# Statistical learning: 4th assignment

Ali Zamani(96123035) & Aryan Nasehzadeh

January 20, 2019

# 1 Machin leraning algorithms

In this section, we will use different machine learning algorithms such as LDA, QDA, KNN, Decision tree, and kernel SVM. To access source code you can check the link:
https://github.com/zamaniali1995/Machine-learning-and-Neural-Network
The sklearn package is used. Dataset has numbers of unknown features, one way is dropping all data with the unknown feature but in this approach, we lose a lot of data thus we should use other approaches. We replace unknown features with mean or min or max of features on that class, for example, if the unknown feature is associated with data in class 1, we replace that feature with the mean of features on class 1. we also train our model with scaled and unscaled data and results are reported in tables 1, 2, 3 , 4 , 5. In the SVM algorithm non-scaled data are used. As you see maximum accuracy belongs to **polynomial kernel SVM** and it is **58%**.

# 2 Neural Network(NN)

In this section, we will create an NN model with 3 hidden layers. We use shuffled and unscaled data. Table 6 shows results for different neurons numbers in each layer. Figure 3 show another result. As you can see in table 6 maximum test accuracy is 28% and it is less than 58% thus kernel SVM with polynomial kernel has better test accuracy.

# 3 Ideas

We have shown that if the concatenated dataset is used then we do not have good accuracy but if we use the first dataset as training we can have good accuracy. It is shown that our test dataset is correlated with first train dataset. We know that first dataset measures with one device and another dataset measures with another device we can add another feature two show that which device is used. We know that 36 class belong to 36 places that data is sampled so the data can be time series. according to this, we can use LSTM.
If we plot our test dataset we see that the number of class 16 is more than other classes so we faced unbalanced test dataset. according to this, if we make a model that predicts class 16, we have good accuracy in test dataset so it is better to use balanced test dataset and calculate the accuracy of the model.
We replace zeros with mean. Another approach can be used. For example, we can fit a natural cubic spline to our model and after that find value of unknown feature from the natural cubic spline.
To visualize data set a number of plots are placed at the end of the report.

Table 1: QDA

| QDA | | | |
|---|---|---|---|
| Train | Merged | Scaled | ACC |
| | Yes | Yes | 75% |
| | Yes | No | 92% |
| | No | Yes | 54% |
| | No | No | 96% |
| Test | Yes | Yes | 5.2% |
| | Yes | No | 27% |
| | No | Yes | 0% |
| | No | No | 23% |

Table 2: LDA

| LDA | | | |
|---|---|---|---|
| Train | Merged | Scaled | ACC |
| | Yes | Yes | 75% |
| | Yes | No | 92% |
| | No | Yes | 54% |
| | No | No | 96% |
| Test | Yes | Yes | 1% |
| | Yes | No | 27% |
| | No | Yes | 2.2% |
| | No | No | 47% |

Table 3: Decision tree

| Decision tree | | | |
|---|---|---|---|
| Train | Merged | Scaled | ACC |
| | Yes | Yes | 75% |
| | Yes | No | 92% |
| | No | Yes | 54% |
| | No | No | 96% |
| Test | Yes | Yes | 11% |
| | Yes | No | 25% |
| | No | Yes | 2.6% |
| | No | No | 25% |

Table 4: KNN

| KNN | | | |
|---|---|---|---|
| Train | Merged | Scaled | ACC |
| | Yes | Yes | 75% |
| | Yes | No | 92% |
| | No | Yes | 54% |
| | No | No | 96% |
| Test | Yes | Yes | 7% |
| | Yes | No | 39% |
| | No | Yes | 1.4% |
| | No | No | 40% |

Table 5: Kernel SVM

| Kernel SVM | | | |
|---|---|---|---|
| Train | Merged | Kernel | ACC |
| | No | linear | 96% |
| | No | polynomial | 96% |
| | No | rbf | 96% |
| | No | sigmoid | 96% |
| | Yes | linear | 92% |
| | Yes | polynomial | 92% |
| | Yes | rbf | 92% |
| | Yes | sigmoid | 92% |
| Test | No | linear | 51% |
| | No | polynomial | 58% |
| | No | rbf | 4% |
| | No | sigmoid | 0.2% |
| | Yes | linear | 28% |
| | Yes | polynomial | 25% |
| | Yes | rbf | 4% |
| | Yes | sigmoid | 0.2% |

