Supervized Learning: Assignment 1

Thomas Nedelec
MSc Machine Learning
thomas.nedelec.15@ucl.ac.uk

Michal Daniluk
MSc Machine Learning
michal.daniluk.15@ucl.ac.uk

1 Exercise 1: Least Square Regression: effect of the training set size

We generated a noisy random data set, containing 600 samples, as $y_i = x_i'w + n_i$, where each x_i and n_i are drawn from the standard normal distribution. We splitted the data into a training set of size 100 a test set of size 500. We compute the mean squared error on both the training and test sets. Figure 1 shows examples of different training sets and estimated regression w.

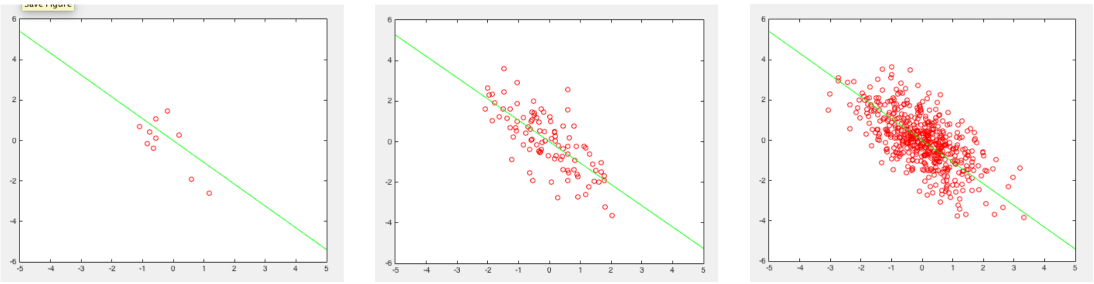
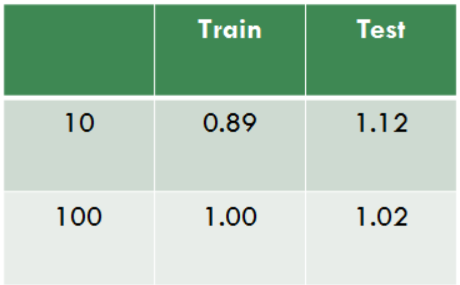


Figure 1: (a) training set of 10 points (b) training set of 100 points (c) the entire data set

Figure 2 illustrates the ”2 x 2” table of averages mean square error for both training and test sets, on both 10 and 100 element-size training set.



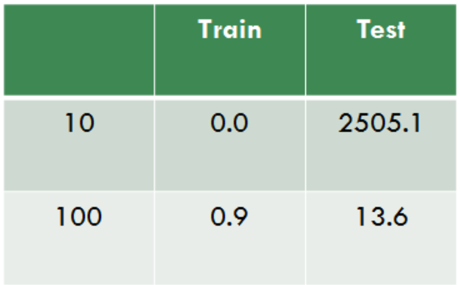
	Train	Test
10	0.89	1.12
100	1.00	1.02

Figure 2: Average mean square error for both training and test sets, on both 10 and 100 element-size training set.

We can observe that increasing the size of training set lets to decrease the average mean error on the test set. However, it increased the average mean error on train set. With increasing the size of training set, the average error on the test and train set became similar. The average error on test set is larger than on train set. It is clear, beacause we have seen the train data before and haven't seen the test data.

2 Exercise 2: Least Square Regression: effect of dimensionality

We repeated the task in exercise 1, but with 10-dimensional data sets. Figure 3 illustrates the "2 x 2" table of averages mean square error for both training and test sets, on both 10 and 100 element-size training set.



	Train	Test
10	0.0	2505.1
100	0.9	13.6

Figure 3: Average mean square error for 10-dimensional data for both training and test sets, on both 10 and 100 element-size training set.

For 10-sample training set, we obain the average mean square error 0 for train set and 2505.1 for test set. We have 10-dimensional data and 10 samples, so we can accurately predict y_i without any errors. However, the prediction is overfitted to the training data and we have a large error for test set. Icreasing number of samples to 100 helps to decrease average error on test set. Average error for train set is bigger than for 10 samples, but our goal is to decrase test error.

3 Ridge regression

The regularization is a way to reduce the freedom of the classifier in order to improve the generalization and reach a better test set average error.
To implement ridge regression, we would like to minimize according to w the cost function,

$$\gamma w^T w + \frac{1}{l} \sum_{i=1}^{l} (x_i^T w - y_i)^2. \tag{1}$$

Using the notation $X = (x_1, x_2, ..., x_l)^T$, a matrix containing the training sample vectors as its rows, we can rewrite the cost function as:

$$\gamma w^T w + \frac{1}{l} Tr((Xw - Y)^T (Xw - Y)) \tag{2}$$

Taking derivative according to w, we reach:

$$2\gamma w + 2\frac{1}{l}(X^T X w - 2Y^T X) = 0 \tag{3}$$

To reach the conditions for optimality, we set the previous equation to zero and we reach:

$$w = (\gamma l Id + X^T X)^{-1} Y^T X \tag{4}$$

because if $\gamma * l \neq 0$ $(\gamma l Id + X^T X)$ is non-singular.

