M124 Coursework 2

Vangelis Tsiatouras
cs2190021@di.uoa.gr

April 26, 2021

1 Naive Bayes

Question:
In many ML approaches, the data are usually preprocessed in such a way as to have zero mean and standard deviation of 1. Is this a good idea in general? Explain your thought. Would this preprocessing help the Naive Bayes Classifier? If so, how?

Answer:
This technique is called *standardization*. By removing the average value of each feature we can center the data. Also, by dividing each of the characteristics with their standard deviation we manage to rescale the data. In that way we can remove the bias from the features, although in Naive Bayes, which is a probabilistic model, the final result is not affected with the usage of this preprocessing technique. The reason is that the estimated probabilities will remain exactly the same, even after centering and rescaling the data because Gaussian Naive Bayes performs standardization internally.

2 Shrinkage Methods: Regularization

2.1 Theoretical

1. Derive the solution of the Ridge Regression optimization problem by hand. Show the intermediate steps.

   The optimization problem of Ridge Regression reads:

$$\underset{\beta}{argmin} \left\|\mathbf{y}_n - X_n\beta\right\|_2^2 + \lambda \left\|\beta\right\|_2^2 \tag{1}$$

   We can extract from the above formula the following function:

$$\begin{aligned} F(\beta, \lambda) &= \left\|\mathbf{y}_n - X_n\beta\right\|_2^2 + \lambda \left\|\beta\right\|_2^2 \\ &= (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\ &= (y^T - X^T\beta^T)(y - X\beta) + \lambda\beta^T\beta \\ &= y^Ty - X\beta y^T - X^T\beta^Ty + X^T\beta^TX\beta + \lambda\beta^T\beta \\ &= y^Ty - 2X^T\beta^Ty + X^T\beta^TX\beta + \lambda\beta^T\beta \end{aligned} \tag{2}$$

In order to derive the RR optimization we must yield the partial derivative with respect to β.

$$\begin{aligned} \frac{\partial F}{\partial \beta} = 0 \xRightarrow{\beta^T\beta \to \beta^2} 2X^TX\beta - 2X^Ty + 2\lambda\beta &= 0 \\ 2\beta(X^TX + \lambda I) - 2X^Ty &= 0 \\ 2\beta(X^TX + \lambda I) &= 2X^Ty \\ \beta(X^TX + \lambda I) &= X^Ty \\ \beta &= (X^TX + \lambda I)^{-1}X^Ty \end{aligned} \tag{3}$$

2. Explain the differences between Least Squares and Ridge Regression, and why RR is more robust to overfitting.

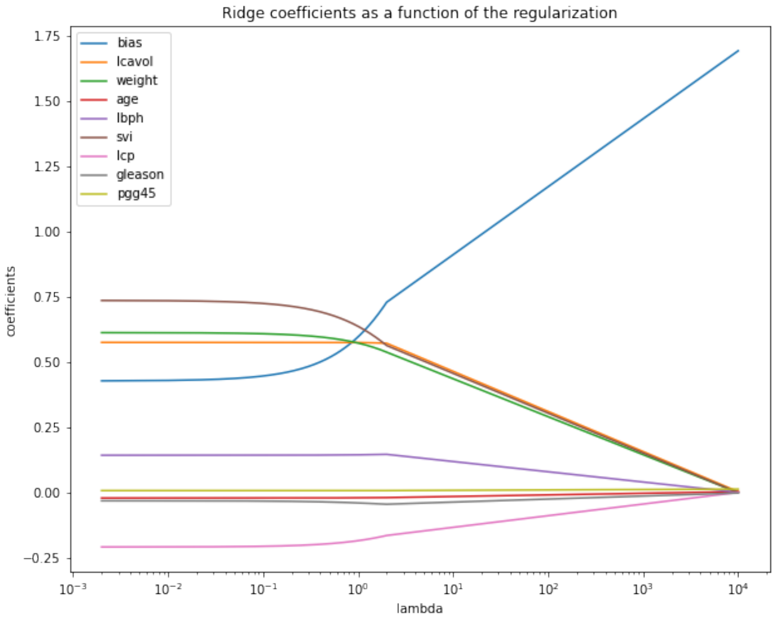
   The main difference between Least Squares and Ridge Regression is that RR is not unbiased, meaning that accepts bias to reduce variance which leads to the reduction of the mean squared error and increased prediction accuracy. Although, the predictions of RR may seem more stable by shrinking coefficients but the model is not very sensitive to the data. More specific, as λ increases the flexibility decreases (increased bias / decreased variance). Also, Linear Regression does not handle very well the presence of correlation between the predictors. Multicollinearity can create inaccurate coefficient estimations, leading to degraded predictability of the Linear Regression model (overfit) contrary to Ridge Regression which overcomes this problem by applying higher bias and decreased variance.

3. We talked about the bias-variance trade-off in Lecture 5. How does the value of λ affect the bias and variance of the estimator?

