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

Abstract

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

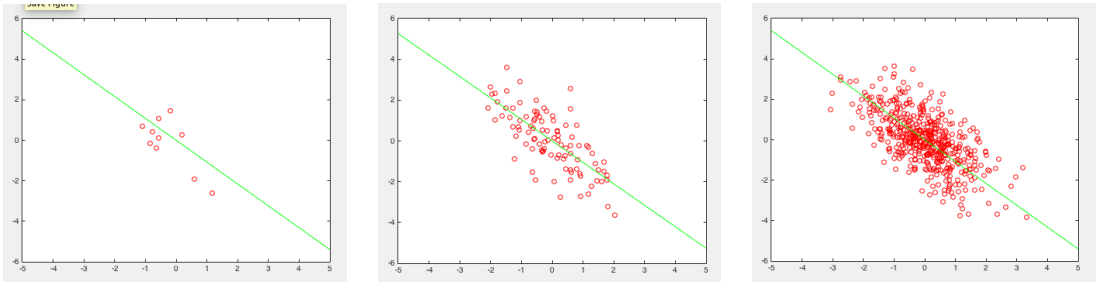

Figure 1: (a) training set of 10 points(b) training set of 100 points (c) the entire data set

```
resultsMatrix =

    0.8927    1.1202
    1.0048    1.0152
```

Figure 2: Average error on the training set (first row: 10 training points, second row: 100 training points

We can observe that increasing the size of training set lets to decrease the average mean error on the test set.

2 Exercise 2: Least Square Regression: effect of dimensionality

```
resultsMatrix =

   1.0e+03 *

    0.0000    2.5051
    0.0009    0.0136
```

Figure 3: Average error on the training set (first row: 10 training points, second row: 100 training points

Interpretation:

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

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

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

4 Effect of the regularisation parameter

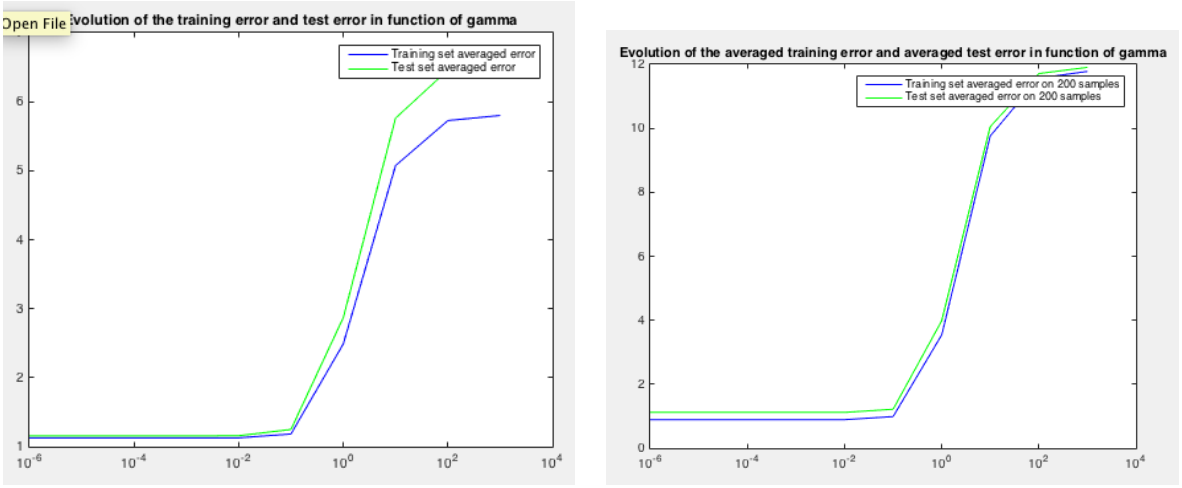

Figure 4: (a) evolution of the training error and test errror for one run (b) evolution of the training error and test error averaged on 200 runs

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