Table 6: Neural Network

| Neural Network (3 layers) | | | | | | |
|---|---|---|---|---|---|---|
| Hidden 1 | Hidden 2 | Hidden 3 | Batch size | Learning rate | Train ACC | Test ACC |
| 40 | 30 | 20 | 40 | 0.001 | 82% | 28% |
| 50 | 40 | 30 | 40 | 0.001 | 88% | 26% |
| 50 | 40 | 10 | 40 | 0.001 | 79% | 26% |
| 40 | 40 | 30 | 20 | 0.001 | 90% | 24% |
| 200 | 150 | 100 | 40 | 0.001 | 97% | 26% |
| 100 | 90 | 80 | 60 | 0.0001 | 97% | 28% |

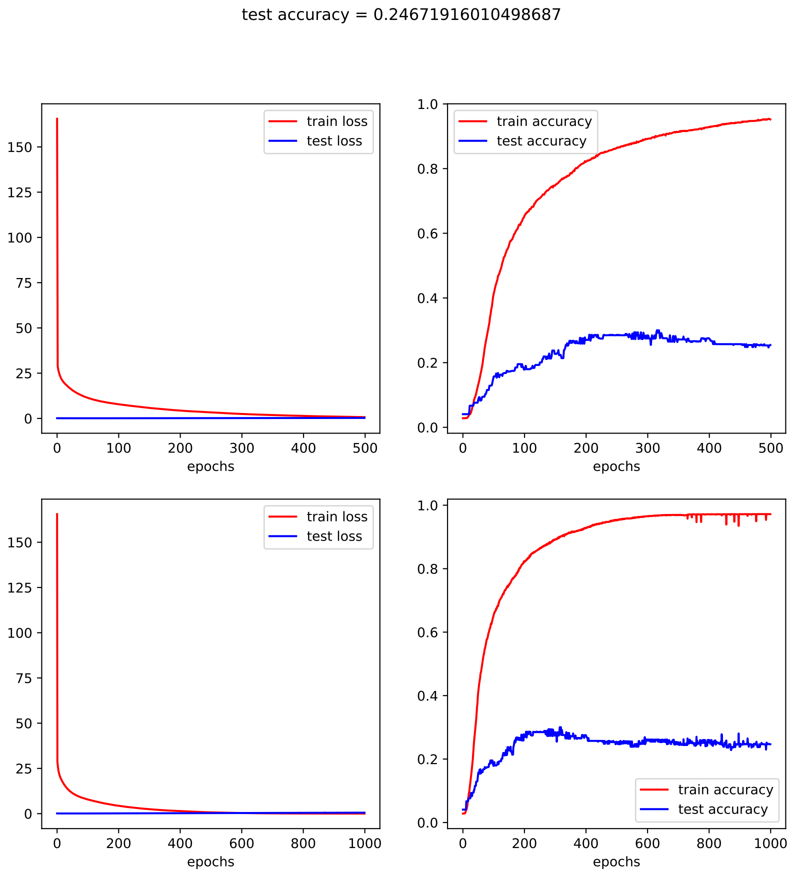

Figure 1: Neural Network

