.

A.1.3 ConvexNonConvex [5]

The ConvexNonConvex dataset consists of convex regions with pixels of value 255. Such regions are constructed via the intersection of a number of half-planes whose location and orientation were chosen uniformly at random. In the meanwhile, the number of half-planes is also sampled randomly according to a geometric distribution with parameter 0.195.

To ensure the validity of the data, a candidate convex image is rejected once there are less than 19 pixels in it.

References

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

[2] “Lisa.” http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/WebHome).

[3] “Molecular biology (splice-junction gene sequences) data set.” http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29.

[4] “Statlog (landsat satellite) data set.” https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite).

[5] “Convexnonconvex.” http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/ConvexNonConvex.