Homework 2
Problem 1
Ayad Masud
733009045

I certify that I have personally done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment. ARM.

1 Part 1: Ordinary Least Squares Solution

As shown in figure 1 we can see that over 100 iterations of training and testing the model we get a normally left skewed distribution. We can see for the most part the R^2 score is quite high, around 0.93, but there are some outliers where the model did not perform as well. This can be explained by the randomly split data set where some of the train-test splits may have more variance or outliers than other iterations.

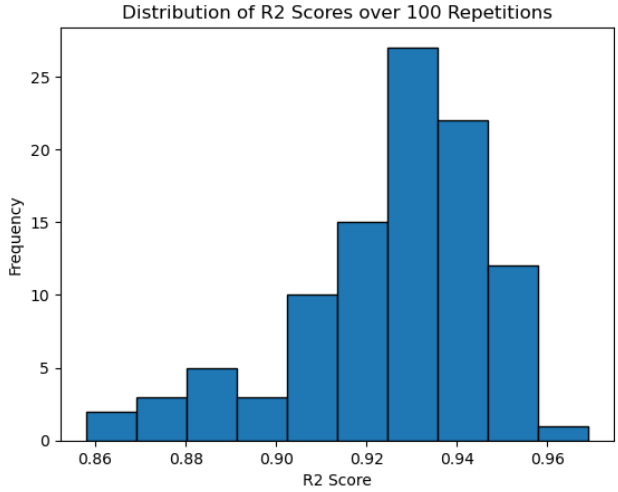

Figure 1. Histogram of R^2 scores over 100 iterations

2 Part 2: How R^2 changes as a function of noisy channels and number of samples.

Figure 2 shows how increasing the number of samples affects the R^2 score. We can see that increasing the number of examples from 100 to 500 shows a much more stable histogram where the R^2 scores don't have much variance. This is shown by the almost equal heights of several of the bars on the plot. Figure 3 shows how changing the number of noisy channels affect the R^2 score. When we increase the number of noisy channels it decreases the R^2 score, we can see the highest R^2 score on the histogram is only 0.4. This is because adding noisy channels does not add any relevant information for the model to use.

