|  |  |  |  |
| --- | --- | --- | --- |
|  | Normal Train-test split  Lambda ranges [0,150] | Normal Train-test split  Lambda ranges [1,150] | Cross-Validation |
| Dataset 1 | A graph with a red line and blue line  Description automatically generated |  | A graph with a green line  Description automatically generated |
| Dataset 2 | A graph with red and blue lines  Description automatically generated | A graph with a blue line and red line  Description automatically generated |  |
| Dataset 3 | A graph with a red line and blue line  Description automatically generated |  |  |
| Dataset 4 | A graph with red and blue lines  Description automatically generated |  |  |
| Dataset 5 |  |  |  |
| Dataset 6 |  |  |  |

A screenshot of a math notebook

Description automatically generated

2.

a. Summary of Best λ Values

* Dataset 1: λ = 8, MSE = 4.16
* Dataset 2: λ = 22, MSE = 5.08
* Dataset 3: λ = 27, MSE = 4.32
* Dataset 4: λ = 8, MSE = 5.54
* Dataset 5: λ = 19, MSE = 5.21
* Dataset 6: λ = 23, MSE = 4.85

**b.** Summary of Best λ Values

* Dataset 2: lambda = 22, MSE = 5.08
* dataset 4: lambda = 8, MSE = 5.54
* dataset 5: lambda = 19, MSE = 5.21

c. When lambda = 0 the model reduces to ordinary least squares regression which leads to overfitting the

training data, including noise/outliers. Also, will involves the issue of high sensitivity to outliers in training data

and inflate the error on test data

3.

a. Summary of Best λ Values:

* dataset 1: lambda = 0, MSE =3.07
* dataset 2: lambda = 7, MSE =3.73
* dataset 3: lambda = 28, MSE =3.88
* dataset 4: lambda = 26, MSE = 5.15
* dataset 5: lambda = 33, MSE =6.39
* dataset 6: lambda = 33, MSE =4.96

b. Comparison of λ and MSE from CV vs. Normal Train-Test Split:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Lambda (CV)** | **MSE (CV)** | **Lambda (Train-Test)** | **MSE (Train-Test)** | **Comparison** |
| **Dataset 1** | 0 | 3.07 | 8 | 4.16 | CV λ is better as it results in lower MSE |
| **Dataset 2** | 7 | 3.73 | 22 | 5.08 | CV λ is better as it results in lower MSE |
| **Dataset 3** | 28 | 3.88 | 27 | 4.32 | CV yields a similar MSE with a slightly higher λ. |
| **Dataset 4** | 26 | 5.15 | 8 | 5.54 | CV λ is better as it results in lower MSE |
| **Dataset 5** | 33 | 6.39 | 19 | 5.21 | CV yields higher MSE |
| **Dataset 6** | 33 | 4.96 | 23 | 4.85 | The MSEs are similar, but CV uses a higher λ |

c. Drawbacks of Cross-validation:

- *Increase in computation cost and time:* The training process is more time-consuming due to multiple training runs (10 in this case), which may not be feasible for very large datasets or complex models.

- *Potential Overfitting:* While CV helps in estimating the model performance, there’s still a risk of overfitting if the same data points are repeatedly validated or if the folds are not representative of the overall dataset.

- *Complex method*: Implementing cross-validation can be more complicated compared to simple train-test splits due to data partitioning and additional code.