1 Implementation of Ridge Regression

Ridge Regression is a linear model for regression that takes into account a regularization term in the objective function. Starting from the linear regression objective function:

$$argmin_w \Lambda(w) = argmin_w \frac{1}{N}(y - Xw)^T(y - Xw)$$

We include the L2-regularization term[1]

$$argmin_w \Lambda(w) + \lambda R(w) = argmin_w \frac{1}{N}(y - Xw)^T(y - Xw) + \lambda w^T w$$

Thus obtaining the closed form solution for w (**Ridge regressor estimate**)

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

In this way, we are able to keep the values of w low, since very large values of w could make our model very sensitive, thus leading to poor generalization. Furthermore, constraints added via regularization may help us in solving ill-posed problems, when $X^T X$ is not possible to invert (when $D >> N$). Therefore, λ becomes an hyper-parameter of our model controlling the strength of the regularization. After obtaining the vector of weights $\hat{w}$ we can perform regression in the same way as a non regularized linear regression in the form: $y = \hat{w}^T x$

1.1 The train function

For the training phase we simply need to compute the closed form solution to obtain our weights. We only need to pay attention in transforming our data in order to absorb the bias term and set $I(0,0) = 0$ before computing the solution, so that we do not regularize the bias term w_0

1.2 The predict function

Again, the predict function is very simple, we just need to transform the data we want to predict in order to take into consideration the bias term, and then compute[2] $Y = Xw$

2 Validation of the model

Since we are working on a dataset (Olympics 100m) with very few samples (29), after splitting it in training set (80%) and test set(20%) using a random stat to obtain reproducibility, we performed **k-fold cross validation leave-one-out** with a **coarse-to-fine** approach to choose the best value of λ. Basically, for each λ we train our model[3] on $N-1$ samples and then we compute the MSE on the prediction of the left-out element. After doing so for all samples in the evaluation set, we compute the mean of the MSEs we obtained, that we will then use to evaluate the goodness of a certain λ. Doing so, we obtain that the best model is with $\lambda = 416.67$, with $MSE = 0.0659$.

3 Test of the model

We also tested our best configuration on a never-seen test set (20% of our dataset) after training on our training set, and we obtained an $MSE = 0.0487$. We can say that our model generalized well.

4 Comparison with scikit-learn

We then also compared our implementation with the one of scikit, using the λ we selected as the best one. As we can see in the plot, the two are practically the same.

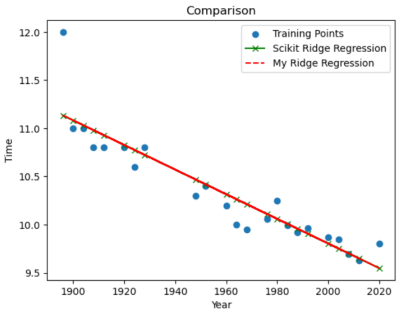

Figure 1: Comparison Plot

[1] Ridge Regression uses L2-regularization term
[2] We are doing matrix multiplications, thus we obtain an array of predicted values
[3] N is the number of samples in our dataset