Ex2 in Machine Learning: Comparison between SVM / Neural Network

# Overview

This exercise is targeting to compare between the two algorithms of SVM and Neural Network on different datasets.

The program can be split into the following major parts:

1. SVM learning process:
   1. Run SVM model on the dataset using different combinations of gamma and C, and find the best combination
   2. For doing this I am running loops on gamma and C with big steps. In the beginging I chose a big range (1 to 1000) with big steps (of 100), then later on I zoom into smaller range with smaller steps.

I stop either if I reach 10 times or if there are no meaningful improvement on the score anymore

1. Neural Network Learning process. We need to find the best combination of <layer, learning rate>.

Note: In this exercise we set the value of learning rate and we keep it constant

1. Comparison between the ‘best’ SVM model and the ‘best’ NN model.

# Datasets

For this exercise I use the following two datasets

## Iris Dataset

URL: <http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>

I selected two classes that are not linear separable : 'Iris-virginica' and ‘Iris-versicolor’. So total we have 100 entries after filtering. I keep all the 4 columns ("sepal length", "sepal width", "petal length", "petal width").

## Breast Cancer Dataset

URL: <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data>

In this dataset we have more than 35 columns. The first column is just an ID, so I ignore it. The second column ‘outcome’ has values of N (not infected) or R (infected) which will be our Y values. The third column is ignorable since it is based on column two.

There were four samples with ‘?’ (question mark letter) in the column ‘Lymph node status’, so I ignored them. That lead us to have a dataset of 194 samples and 32 features.

# Results

## IRIS Dataset

### SVM

|  |  |
| --- | --- |
| C range: [1, 1000], step size: 100  Gamma range: [1, 1000], step size: 100 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1937468337160196.png |
| C range: [0.1, 101], step size: 10  Gamma range: [0.1, 101], step size: 10 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1937515581800452.png |
| C range: [10.1, 30.1], steps 1.0  Gamma range: [0.1, 10.1] , steps: 1.0 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1937575711342596.png |
| C range: [24.1, 26.1], steps 0.1  Gamma range: [0.1, 1.1] , steps: 0.1 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1937635840884740.png |
| C range: [25.8, 26.0], steps 0.01  Gamma range: [0.1, 0.2] , steps: 0.01 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1937695970426884.png |
| Results: | Best C: 25.9  Best gamma: 0. 1  Best score (average): 0.985  C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1947471315992580.png |

### Neural Network

For finding the best parameters for neural network model I run through all the combinations of:

* Layer : we scan the range of [2, 50]
* Learning rate: we scan the range of [1/50, 1.0]. each time we take a step of 1/50.

The best score found was 0.98

Recommended rate is 0.02 and number of layers 39.

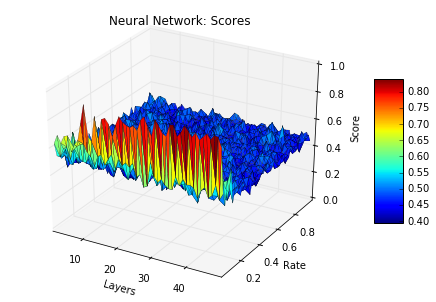


Figure 1: Neural Network scores per <rate, layer> (IRIS dataset)

## Breast Cancer Dataset

### SVM

|  |  |
| --- | --- |
| C range: [1, 1000], step size: 100  Gamma range: [1, 1000], step size: 100 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1947587280109572.png |
| C range: [301, 501], step size: 10  Gamma range: [601, 801], step size: 10 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1947634524749828.png |
| C range: [391.0, 411.0], steps 1.0  Gamma range: [631.0, 651.0] , steps: 1.0 | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1951418390937604.png |
| Results: | No further improvement on the best score so no need to zoom in.  Best C: 401  Best gamma: 641.0  Best score (average): 0.818  C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1951491405381636.png |

### Neural Network

For finding the best parameters for neural network model I run through all the combinations of:

* Layer : we scan the range of [2, 50]
* Learning rate: we scan the range of [1/50, 1.0]. each time we take a step of 1/50.

The best score found was 0.8282

Recommended rate is 0.3 and number of layers 5.



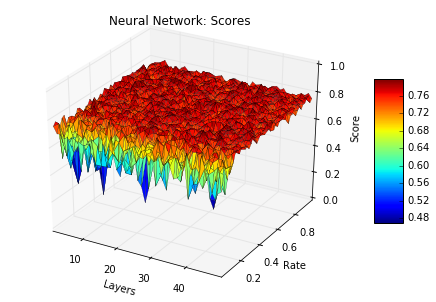


Figure 2: Neural Network scores per <rate, layer> (Breast Cancer dataset)

## Comparison

Below are the durations for searching the best parameters

|  |  |  |
| --- | --- | --- |
| IRIS dataset | SVM | 12.39 seconds |
|  | Neural Network | 127.76 seconds |
| Breast Cancer dataset | SVM | 47.82 seconds |
|  | Neural Network | 95.79 seconds |

The learning time of neural network is taking more time. The process on the dataset of IRIS is much slower to team neural network model.

Looking at the results and accuracy:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Fit time (seconds) | Test time (seconds) | Confusion matrix | Score (Avg.) | (Std.) |
| IRIS Dataset | SVM | 0.01 | 0.0 | [[ 9.8, 0.6]  [ 0.3, 9.3]] | 0.955000 | 0.035000 |
| NN | 0.48 | 0.0 | [[ 6.8, 3.5]  [ 0.1, 9.6]] | 0.820000 | 0.226053 |
| Breast Cancer Dataset | SVM | 0.02 | 0.02 | [[ 30.4, 0. ]  [ 8.6, 0. ]] | 0.779487 | 0.052798 |
| NN | 0.03 | 0.0 | [[ 29.1, 0. ]  [ 9.9, 0. ]] | 0.746154 | 0.064307 |

|  |  |
| --- | --- |
| C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1951646024204292.png  Figure 3: IRIS Dataset | C:\Users\samera\Google Drive\iSchool\תקופה 6\MachineLearning\good_results5_files\qt_img1951753398386692.png  Figure 4: Breast Cancer |