An RBF was employed to do regression for the Boston House-Prices dataset. The training data was split into a train and a test set with ratio 80:20. Model selection was done by doing 5-fold cross-validation on the training set. Data was shuffled before being partitioned. The choice of k was made considering the small number of training samples. In order to account for variance in the results due to the particular split of the data, the results for a parameter set were averaged over 5 cross-validation runs.

# Pre-processing

The types of pre-processing used for the regression were normalisation, whitening, and normalisation followed by PCA. Dimensionality reduction was done by sequentially dropping a principal component. Feature selection was done by dropping variable #4 Charles River dummy variable, due to prior knowledge – the variable would not affect the house price. A covariance matrix between the features and the output was done, in order to establish which features are correlated the most with the output. Feature selection was performed by keeping the three features which had the highest correlation values.

# Stage 1 of training: Tuning the free parameters of the RBF

Finding the centers of the network was done via k-means. K was chosen in the range [5;300] with a step in {5,10,30,50} which was incrementally increasing with the number of centers. The upper boundary of the range was chosen to be close to the number of training samples resulting from the cross-validation split.

The activation function was optimised between the identity function, a Gaussian with a width parameter, or a multivariate Gaussian. The Gaussian function’s width was set to be a single value or a different value for each cluster. The single width was calculated as twice the mean distance between the centers. The width per cluster was calculated as the average distance between the data points and the center; twice this distance; or the standard deviation of the data points in the cluster. The multivariate Gaussian used the covariance matrix of the features of the data points per cluster.

# Stage 2 of the training: Finding the weights of the RBF

The weights of the RBF were found using regularisation. The parameter lambda was optimised for the values in {0,0.3,0.6,0.9}, 0 meaning no regularisation applied; and for [0.1;0.4] with a step of 0.1.

# Results from training

Using the multivariate Gaussian with number of clusters above 12 resulted in numerical instabilities due to some clusters having size of 1, matrices which could not be inverted, or terms which were too big for the exponential function. It was tested with centers in the range [3;11], but overall resulted in performance, worse than the linear or the Gaussian function.

Normalisation, as well as normalisation with PCA resulted in better performance overall than whitening. Dropping variable #3 improved the results. Keeping the variable with the biggest correlations did not lead to improvement, regardless of the activation function used or the number of centers. Sequentially dropping principal components resulted in a slight improvement of the results when principal component #7 was dropped.

The linear activation function led to better results, the higher the number of centers was chosen. It also performed better than the Gaussian function on the higher number of centers. If the number of centers was below 100, the Gaussian function with width, set to the average distance between data points to the center per cluster, performed better than the other two width methods, and the linear function. However, the linear function led to the smallest MSE overall. The Gaussian function with width twice the average distance led to results with MSE twice as big as the width set to the average distance for smaller numbers of centers, but resulted in close MSE, when the number of centers was higher. Using a single width parameter for all clusters for the Gaussian resulted in sub-optimal performance overall.

Regularisation improved the results of the linear function, but did not improve the results of the Gaussian. The finer grid-search for the regularisation parameter lambda was done for a model with 300 number of centers and a linear activation function, finding the parameter 0.1 to be the best one.

# Final Model & Outputs

The final model was chosen with number of centers set to 300, a linear activation function, lambda set to 0.1. The pre-processing done was dropping feature #3; and normalisation with PCA, dropping the principal components #6 and #7. The predictions in the order of the given data points are:

14.1492

25.0941

20.2592

20.3016

25.5496

27.4354

22.9767