COMPM055: SUPERVISED LEARNING

Coursework 2

Klaudia Ludwisiak
ucabkll
16117698

Minas Sifakis
ucabmkm
16104127

28 December 2016

Exercise 1- Linear Example

a). Generated the dataset according to the instructions and performed an 80/20 (train/test) data split. Standardised the data (based on the training set) and saved the mean and standard deviation vectors.

Learned the weight vector by performing Ridge Regression (RR) in the primal form (on the training data set), with the regularisation parameter set to the true noise variance, $\sigma^2 = 0.01$. Formed predictions, $\hat{y}_i$, for the test data set, based on the vector of learned weights ($\hat{w}$) and computed the test set mean squared error (MSE), $\mathbf{e_a = 0.0054}$, and the Euclidean distance (norm) to the true labels: $\mathbf{d_a = 0.3286}$. Using the saved normalization parameters, we rescaled the weights back to the original data scale, computed the bias term and the raw difference between the calculated and the true weight vector ($\hat{w}$-w) - presented in table 1 that follows together with percentage differences (over the true weights).

Table 1: True v.s. predicted weights (bias value for RR model $w_0 = -0.7612$)

w	-0,1311	0,1609	-0,4476	0,9657	1,2269	0,2505	-0,4442	-0,1869	-0,5693	-0,5355
$\hat{w}$	-0.1345	0.1444	-0.4662	0.9578	1.2201	0.2538	-0.4446	-0.1897	-0.5531	-0.5375
$\Delta w = w - \hat{w}$	0.0033	0.0165	0.0186	0.0079	0.0069	-0.0033	0.0004	0.0028	-0.0162	0.0020
$100 * \Delta w / w$	-2.5505	10.2298	-4.1622	0.8162	0.5605	-1.3314	-0.0842	-1.4817	2.8481	-0.3799

b) Learned the alpha (α) weight vector by performing Ridge Regression in the dual form (on the training data set) with the regularisation parameter set to the true noise variance, $\sigma^2 = 0.01$. Multiplied each training data point with the calculated α_i value and obtained the vector of dual weights, $\hat{w}_{ai}$. Formed predictions, $\hat{y}_i$, for each point in the test data set, by taking the sum of the inner products of each test point x_{ti} with all the (80) dual weights, $\hat{w}_{ai}$ (using the linear kernel ($K(x, z) = \langle x, z \rangle = x_j z_j$) . Calculated mean squared test set error, $\mathbf{e_b}$, and the Euclidean distance, $\mathbf{d_b}$, to the true labels: $\mathbf{e_b = 0.0054}$, $\mathbf{d_b = 0.3286.}$

As expected, the calculated errors values and distances are exactly the same in both cases. The primal and dual formulations are equivalent and since we applied them to the exact same dataset, we expected them to yield the same results. Differences due to numerical issues (each method involves inversion of a different matrix and the dual method more multiplications for prediction) were not evident up to 15 decimal points. The achieved quality of fit is very satisfactory as can be manifested by the very small values of both **e** and **d** and the small amplitude and random distribution of the residual plots (see figure 1 that follows).

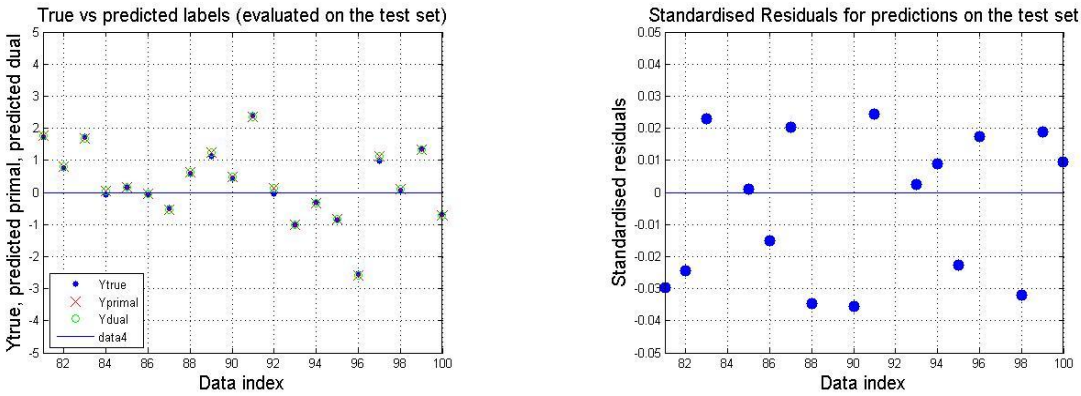

Figure 1: (a). % Difference between the true and the predicted weights learned with RR in the primal form,
(b). Standardised residuals plot

Exercise 2- Non-linear Kernel

a). We generated a dataset consisting of 5000 two-dimensional points, from a quadratic kernel without noise. Interpolated data (using cubic interpolation) and plotted for visual inspection (figure 2 sideways).

