Homework 2

Data Analysis and Machine Learning FYS-STK3155/FYS4155

Department of Physics, University of Oslo, Norway

Sep 5, 2019

Exercise 4

This exercise is a continuation of exercise 2 from homework 1. We will use the same function to generate our data set, still staying with a simple function $y(x)$ which we want to fit using linear regression, but now extending the analysis to include the Ridge and the Lasso regression methods. You can use the code under the Regression as an example on how to use the Ridge and the Lasso methods, see the regression slides).

We will thus again generate our own dataset for a function $y(x)$ where $x \in [0, 1]$ and defined by random numbers computed with the uniform distribution. The function y is a quadratic polynomial in x with added stochastic noise according to the normal distribution $\mathcal{N}(\prime, \infty)$.

The following simple Python instructions define our x and y values (with 100 data points).

```
x = np.random.rand(100,1)
y = 5*x*x+0.1*np.random.randn(100,1)
```

1. Write your own code for the Ridge method (see chapter 3.4 of Hastie *et al.*, equations (3.43) and (3.44)) and compute the parametrization for different values of λ. Compare and analyze your results with those from exercise 2. Study the dependence on λ while also varying the strength of the noise in your expression for $y(x)$.

2. Repeat the above but using the functionality of **Scikit-Learn**. Compare your code with the results from **Scikit-Learn**. Remember to run with the same random numbers for generating x and y.

3. Our next step is to study the variance of the parameters β_1 and β_2 (assuming that we are parameterizing our function with a second-order polynomial). We will use standard linear regression and the Ridge regression. You can now opt for either writing your own function or using **Scikit-Learn** to find the parameters β. From your results calculate the variance of these

paramaters (recall that this is equal to the diagonal elements of the matrix $(\hat{X}^T\hat{X}) + \lambda\hat{I})^{-1})$. Discuss the results of these variances as functions of λ. In particular, try to link your discussion with the discussion in Hastie *et al.* and their figure 3.11.

4. Repeat the previous step but add now the Lasso method, see equation (3.53) of Hastie *et al.*. Discuss your results and compare with standard regression and the Ridge regression results. You can write your own code or use the functionality of **scikit-learn**. We recommend the latter since we have not yet discussed how to solve the Lasso equations numerically.

5. Finally, using **Scikit-Learn** or your own code, compute also the mean square error, a risk metric corresponding to the expected value of the squared (quadratic) error defined as

$$MSE(\hat{y}, \hat{\tilde{y}}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2,$$

and the R^2 score function. If $\tilde{\hat{y}}_i$ is the predicted value of the $i-th$ sample and y_i is the corresponding true value, then the score R^2 is defined as

$$R^2(\hat{y}, \tilde{\hat{y}}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2},$$

where we have defined the mean value of $\hat{y}$ as

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i.$$

Discuss these quantities as functions of the variable λ in the Ridge and Lasso regression methods.

Exercise 5

The theory behind this exercise will be covered during the lectures of week 36. It requires reading chapter three of Hastie et al, in particular the derivations preceeding equation (3.49) as the well as the material in the Regression slides that deal with the singular value decomposition.

Using the singular value decomposition, show that the variance of the direction vector $\hat{z}_i = \hat{X}\hat{v}_i = \hat{u}_1 d_1$ is equal to (equation (3.49) of Hastie *et al.*)

$$\mathrm{Var}(\hat{z}_i) = \frac{d_i^2}{N},$$

where d_i are the singular values of the matrix $\hat{X}$. In Hastie *et al*, the matrix elements of X are centered. The consequence is that the mean values of for example $\hat{u}_i$ are zero.

Give an interpretation of these results, in particular in connection with the variance of the coefficients you obtained in the previous exercise.

Solution to the last exercise

A possible way to show why $\langle \hat{u}_i \rangle = 0$ given that the columns of $\hat{X}$ is centered is by considering $\left\langle \hat{X}\hat{v}_i \right\rangle$:

$$\begin{aligned}
\left\langle \hat{X}\hat{v}_i \right\rangle &= \frac{1}{N}\sum_j (\hat{X}\hat{v}_i)_j \\
&= \frac{1}{N}\sum_j \sum_k x_{jk}\hat{v}_i(k) \\
&= \frac{1}{N}\sum_k \hat{v}_i(k) \sum_j x_{jk} \\
&= \sum_k \hat{v}_i(k) \left(\frac{1}{N}\sum_j x_{jk} \right) \\
&= \sum_k \hat{v}_i(k) \left\langle \hat{x}_k \right\rangle
\end{aligned}$$

where x_{jk} being the element of $\hat{X}$ at row j and column k, $(\hat{X}\hat{v}_i)_j$ the j-th element of the vector $\hat{X}\hat{v}_i$, $\hat{x}_k$ being the k-th column vector of $\hat{X}$, and $\hat{v}_i(k)$ the k-th element of the vector $\hat{v}_i$.

Since the columns of $\hat{X}$ are assumed to be centered, $\langle \hat{x}_k \rangle = 0$ for all k. This gives that $\left\langle \hat{X}\hat{v}_i \right\rangle = 0$.

But $\left\langle \hat{X}\hat{v}_i \right\rangle = \langle \hat{u}_i d_i \rangle = d_i \langle \hat{u}_i \rangle$.

Since $\left\langle \hat{X}\hat{v}_i \right\rangle = 0$, then $d_i \langle \hat{u}_i \rangle = 0$ also. Assuming that $d_i \neq 0$ (otherwise the variance in the exercise would just be zero), gives that $\langle \hat{u}_i \rangle = 0$.

Regarding $\hat{V}$ and using the similar approach as above by computing $\left\langle \hat{X}^T\hat{u}_i \right\rangle = d_i \langle \hat{v}_i \rangle$, we have

$$\begin{align}
\left\langle \hat{X}^T \hat{u}_i \right\rangle &= \frac{1}{N} \sum_j (\hat{X}^T \hat{u}_i)_j \tag{1}\\
&= \frac{1}{N} \sum_j \sum_k x_{kj} \hat{u}_i(k) \tag{2}\\
&= \frac{1}{N} \sum_k \hat{u}_i(k) \sum_j x_{kj} \tag{3}\\
&= \frac{1}{N} \sum_k \hat{u}_i(k) \sum_j x_{kj} \tag{4}\\
&= \sum_k \hat{u}_i(k) \left(\frac{1}{N} \sum_j x_{kj} \right) \tag{5}\\
& \tag{6}
\end{align}$$