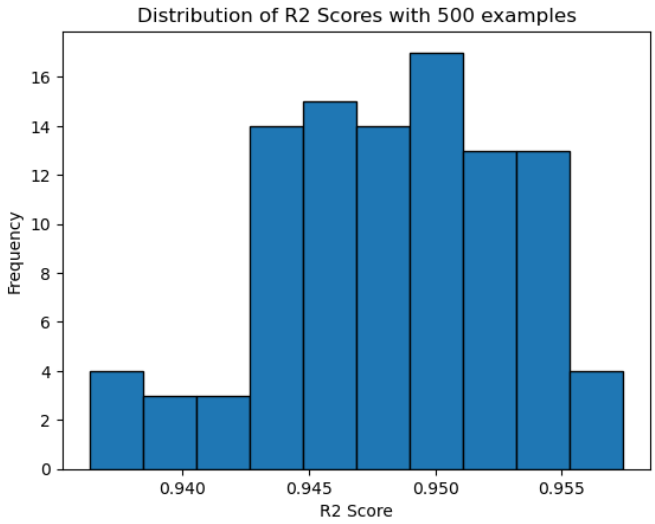

Figure 2. Distribution of R^2 score with 500 examples

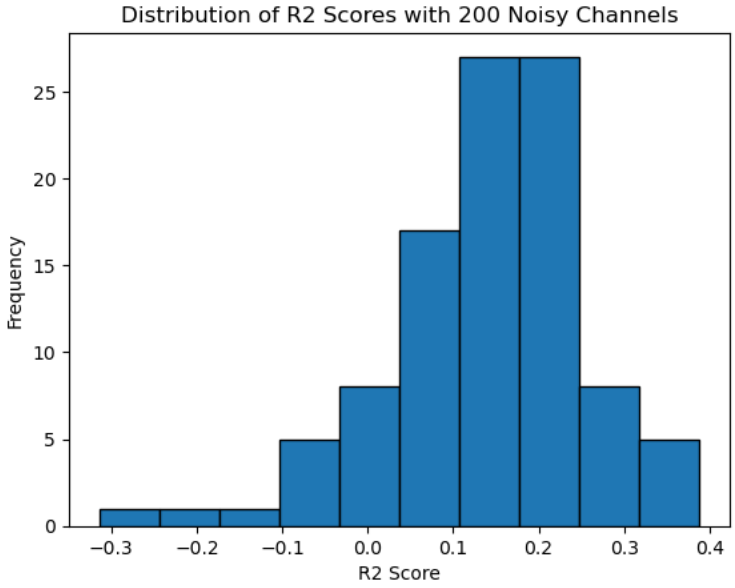

Figure 3. Distribution of R^2 scores with 200 noisy channels

3 Part 3: Regularized OLS (Ridge Regression)

Figure 4 shows the how the R^2 score is affected as we increase alpha, or the regularization constant used in ridge regression. It has negative correlation as when we increase alpha the value of R^2 decreases. At alpha=0 we have the ordinary least squares solution that we see in linear regression because there is no regularization constant applied. At high values of alpha, we can see that R^2 goes down. This is because the model becomes overly simple and cannot capture the details needed to make an accurate prediction. The most optimal value of alpha is found by balancing both the bias and variance.

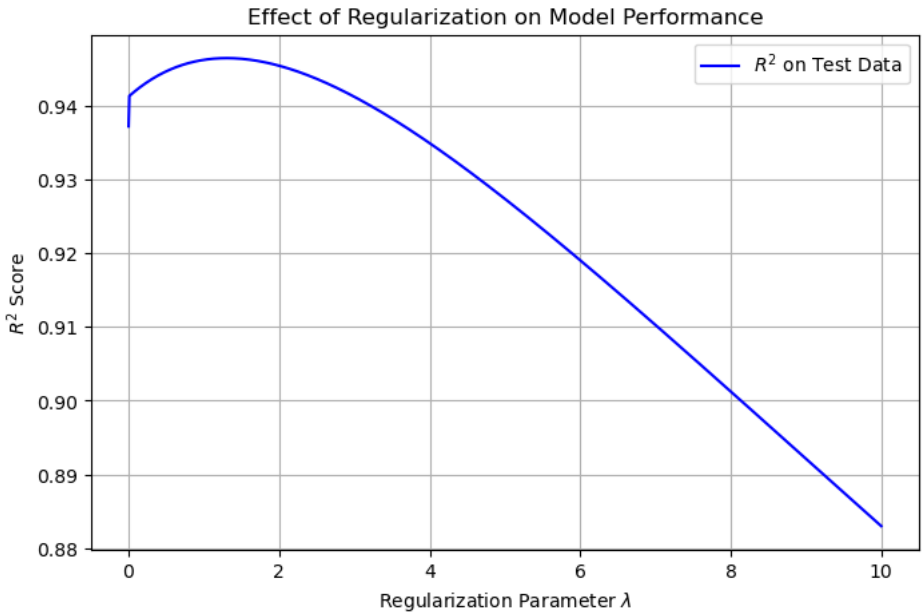

Figure 4. R^2 score vs Regularization Parameter Alpha

4 Resources used to achieve this goal

Canvas: Homework template

Python Libraries: NumPy, pandas, matplotlib, Scikit Learn

5 References

- "1.1. Linear Models — Scikit-Learn 0.24.0 Documentation." *Scikit-Learn.org*, scikit-learn.org/stable/modules/linear_model.html.
- "Linear Regression Example — Scikit-Learn 0.20.3 Documentation." *Scikit-Learn.org*, 2018, scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py.
- "Ridge." *Scikit-Learn*, 2024, scikit-learn.org/1.5/modules/generated/sklearn.linear_model.Ridge.html.