$\gamma l Id + X^T X$ is symetric.

We consider $u \in \mathbb{R}^d$:

$$\begin{aligned} \langle(\gamma l Id + X^T X)u, u\rangle &= \langle \gamma l u, u\rangle + \langle X^T X, u\rangle \\ &= \gamma l \langle u, u\rangle + \langle Xu, Xu\rangle \\ &= \gamma l ||u||^2 + ||Xu||^2 > 0 \; \textit{for } u \neq \textit{0} \end{aligned}$$

Thus $\gamma l Id + X^T X$ is definite positve.

4 Effect of the regularisation parameter

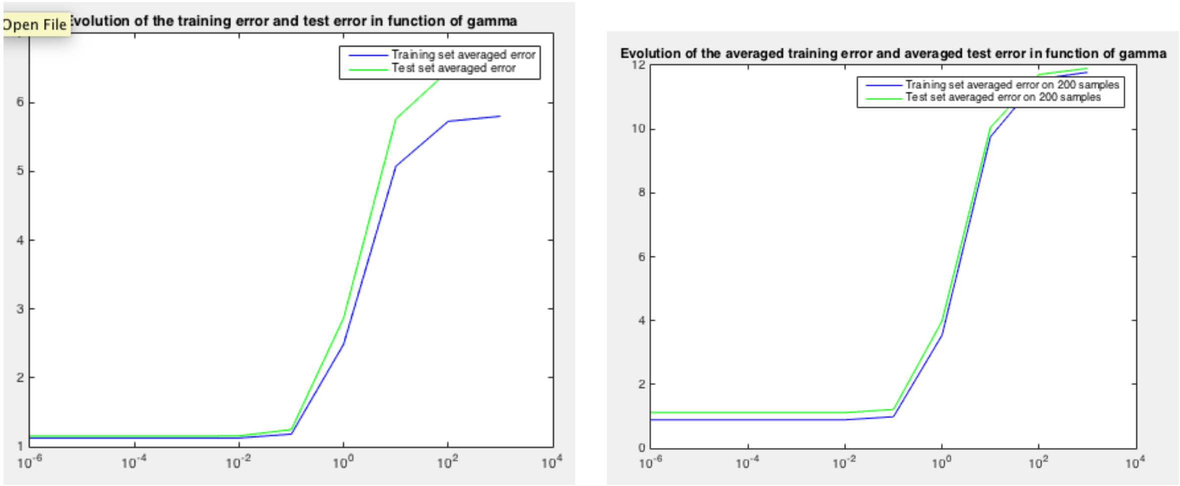


Figure 4: (a) evolution of the training error and test errror for one run (b) evolution of the training error and test error averaged on 200 runs

First, the graph shows that if gamma is too high, the training error and test error increase dramatically. It is intuitive because when gamma is high the algorithm is far more likely to minimize $||w||$ rather than the training error.

We have an optimal point corresponding to inflection point of the curve in order to select γ. Nevertheless, the training set error is not a sufficiently good guidance to select the regularization parameter because the charts are quite different.

However, when we average the training error on 200 runs, we can observe that now the test error and the training error are now almost identical.

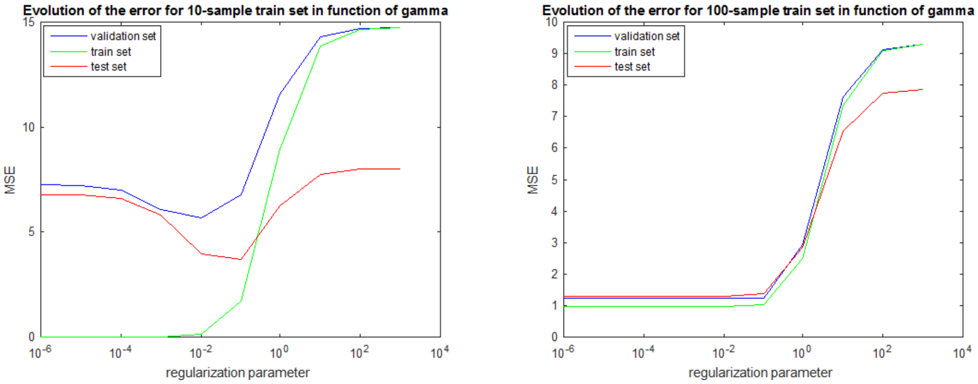
A way to select the optimal γ would be to plot the error on the training set averaged on a certain number of runs of the algorithm and select the optimal γ on this curve.

5 Tuning the regularization parameter using a validation set

6 Tuning the regularization parameter using cross valdidation

We use 5-fold cross validation to tune the regularization parameter. Figure 5 shows cross-validation score on top of the training and test set error for different values $\gamma = \{10^{-6}, 10^{-5}, \ldots, 10^{3}\}$ of the regularization parameter.

TO DO : Interpretation



(a) 10-sample training set

(b) 100-sample training set

Figure 5: Evaluation of the mean square error in function of gamma for different number of training examples.