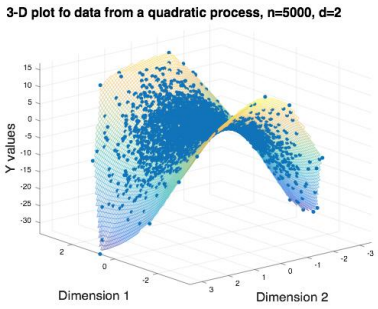

Figure 2: 2-D quadratic surface plot

b). We generated a dataset from a quadratic kernel with parameters: d = 20, n = 500, σ = 5 and split the data with a train/test data split of 25/75. Applied nonlinear kernel Ridge Regression with a quadratic kernel and the regularization parameter set to the true noise variance, $\sigma^2 = 0.01$. Calculated the test set mean square error (MSE): $\mathbf{e_1 = 267.49}$.

The calculated value of the MSE is large. However, the dataset that we are dealing with is high dimensional, the noise variance high ($\sigma^2 = 25$) and the train/test split that we have used far from optimal (i.e. the training data set is not large enough). To investigate this issue further we inverted the data split to 75% training set – 25% test set and repeated the fit. We observed a substantial reduction of the test set mse to $\mathbf{e_b' = 60.58.}$

c). We generated 100 different datasets with the same parameters as part b above and split the data in three parts (train/validation/test) with a split ratio of 25/15/60. For each dataset, we fitted a quadratic kernel RR model on the training sets and evaluated performance on the validation set, in order to choose the best regularization parameter λ from within the specified range ($5x10^{-4}$, $5x10^{-3.75}$…$5x10^{-3.75}$, $5x10^{4}$).

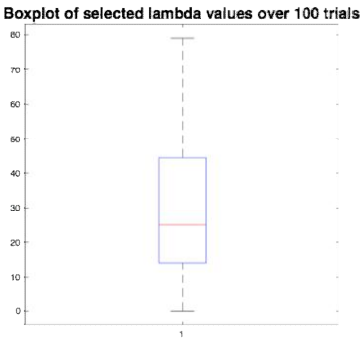

Figure 3: Boxplot of selected λ values

The *mean value* of selected regularization parameter over the 100 trials was: $\mathbf{\lambda_{meanc} = 37.16}$. The *median* of the distribution of selected values, was: $\mathbf{\lambda_{medc} = 25}$ i.e. at the theoretical value ($\mathbf{\lambda_c = \sigma^2 = 25}$) , the distribution is relatively symmetric around this value as can be seen from figure 3 and the outliers are not alarming taking into account the large noise variance and the limited size of the training set. The calculated distribution of the mean squared error over the 100 runs is presented in figure 4. It is a relatively symmetric distribution around the global mean value of $\mathbf{e_{100} = 216.44.}$ The standard deviation is: $\mathbf{std_{100} = 26.35}$. From the plot of predicted (ŷ) v.s. true (y) values (figure 4.b) we observe that despite the large MSE the actual fit is indeed satisfactory.

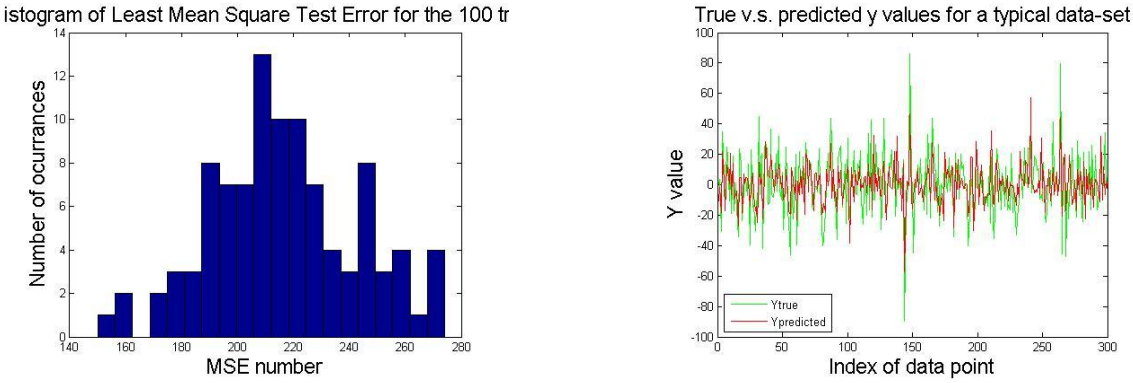

Figure 4: (a). Histogram of the calculated test set MSE (with the best selected λ for each trial) for 100 trials
(b). Plot depicting the quality of fit between the actual and predicted y values (evaluation on the test set)

