Instituto Superior Técnico

Machine Learning
Lab 1 - Linear Regression

André Pereira 90016

Anton Erikson 98037

Grupo 16

2 Pm, Friday

Teacher: Catarina Barata

Least Squares Fitting

2.1

2

The matrix expression of the LS estimate of the coefficients is:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y.$$

Where:

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & \dots & x_p^{(n)} \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

The corresponding sum of squares error:

$$\text{SSE}(\beta) = \|y - \hat{y}\|^2 = \|y - X\beta\|^2.$$

3

(a)

A straight line is equivalent to a first degree polynomial. The line seems to be well fitted to the data points.

4

(a)

As expected the raw data looks a bit like a cosine function. However with some noise. The LS-estimation seems to fit well, since the obtained result of the regression is a cosine function that hovers around the data.

(b)

The SSE was calculated to be about 1.342. Looking at the plot no points seems to be more than about 0.25 away in the y-direction from the curve. Therefore an SSE of 1.342 seems resonable considering there is a total of 50 points.

The coefficients of the fitted polynomial were 0.976, -0.026 and -1.532. These are also reasonable. The first coefficient can be read from the graph by looking at what y the curve crosses $x = 0$. The second coefficent is close to zero, which also the peak of the curve is. And lastly the third coefficient is negative, which is expected since the curve has a maximum.

5

(a)

The curve in the plot looks very similar to the one in task 4. However two points have drastically different values than the corresponding value on the curve. These can be seen as outliers.

(b)

By removing the outliers we get a plot that is nearly identical to the one in task 4. Calculating the SSE with the outliers gives a value of about 9.890. Without the outliers we get a value of about 1.331. Only two points made that big of a difference. We can therefore assume that the LS-algorithm is very sensitive to outliers. Outliers effects the total SSE a lot.

2.2

2

The Ridge regularization method and the Lasso regularization method come in play when the X^TX matrix is singular. When this happens,the least squares model (||y-xβ||^2) is not unique, and infinite solutions to β ($(X^TX)^{-1}X^TY$) are available. Therefore, we can compute the coeficients of the models either by:

- Ridge Regression: $\beta_{hat} = arg\,min_\beta\,||y - x\beta||^2 + \lambda||\beta||^2$. The Ridge method aims to represent all the data, keeping the coefficients small. When a coefficient is large, the term $||\beta||^2$ penalizes it, which keeps the coefficients small, but the data is still present.The coeficients will shrink to 0, but never reach it. λ will serve as a trade-off between both objectives(speed which it minimizes the coefficients of the model and representation of data).

- Lasso Regression: $\beta_{hat} = arg\,min_\beta\,||y - x\beta||^2 + \lambda||\beta||_1$. The Lasso regression aims to minimize the sum of squared errors.In this case, the regularization term will be $\lambda||\beta||_1$.

Now, why can we use the Lasso model for feature selection but can´t use Ridge? It all comes down to the way the coefficients are handled. In the Lasso Regression, the coefficients will shrink all the way to 0. This means that, when minimizing the function, and the sum of the absolute value of the coeficients is higher than a fixed value, the lasso regression will shrinks the coefficients , to the point where some will turn to 0, effectively ignoring those coefficients in the linear regression. This acts as a feature selection, because you can fit the model t make a linear regression solely focusing on some parameters and ignoring others, like it is a simpler model that fits said parameters. This will create a linear regression that fits better the ”features” remaining,because all the other features that are nullified won´t come in account when performing the linear regression.

6

In the plots obtained, we can see that the Lasso model shrinks it´s coefficients much faster than the Ridge model. While all the coefficients obtained by the Lasso model have already shrinked to 0 by the time alpha is less than 10, the coefficients obtained by the Ridge method are starting to decrease slowly by the time alpha is close to 10. The irrelevant feature here will be the coeficient 2, or the red line in both graphs, since it is the first one to be shrinked (and in the Lasso model it goes quite fast to 0).

7

To pick the adequate value of α, we picked the value that first make the coefficient 2 (red line) 0, since that will be when there will occur feature selection, which leads to a better linear regression fitted to the other parameters. Besides, that point will be when the two other coefficients are the ”closest” to their LS counterparts. We can see that, even though both the LS model and the Lasso model didn´t exactly overlapped the values in the data set, they were relatively close to said values, which is a satisfying result. In this case, the value of SSE from the LS model was lower than the SSE value of the Lasso model. But, yet again, both were somewhat close (and low) values, which is a prove that the models are reaching good approximations when given a set of data to extrapolate.