Machine Learning: Coursework 1

MLP or RBF?

In my work I decided to use an RBF, this was because MLP's can be harder to optimise than RBF's, and seeing as the quality of output is important for this task, I made the decision to use an RBF. It may be that an MLP could in fact have performed this task better if properly optimised, which is obviously a flaw that I will talk about in my conclusions.

Which kind of RBF?

For my investigation into which kind of RBF would be most effective I decided to model 4 different kinds; exact identity, exact Gaussian, identity and Gaussian. I implemented each of these in MATLAB (you can see my implementation in the tests.m file), fed in 90% of my training data to each of them and then used their results and the testing results to calculate the Mean Squared Loss for each different RBF. I then rotated the data that I was feeding into the RBF's in a K fold system, this allowed me to collect five different mean squared losses for each of the RBF's which I could then average. This resulted in the following table:

	Exact Identity	**Exact Gaussian**	**Identity**	**Gaussian**
Test 1	0.1229	0.29	0.4681	0.4681
Test 2	0.0852	0.234	0.4189	0.4189
Test 3	0.0977	0.2357	0.4658	0.4658
Test 4	0.1119	0.2757	0.4713	0.4713
Test 5	0.0978	0.2716	0.4653	0.4653
Mean	0.1031	0.2614	0.45788	0.45788

As you can see the Exact Identity RBF has significantly lower mean squared loss than all of the other RFB's, this is an indication that we should be focusing our efforts into optimising this RBF going forward. It is worth noting that before these results were collected I spent some time changing the sigma values for the Gaussian functions and the number of centres for the clustering algorithm to optimise the mean squared loss of their respective functions. That means that this table represents the mean square losses of the RBF's in their best-optimised state.

Pre-Processing Whitening or Dimensionality Reduction?

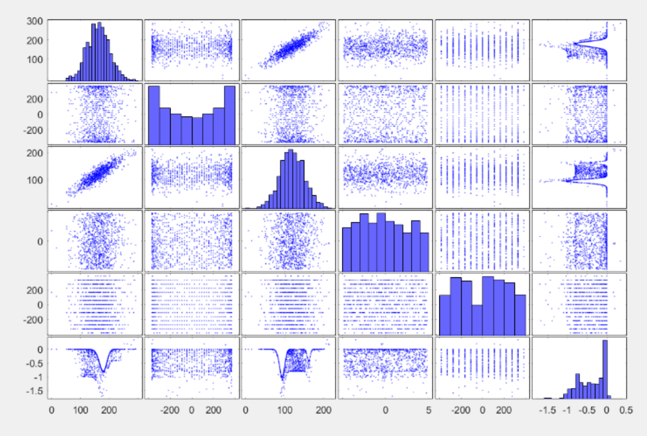
The first kind of pre-processing that I tried was ZCA whitening on my data before being processed by my RBF's. The function I used for my whitening was written by Colorado Reed and can be found here: https://uk.mathworks.com/matlabcentral/fileexchange/34471-data-matrix-whitening/content/whiten.m

The whitening did improve the quality of my results somewhat as can be seen below:

	Exact Identity	Exact Gaussian	Identity	Gaussian
Test 1	0.077	0.29	0.4571	0.5357
Test 2	0.0573	0.234	0.3819	0.5077
Test 3	0.0479	0.2357	0.4485	0.4613
Test 4	0.0618	0.2757	0.4772	0.5463
Test 5	0.0613	0.2716	0.4705	0.5224
Mean	0.06106	0.2614	0.44704	0.51468

I included all of my RBF's in this test because it was easy to do and it could potentially be that others perform better after pre-processing. Obviously the exact identity RBF is still on top in terms of loss.

The Next thing that I tried was dimensionality reduction, I achieved this by plotting each value of my input data to see which contained trends and which just contained noise:



As you can see from this graph, columns 1 and 3, when plotted against my data produce visible trends whilst the others just produce noise. That means that we can remove columns 2, 4 and 5 of the data and give the resulting 2 dimensional dataset to my RBF's for processing.

Doing this resulted in a collection of mean squared losses that looked like this:

	Exact Identity	Exact Gaussian	Identity	Gaussian
Test 1	5.1151e-4	0.2716	0.3598	0.3598
Test 2	4.0538e-4	0.2757	0.4047	0.4048
Test 3	4.0205e-4	0.2357	0.3771	0.3920
Test 4	3.9882e-4	0.2340	0.3850	0.3972
Test 5	8.6146e-4	0.29	0.3903	0.3903
Mean	5.16e-4	0.2614	0.3834	0.3888

This has caused the mean squared loss of my Exact Identity RBF to become very low, and this was the set of conditions I used to create my finished model, trained on the full data set with an exact identity RBF and dimensionality reduction.

Conclusions

I believe that my completed model is very accurate and that it will perform well on the data when given to it. It could be that an MLP would have given a better solution to the problem. However, with an average mean squared loss of 5.16e-4, I am very pleased with my best model.