Exercise 3- Error Bar and Fitting

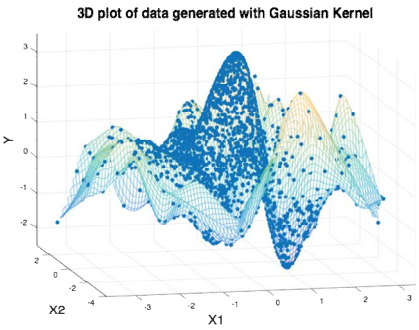

Figure 5: 2-D Gaussian RBF surface plot

a). As per question requirement, we generated a 2-D Gaussian Process dataset with a radial basis function (Gaussian) kernel. We computed 5000 points with a kernel width v = 0.5 (without additive noise) and used cubic interpolation to produce figure 5, which is, we believe, typical for a Gaussian process.

b). We generated 100 datasets using the Gaussian kernel with parameters: d = 10, n = 500, v0 = 5, σ = 1.3 and split each dataset to 3 parts, train/validation/test, with proportions 12.5/12.5/75. With the regularization parameter, λ, set to the true value (σ^2) we implemented non-linear Kernel Ridge Regression with a Gaussian radial basis function (RBF) kernel. For each of the 100 generated datasets we looped over a (logarithmic) range of candidate kernel widths, v, {$5x10^{-2}$, $5x10^{-1.5}$, . . . , $5x10^{1.5}$, $5x10^{2}$} and selected the ”best” value based on minimizing the distance of the empirical distribution of the normalized errors from the CDF of a Gaussian (typical plot displayed in Figure 6) . We re-run the regression with the best v selected and stored the test set MSEs.

The average “best-v” MSE (over the 100 trials), was $\mathbf{e_{100} = 2.12}$ (which is satisfactory taking into account that the datasets have a noise variance of σ^2 = 1.69) with a standard deviation of: $\mathbf{std_{100} = 0.41}$. The shape of the corresponding normal plot of the MSE resembles a normal distribution, with a mild right skew and fatter upper tail. The mean value was, $\mathbf{v_{mean100} = 13.11}$, since it was influenced by the few large values of selected best-v that occurred (the mean is not robust to outliers). The median value of the ‘best’ selected v was: $\mathbf{v_{med100} = 1.58}$. Further investigations: We doubled the size of the training and validation sets and observed a reduction of the standard deviation of the MSE and a smaller valued of selected mean λ. We also investigated the effect of reduced noise variance (i.e. improved the signal to noise ratio). A reduction of the standard deviation of the noise by 50% resulted in a large reduction of the amplitude of the MSE, (to e'_{100} = 0.70, std'_{100} = 0.27) and to an increase of the value of the selected width (v'_{med100} = 5, $v'_{mean100}$ = 14.00). We also visually investigated the quality of the fit by plotting the y true v.s. the predicted ŷ for various different levels of variance σ. Finally, we implemented selection based on validation set MSE. This method yielded superior results, both in terms of average MSE (~10% less) and in terms of the selection of the kernel width.

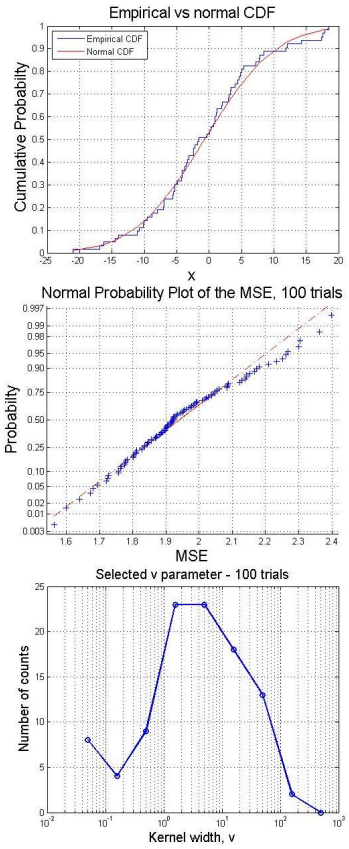

Figure 6: (i). ECDF v.s. Normal CDF plot, (ii). Normal Probability plot of the MSE, (iii). Selected kernel width, v

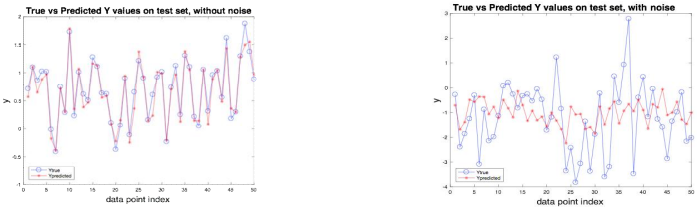

Figure 7: True v.s. predicted Y for different levels of noise variance : (ı). $\sigma=2.5*10^{-5}$, (ii). σ= 2.5

