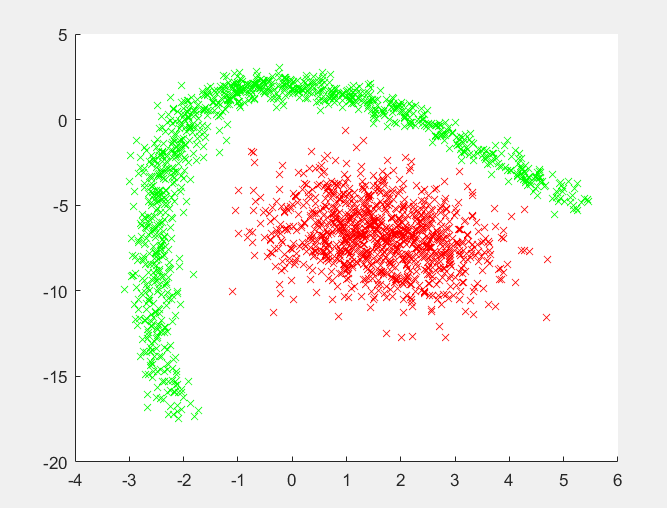
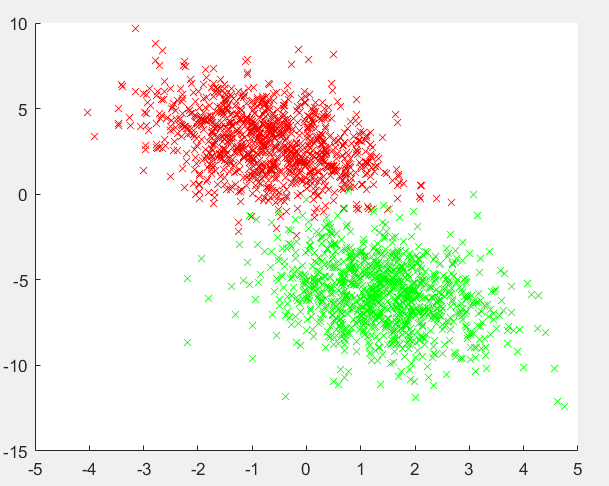
TBMI26 – Computer Assignment Report  
Supervised Learning

Deadline – February 12 2018

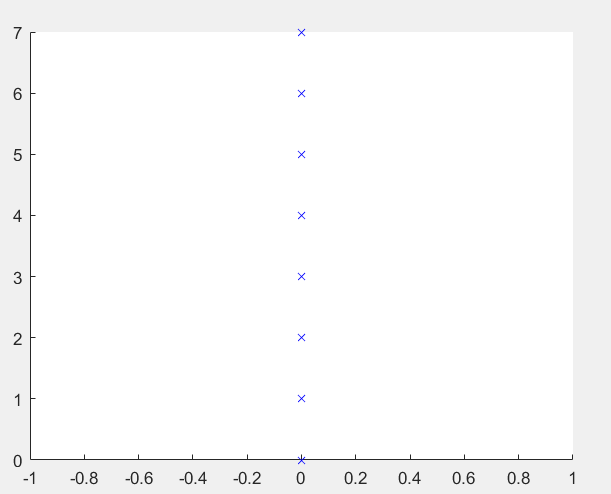
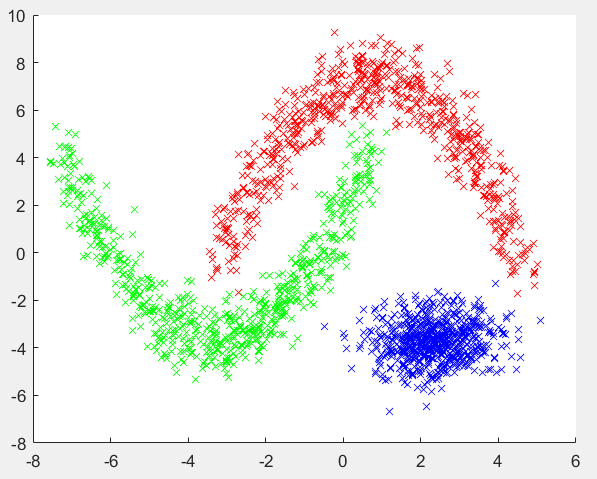
Author/-s:

In order to pass the assignment you will need to answer the following questions and upload the document to LISAM. If you meet the deadline we correct the report within one week after the deadline. Otherwise we give no guarantees when we have time.

1. **Give an overview of the data from a machine learning perspective. Consider if you need linear or non-linear classifiers etc.**



**Data sets 1 and 2**



**Data sets 3 and 4**

The first data set could be classified with a linear classifier since the green and the red classes could be separated with a line. The second data set could not since its impossible to draw a straight line that separates the classes. Same thing with the third data set. The forth data set is different, since it consists of 64 features the 2-dimensional plot doesn’t really show the complexity of the data.