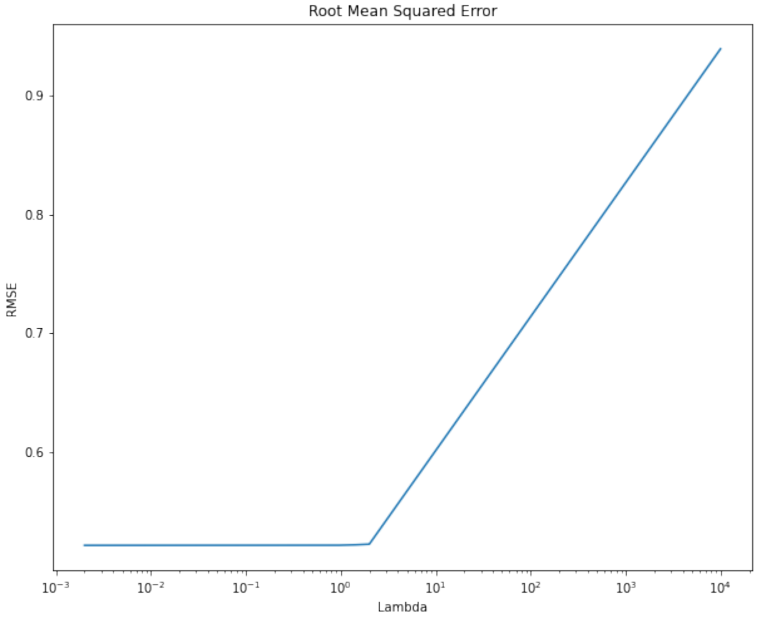
   The λ parameter affects crucially the bias and the variance of the estimator. As the λ increases the bias increases and variance decreases. In the next chapter of this assignment this behaviour is observed and is provided a plot that describes the relationship between λ, bias and the variance of the coefficient estimations.

2.2 Ridge Regression Implementation Report

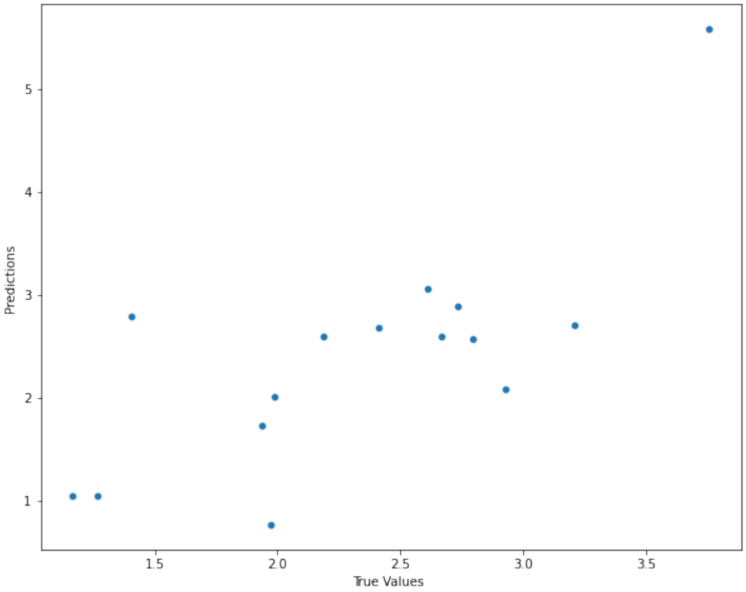
1. In the following plot it can be seen how the λ affects the bias and the variance of the model. As the λ increases, the bias increases and the coefficients of the features tend to 0 (variance decreases).



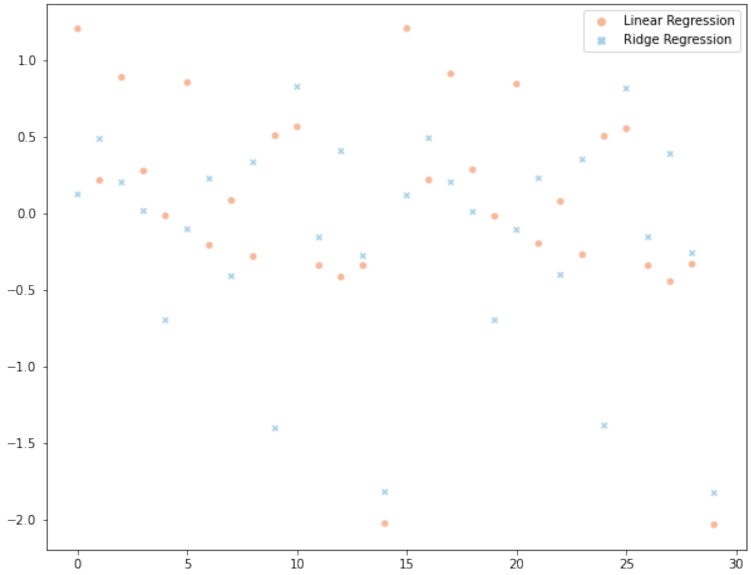
2. In the following plot, it is visualized the RMSE (Root Mean Squared Error) to the regularization parameter λ. It is noticed that for large values, the RMSE is increased significantly. Also the RMSE minimizes for $\lambda = 0.354$ with value $RMSE = 0.5211839118146687$.



3. In the next plot we display for 15 random points of the test set their predicted value contrast to their actual value, with $\lambda = 0.354$.



4. Finally, the next plot shows the residuals of Ridge Regression contrary to Linear Regression. It is worth to mention that the Linear Regression model achieves $RMSE = 0.5212740055076177$.



Concluding, the following table shows the coefficients of each model (Ridge Regression uses $\lambda = 0.354$). The assumption is that Ridge Regression yields better results than Linear Regression.

	Ridge Regression	Linear Regression
bias	0.4939205964598763	0.42917013
lcavol	0.576747003107279	0.57654319
weight	0.5993205294774145	0.61402
age	-0.018674384427299545	-0.01900102
lbph	0.1453820370773281	0.14484808
svi	0.6983674176460376	0.73720864
lcp	-0.19718780977168418	-0.20632423
gleason	-0.03293609509756401	-0.02950288
pgg45	0.009497899097753473	0.00946516