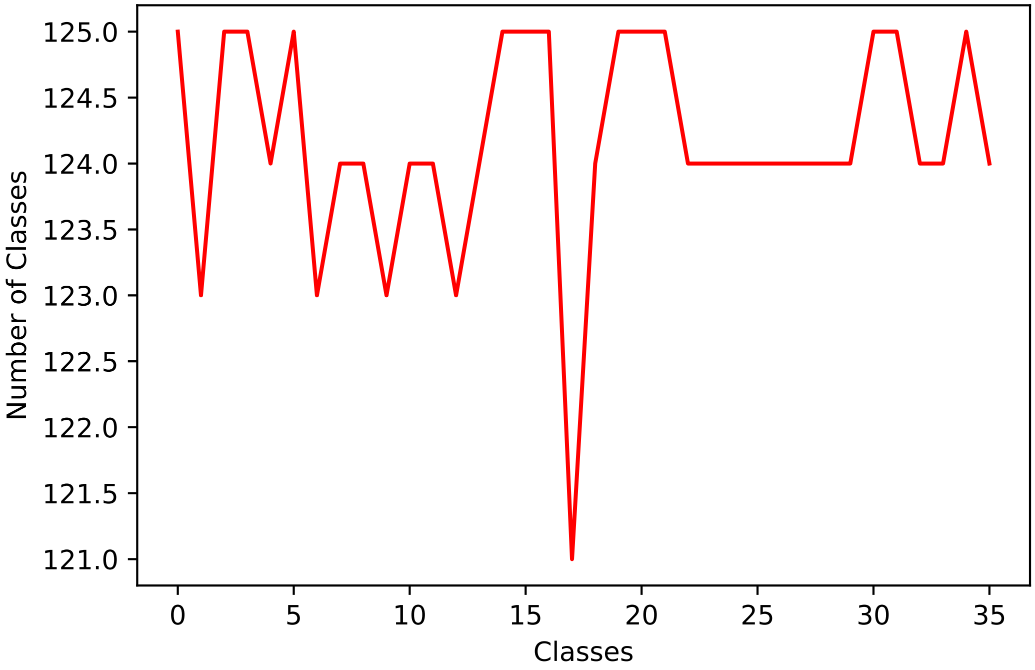

Figure 2: Number of data in each class

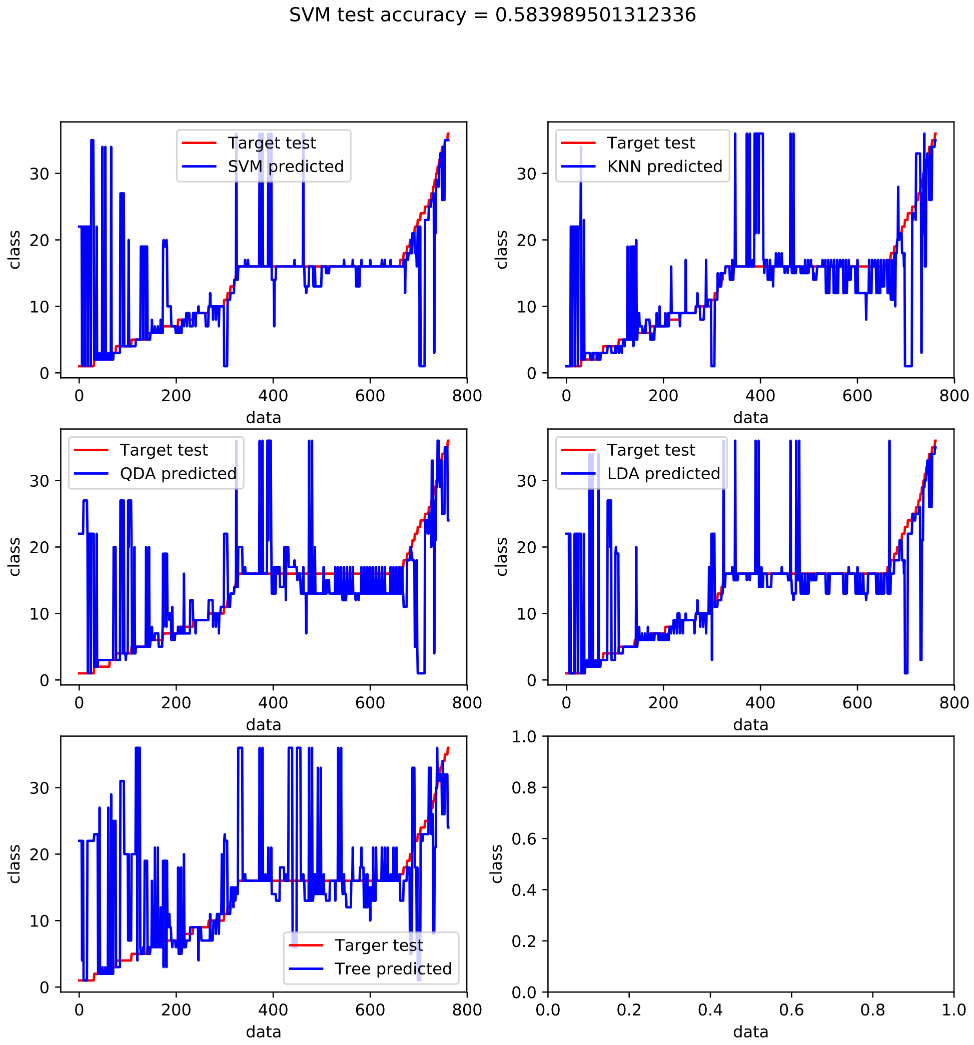

Figure 3: Predicted test in each algorithm

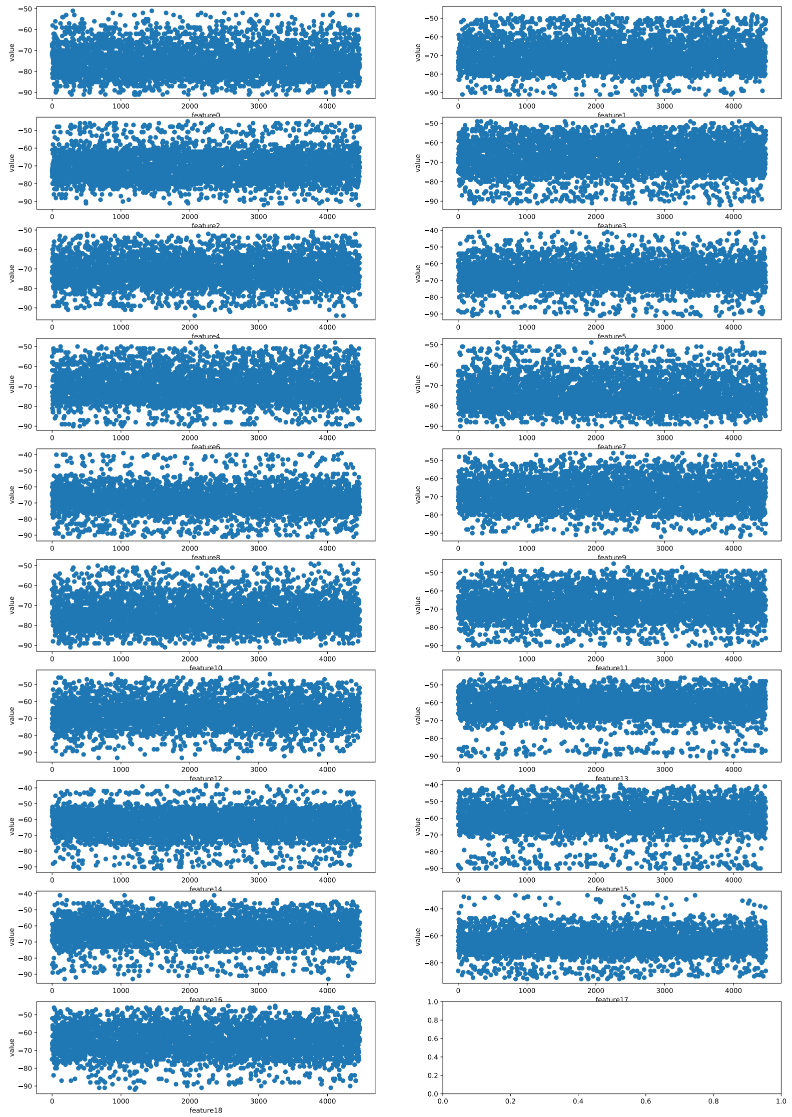

Figure 4: Value of each feature