Exercise 4.Evidence based model selection

Generated 100 datasets using a Gaussian RBF kernel and parameters: d = 10, v = 5, σ = 0.5. Performed train/validation/test split of 12.5/12.5/75.

a). We used the log-likelihood of the data to select the optimal value of the regularisation parameter (λ) among the set of candidate values provided { $\sigma^2 \cdot 10^{-2}$, $\sigma^2 \cdot 10^{-1.9}$, . . . , $\sigma^2 \cdot 10^{1.9}$, $\sigma^2 \cdot 10^{2}$}). Since the evidence is computed on the training data the validation set was not used. In figure 7, we have plotted typical curves of the calculated log-likelihood against λ. One can see that in most datasets there was a clear maximum of the likelihood function within the provided range of λ. However, there were datasets that a likelihood plateau was observed, followed by reducing values at higher λ (our algorithm selected the lowest λ in these cases).

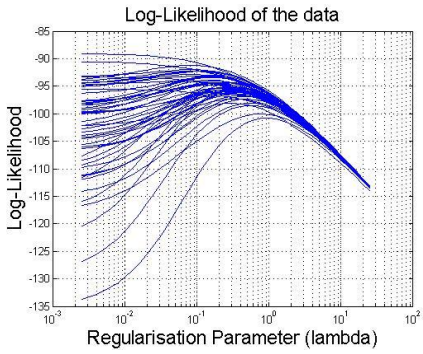

Figure 8: Log-Likelihood against λ

The calculated average test MSE over the 100 datasets was: $\mathbf{e_{100a} = 0.4447, (std_{100a} = 0.1199)}$, the corresponding mean selected-λ was: $\boldsymbol{\lambda_{mean100a} = 0.2481,}$ and the median: $\boldsymbol{\lambda_{median100a} = 0.25}$. The value of the MSE is considered satisfactory (taking into account that the data have an inherent error variance of $\sigma^2 = 0.25$), while the mean selected-λ is very close to the theoretical value (i.e. $\sigma^2 = 0.25$) and, once more, the median value of selected-λ is at the exact theoretical value.

b). In this part of the exercise we performed λ-selection, based on minimizing the MSE on the validation set. We then run KRR on the test set and stored the test set MSE.

The average values obtained with this approach are the following: $\mathbf{e_{100b} = 0.4745, (std_{100b} = 0.1252)}$, $\boldsymbol{\lambda_{mean100b} = 2.0376,}$ $\boldsymbol{\lambda_{median100b} = 0.25}$. The calculated average MSE is effectively similar with the one calculated from the models with evidence based selection. The mean selected regularization parameter was almost an order of magnitude larger than the theoretical value (inspection of the data revealed that the mean value was influenced by a few large values of λ that occurred). Once more, the median was exactly at the theoretical value.

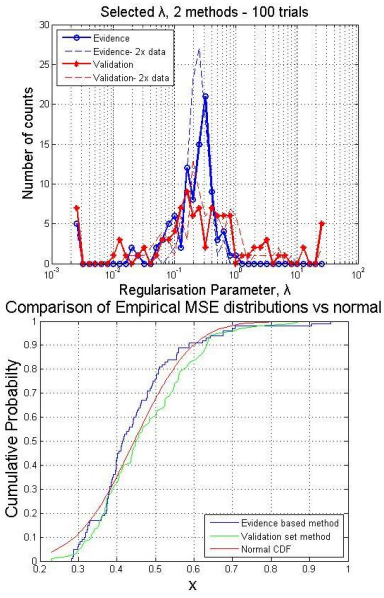

Figure 9: (i). Selected λ – 100 trials, (ii). ECDFs for MSE v.s. normal

We investigated the effect of doubling the size of the training and validation sets (each to 25% of the total dataset). Both the amplitude and the standard deviation of the MSE were reduced ($e'_{100a} = 0.3622$, $std'_{100a} = 0.0928$, $e'_{100b} = 0.4003$, $std'_{100b} = 0.1424$). The selected mean-λ further approached the theoretical value and the distribution around the theoretical value became tighter ($\lambda'_{mean100a} = 0.2526$, $\lambda'_{mean100b} = 1.1188$) – see also figure 8. We repeated the whole investigation (parts a, b) with many different feeds of the random number generator. The evidence approach consistently chose values of λ that were smaller and closer to the true value and model fits that were slightly better (in terms of test set MSE).