Statistical Methods for Machine Learning
Assignment 3

Tudor Dragan (xlq880)
Nicolae Mariuta (rqt629)
Gabriel Carp (slp670)

March 17, 2015

III.1 Neural Networks

III.1.1 Neural network implementation

We have constructed a neural network with a hidden layer which contains three neurons and for the forward propagation we have implemented the following functions:

(i) for applying the weights between the input neurons and the neurons in the hidden layer

$$Z^{(2)} = XW^{(1)}$$

(ii) for applying the activation function in the hidden layer

$$a^{(2)} = f(Z^{(2)})$$

(ii) for applying the weights between the neurons in the hidden layer and the output neuron

$$Z^{(3)} = a^{(2)}W^{(2)}$$

Derivative of $h(a)$

$$h(a) = \frac{a}{1+|a|}$$

The derivate of the $h(a)$ function has been calculated by applying the following formula:

$$\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}$$

and the results for both $a >= 0$ and $a < 0$ are:

$$h'(a) = \frac{1}{(1+a)^2}, a >= 0$$

$$h'(a) = \frac{1}{(1-a)^2}, a < 0$$

from which we can conclude that

$$h'(a) = \frac{1}{(1+\mathsf{a})^2}$$

Unfortunately we had problems when verifying the gradient computation using numerically estimated gradients.

III.1.2 Neural Networks Training -TODO

III.2 Support Vector Machines

III.2.1 Data Normalization

The means for the train data set is represented by the $Means$ variable and the standard deviation is displayed in $Stds$. For the normalized test data the means can be seen in $testMeansNorm$ and the standard deviations in $testStdsNorm$.

III.2.2 Model selection using grid-search

We used the LIBSVM library for Support Vector Machines. Training the SVM will result in a model which is used for classifying the test data.

```
model = svmtrain(trainX, trainY, 'kernel_function','rbf',
                 'rbf_sigma',sigmas,'boxconstraint',C,'autoscale',false);
```

For the cross validation we used the same algorithm from the first assignment and we split the data into 5 parts. Each part was used to create a model and used the remaining data as a test set.
For each group we calculated the average error and we applied this algorithm for all the possible combination between the regularization parameters C and the kernel parameter γ.
For the non normalized data we obtained the best results for

$$C = 0.01$$

$$\gamma = 10^{-4}$$

For the normalized data we obtained the best results for

$$C = 1000$$

$$\gamma = 10^{-3}$$

Training data accuracy	0.7857
Test data accuracy	0.7526
Normalized training data accuracy	0.8980
Normalized test data accuracy	0.8144

After the normalization of the data set we have higher accuracy because each parameter has almost the same weight in creating the model.

III.2.3.1 Support vectors

We obtained the number of free and bounded support vectors for each value of C. We noticed that we get a large number of bounded support vectors for the lower values of C but also a higher number of free vectors which can lead to classification errors.

III.2.3.2 Scaling behavior - TODO