CS5691: Pattern Recognition and Machine Learning
Assignment 2

K Kamalesh Kumar
CE20B054

April 1, 2023

In this report, Linear regression and its variants (Ridge regression and Lasso regression) are experimented, and the observations and results are detailed in the sections below.

1 Question 1

1.1 Part 1

Given the data matrix, X and the ground truth targets Y, the maximum likelihood solution for the linear regression problem is given by:

$$w_{ML} = (XX^T)^{-1}XY$$

1.2 Part 2

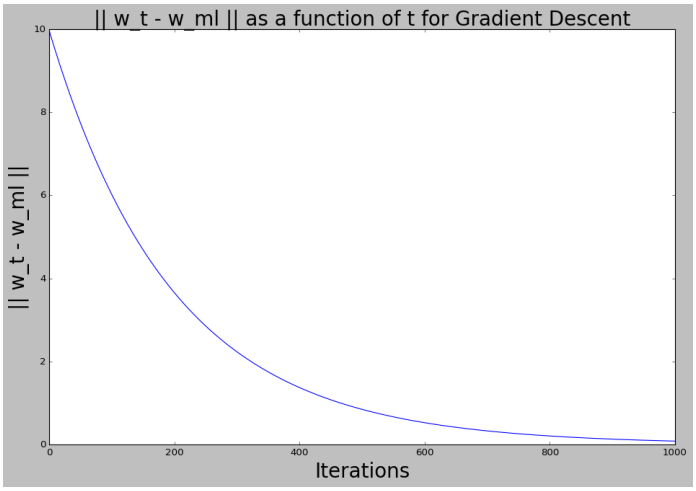

Figure 1: $\|w_t - w_{ML}\|$ as a function of t

The above figure plots the error of the least squares solution obtained through gradient descent from the closed-form solution of the same, as a function of iteration t.

Clearly, the gradient descent estimate converges to the actual value w_{ML} by around 1000 iterations, with a step size of 3×10^{-2}. The convergence is sensitive to the step size, with larger values producing diverging results. The gradient and gradient update are as follows:

$$f(w) = \frac{1}{n} \|Xw - Y\|^2$$
$$\nabla_w = \frac{1}{n}(2XX^T - 2XY)$$
$$w^{t+1} = w^t - \eta \nabla_w$$

1.3 Part 3

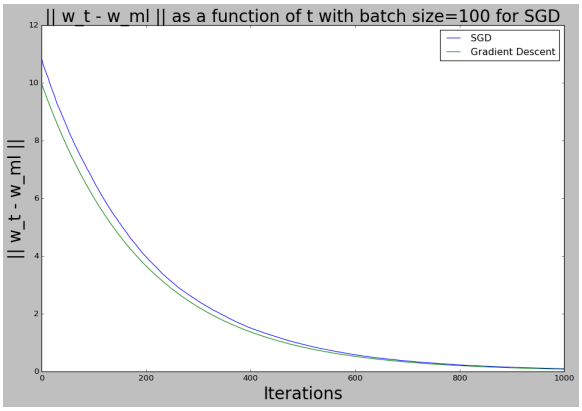

Figure 2: $\|w_t - w_{ML}\|$ as a function of t for SGD with batch size 100

The above plots show $\|w_t - w_{ML}\|$ as a function of t for both stochastic gradient descent with batch size 100 and full batch gradient descent. Clearly, both the algorithms converge reasonably within 1000 iterations, although the full batch gradient descent converges faster (the error is lower than SGD), which is expected since it uses the true gradient while updating w

For a batch size of 100, both SGD and true gradient descent seem alike and not very different. By decreasing the batch size to 2, the variance increases for SGD and is explicit in the below plot

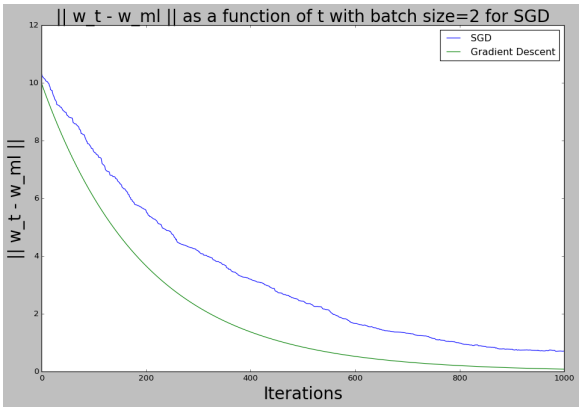

Figure 3: $\|w_t - w_{ML}\|$ as a function of t for SGD with batch size 2

2 Question 2

2.1 Part 1

The gradient for the ridge regression problem is given by

$$\nabla_w = \frac{1}{n}(2XX^T - 2XY) + 2\lambda w$$

The implementation details and code are present in the jupyter notebook submitted.

2.2 Part 2

For the cross-validation, a validation size of 2000 was considered, that is with $k = 5$. The choices of the regularization parameter considered were $\lambda = 0.01, 0.1, 1, 10, 100$. The validation error $\frac{1}{\text{val size}} \left\| \hat{Y}_{val} - Y_{val} \right\|^2$ was plotted as a function of iterations t for each choice of λ, which is shown below:

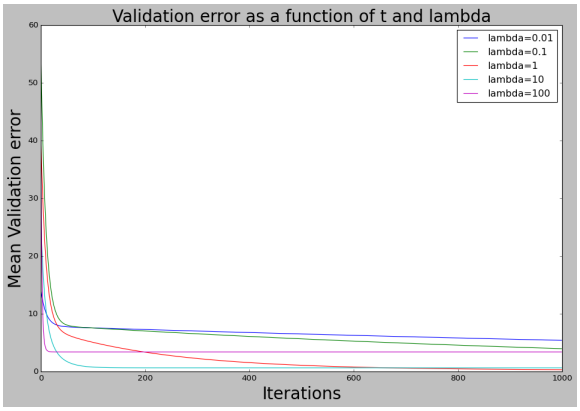

Figure 4: Validation error as a function of λ and t

At the end of $T = 1000$ iterations, the validation error for each choice of λ were [5.39622278, 3.96015787, 0.32030762, 0.63658711, 3.38388932]. Clearly, from this and the above plot, $\lambda = 1$ had the lowest mean validation error. Now using $\lambda = 1$, the closed form solution for the ridge regression problem is given by:

$$w_R = (XX^T + I)^{-1}XY$$

And the prediction given by using w_R and w_{ML} on the test data is given by:

$$\hat{Y}_{test,R} = X_{test}^T w_R = X_{test}^T (X_{test}X_{test}^T + I)^{-1} X_{test}Y_{test}$$
$$\hat{Y}_{test,ML} = X_{test}^T w_{ML} = X_{test}^T (X_{test}X_{test}^T)^{-1} X_{test}Y_{test}$$

And the test error were found to be:

$$\text{Test error using } w_R = \frac{1}{\text{test size}} \left\| \hat{Y}_{test,R} - Y_{test} \right\|^2 = 0.3703039794569779$$
$$\text{Test error using } w_{ML} = \frac{1}{\text{test size}} \left\| \hat{Y}_{test,ML} - Y_{test} \right\|^2 = 0.37072731116978747$$

The test error for both cases is nearly the same, with the ridge regression giving a slightly lower test error than the maximum likelihood solution, mainly due to its better generalization on the test data, and overfitting issues that arise with the latter.