

# Statistical Machine Learning GU4241/GR5241

Spring 2018

<https://courseworks.columbia.edu/>

## Homework 3

Due: Thursday, Mar. 22th, 2018

**Homework submission:** Please submit your homework electronically through Canvas by 11:59pm on the due date. You need to submit both the pdf file and your code (either in R or Python).

### Problem 1 (Ridge Regression and Lasso for Correlated Variables, ISL 6.5, 10 points)

It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that  $n = 2$ ,  $p = 2$ ,  $x_{11} = x_{12}$ ,  $x_{21} = x_{22}$ . Furthermore, suppose that  $y_1 + y_2 = 0$  and  $x_{11} + x_{21} = 0$  and  $x_{12} + x_{22} = 0$ , so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero:  $\hat{\beta}_0 = 0$ .

#### Homework Problems.

1. Write out the ridge regression optimization problem in this setting.
2. Argue that in this setting, the ridge coefficient estimates satisfy  $\hat{\beta}_1 = \hat{\beta}_2$ .
3. Write out the lasso optimization problem in this setting.
4. Argue that in this setting, the lasso coefficients  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are not unique—in other words, there are many possible solutions to the optimization problem in 3. Describe these solutions.

### Problem 2 (Smoothing Splines, ISL 7.5, 10 points)

Consider two curves,  $\hat{g}_1$  and  $\hat{g}_2$ , defined by

$$\begin{aligned}\hat{g}_1 &= \arg \min_g \left( \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 dx \right), \\ \hat{g}_2 &= \arg \min_g \left( \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int [g^{(4)}(x)]^2 dx \right),\end{aligned}$$

where  $g^{(m)}$  represents the  $m$ th derivative of  $g$ .

#### Homework Problems.

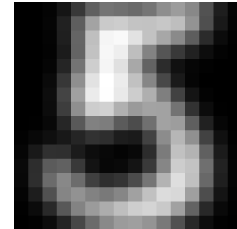
1. As  $\lambda \rightarrow \infty$ , will  $\hat{g}_1$  or  $\hat{g}_2$  have the smaller training RSS, or there is no definite answer? Explain briefly.
2. As  $\lambda \rightarrow \infty$ , will  $\hat{g}_1$  or  $\hat{g}_2$  have the smaller test RSS, or there is no definite answer? Explain briefly.
3. As  $\lambda \rightarrow 0$ , will  $\hat{g}_1$  or  $\hat{g}_2$  have the smaller training and test RSS, or there is no definite answer? Please explain.

### Problem 3 (SVM, 20 points)

In this problem, we will apply a support vector machine to classify hand-written digits. You do not have to implement the SVM algorithm: The R library `e1071` provides an implementation, see

<http://cran.r-project.org/web/packages/e1071/index.html>

Download the digit data set from the course website. The zip archive contains two files: Both files are text files. Each file contains a matrix with one data point (= vector of length 256) per row. The 256-vector in each row represents a  $16 \times 16$  image of a handwritten number. The data contains two classes—the digits 5 and 6—so they can be labeled as -1 and +1, respectively. The image on the right shows the first row, re-arranged as a  $16 \times 16$  matrix and plotted as a gray scale image.



- Randomly select about 20% of the data and set it aside as a test set.
- Train a linear SVM with soft margin. Cross-validate the margin parameter.
- Train an SVM with soft margin and RBF kernel. You will have to cross-validate both the soft-margin parameter and the kernel bandwidth.
- After you have selected parameter values for both algorithms, train each one with the parameter value you have chosen. Then compute the misclassification rate (the proportion of misclassified data points) on the test set.

### Homework questions:

1. Plot the cross-validation estimates of the misclassification rate. Please plot the rate as
  - (a) a function of the margin parameter in the linear case.
  - (b) a function of the margin parameter and the kernel bandwidth in the non-linear case (you are encouraged to use heat map here).
2. Report the test set estimates of the misclassification rates for both cases, with the parameter values you have selected, and compare the two results. Is a linear SVM a good choice for this data, or should we use a non-linear one?