Instituto Superior Técnico

MEEC

Machine Learning

Lab 1

Linear Regression

Group 9

Manuel Diniz, 84125
Alexandre Rodrigues, 90002

Turno: 4ªf 11h00

Contents

1.1 Matrix expressions

The *Least Squares* method relies on minimizing the sum of squared errors of a given model:

$$SSE = ||y - X\beta||^2 \tag{1.1}$$

Where X is a matrix defined, in this case, by:

$$\begin{bmatrix} 1 & x_1^1 & ... & x_1^p \\ ... & ... & ... & ... \\ 1 & x_n^1 & ... & x_n^p \end{bmatrix} \tag{1.2}$$

Since we are working with a polinomial model who's dataset has n elements and is of order p.

By calculating the zero of the gradient of the *SSE*, it is possible to analytically solve the model, through the normal equation:

$$(X^T X)\beta = X^T y$$
$$\beta = (X^T X)^{-1} X^T y$$

Thus obtaining the coefficients of the model.

1.2 Least Squares fitting

For files 1 and 2, the models fits the data quite accurately. They also ignore the noise effectively because of the low order of the polynomials, with no signs of overfitting.

As for 2a, the model fits the data with decent accuracy. The inlier-only SSE is significantly larger than the one of the previous case. Seeing as both of them have the same data with the exception of the outliers, we can infer that the high sensitivity of the LS method to outliers was what caused the increase in *SSE* for the inliers.

It is also possible to view the effect of the outliers through the plot. The model follows the part of the wave with a negative slope with a significant error, the curve being above the inliers significantly. In layman's terms, it can be said that the two outliers above the wave "pulled" the fit closer to them, and away from the inliers.

It can also be explained mathematically, due to the fact the method gives greater "importance" to outliers, as the those points contribute to the cost function with a large error, which is then squared.

1.3 Ridge regression and Lasso regression

Ridge regression is a form of regression which has a regularization term, which penalizes coeficients with large values, a way of preventing overfitting and selecting for relevant features. It is simmilar to the LS method, with it's optimization function being given by the following expression:

$$min(SSE + \lambda||\beta||^2) \tag{1.3}$$

The first term being the SSE, and the second the regularization term. A large coefficient vector leads to a large cost function, which is contrary to the objective of the algorithm, which is to minimize it.

Lasso regression is simmilar, but uses a different regularization term, given by:

$$\lambda||\beta||_1^2 \tag{1.4}$$

Where the norm is the *l1* norm, a simple sum of the absolute values of the coefficients.

This forces the sum of the coefficients to be below a certain value, directly related to λ, which then forces some of the coefficients to be zero if λ is sufficiently large. It is superior at feature selection when compared to ridge regression, as it reduces the absolute value of the coefficients equally and independently of their value. The former has "diminishing returns" for small coefficients, meaning it will reduce the length of the coefficient vector, but generally doesn't eliminate coefficients entirely. As such, ridge regression doesn't select for features as well.

Relating to file 3, the results match the what was previously stated, with lasso regression quickly selecting for the relevant features. The feature who's coefficient was nullified first was the second, meaning it is the irrelevant feature.

Another thing to note is that the coefficients of the two methods are identical to the coefficients of the least squares method when λ is zero. This makes sense from a mathematical standpoint, as it nullifies the second term of the cost function, making it identical to the sum of squared errors.

Now considering only lasso regression, the chosen value for λ is 0.071, which is the first value shown to nullify the irrelevant term completely.

The lasso method results in a slightly larger squared error when compared to the least squares method, which is natural seeing as how the λ term also influences the other coefficients, making them smaller, but also making the model less accurate as a result. Still, the difference in *SSE* is negligible, especially when compared to the computational power the lasso method saves when predicting. Seeing as how one of the coefficients was completely nullified, the processing power required to compute the prediction goes down by one third, as one of the three features is ignored. This would have a significant advantage if the model had to be applied a large number of times, like is often the case.