1. **Explain why the down sampling of the OCR data (done as pre-processing) result in a more robust feature representation. See** [**http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits**](http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits)

The pictures are divided into “subpictures” and the pixels in those areas are counted. That means that the feature dimension have been reduced to 8\*8=64 features. This also means that an average from 4\*4 pixels have been taken which means that noise or wrongly labeled pictures have been canceled out. This leads to a more robust data set.

1. **Give a short summery of how you implemented the kNN algorithm.**

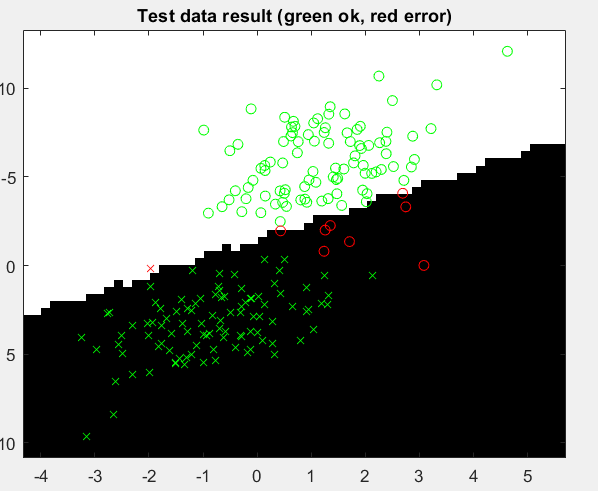
The distances are calculated for every data point in the training and test set. Then the distances are sorted, and the k smallest ones are used to predict the class label.

1. **Explain how you handle draws in kNN, e.g. with two classes (k = 2)?**

In cases where 2 neighbors are used, and they give a different class label, the shortest distance is used.

1. **Explain how you selected the best k for each dataset using cross validation. Include the accuracy and images of your results for each dataset.**

K values between 1 and 10 are evaluated for three different subsets of the data. The average of the accuracy between the three subsets are calculated and used as an overall-sample-accuracy. Then the maximum accuracy is used to decide which mean is the best.



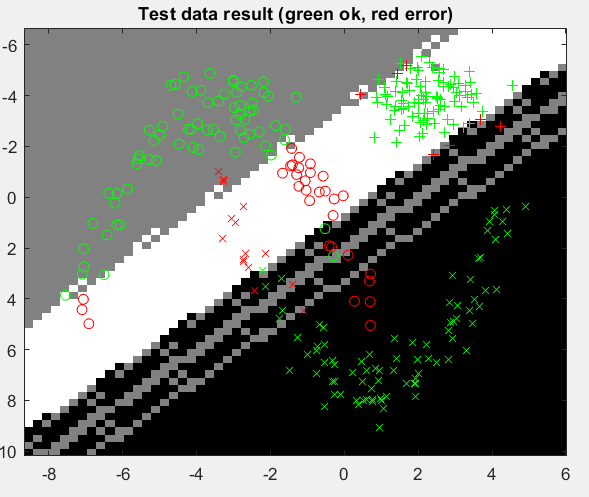
Accuracy = 0.95

This looks quite good. It classifies 95 percent correct. Although, you could probably get an even better classification with a better decision boundary.



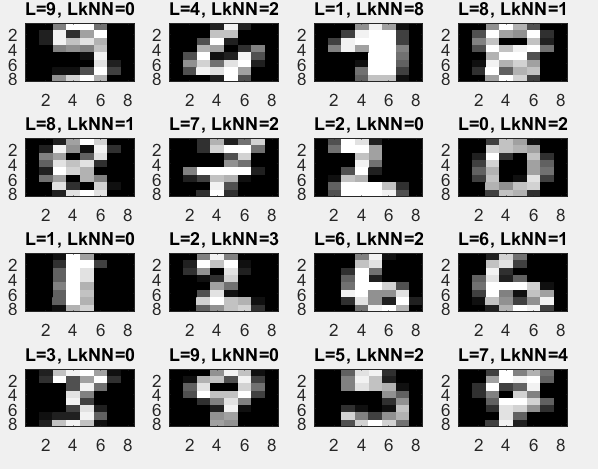
Accuracy = 0.93

Looking at the picture, kNN cannot find the concave shape, but the accuracy is still quite high.



Accuracy = 0.8267

The accuracy is quite high, but the kNN cannot find the convex shape here either.



Accuarcy = 0.11

The hand-written digits can’t really be classified using kNN. The accuracy is really low and in the picture, you can see that the LkNN doesn’t not give correct guesses.

1. **Present the results from the backprop training and how you reached the accuracy criteria for each dataset. Motivate your choice of network for each dataset. Explain how you selected good values for the learning rate, iterations and number of hidden neurons. Include images of your best result for each dataset, including parameters etc.**
2. **Present the results, including images, of your example of a non-generalizable backprop solution. Explain why this example is non-generalizable.**
3. **Give a final discussion and conclusion where you explain the differences between the performances of the different classifiers. Pros and cons etc.**

The pros of kNN is that it is very simple to implement. It could possibly have given a better result with a different distance metric.

1. **Do you think there is something that can improve the results? Pre-processing, algorithm-wise etc.**

Maybe an additional layer in the multilayer neural net could find the more complex boundaries.