Assignment 1

Fit a polynomial model.

David Padilla
Ignacio Pastore Benaim

September 27, 2024

1 Introduction

In this report, we present the results of the first assignment of the Machine Learning course. The assignment consists of fitting several regression models to a noisy dataset generated from a polynomial function. The figure 1 shows the dataset which contains 500 data points. The majority of the methods seen in class were explored for the sake of completeness. A comparison of the results is presented along with the election of the optimal model.

We explore multiple methods, including k-NN, linear regression variants, and regularized polynomial regressions (Ridge and Lasso). The optimal model is selected based on cross-validation, and final test results are reported. Finally, a short discussion and conclusion are presented.

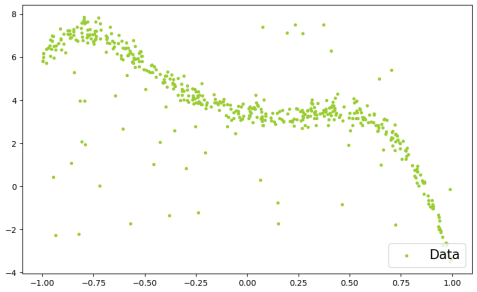

Figure 1: Polynomial function generated with random noise.

2 Methods

The original dataset consists of 500 data points, which were split into training (81%), validation (9%), and test (10%) sets. No outlier removal was performed. All models were trained using the training set, evaluated with the validation set for model selection, and finally evaluated on the test set after model selection. The evaluation process involved three key steps: initial screening with validation, cross-validation of selected models, and final evaluation on the test set.

2.1 Initial Screening with Validation

Various models, including k-NN, linear regression variants, and polynomial regression with regularization, were first evaluated using a validation set. The purpose of this step was to quickly assess model performance and identify promising candidates.

2.2 Cross-Validation of Selected Models

Based on the validation results, a subset of models was chosen for cross-validation. Cross-validation was used to provide a more robust evaluation with $k = 10$ folds. The average performance across the folds was used to further narrow down the best-performing models.

2.3 Final Testing on Test Set

After cross-validation, the best models were evaluated on the test set to determine their generalization performance.

2.4 Models

We tested several different models, including k-Nearest Neighbors (k-NN), various linear regression models(Huber and RANSAC), and polynomial regression with regularization (Ridge and Lasso).

2.4.1 k-Nearest Neighbors (k-NN)

The k-NN algorithm was tested with CAMBIAR Y AGREGAR GRIDSEARCH $k = 1$, $k = 5$, $k = 10$, $k = 15$, and $k = 20$. The optimal k value was chosen based on validation set performance.

2.4.2 Linear Regression Variants

Despite knowing that a linear model might not be the outmost appropiated for this task, 3 variants were implemented to compare with other methods:

- Standard Linear Regression.
- Huber Regression.
- RANSAC Regression.

2.4.3 Polynomial Regression with Regularization

Polynomial regression models were fitted with degrees ranging from 2 to 12. Additionally, Lasso and Ridge regression were applied with a regularization parameter $\lambda = 0.01$, obtained from a grid search over $\{0.01, 0.1, 1, 10, 100\}$.

2.5 Evaluation Metrics

The models were evaluated using two primary metrics:

- Mean Squared Error (MSE).
- R^2 Score.

3 Results

The following steps were taken to assess model performance:

3.1 Initial Validation Results

The first screening of models was done using the validation set. Table **??** shows the MSE and R^2 scores for the various models during the validation phase. Based on these results, several models were selected for cross-validation.

3.2 Cross-Validation Results

The selected models (k-NN with $k = 15$ CAMBIAR 15, polynomial regression with degree 6, and Ridge regression with degree 6 and $\lambda = 0.01$) were then evaluated using cross-validation. The cross-validation results are summarized in Table **??**. Cross-validation confirmed that k-NN and polynomial regression were the best candidates.

3.3 Test Set Results

Finally, the selected models were tested on the test set. Table 3 shows the final MSE and R^2 scores on the test set for the top-performing models. k-NN with $k = 15$ CAMBIAR ESTE 15 provided the best generalization, slightly outperforming polynomial regression with degree 6.

4 Discussion

5 conclusion

Table 1: Model Performance with MSE and R^2

Model	MSE	R2
KNN (k=17)	0.514	0.916
Linear	0.978	0.841
Huber	1.059	0.828
RANSAC	1.266	0.794
Polynomial (degree=2)	1.002	0.837
Ridge (degree=2, $\lambda = 0.1$)	1.002	0.837
Lasso (degree=2, $\lambda = 0.1$)	1.097	0.821
Polynomial (degree=3)	0.757	0.877
Ridge (degree=3, $\lambda = 0.1$)	0.756	0.877
Lasso (degree=3, $\lambda = 0.1$)	0.862	0.860
Polynomial (degree=4)	0.447	0.927
Ridge (degree=4, $\lambda = 0.1$)	0.438	0.929
Lasso (degree=4, $\lambda = 0.1$)	0.759	0.877
Polynomial (degree=5)	0.435	0.929
Ridge (degree=5, $\lambda = 0.1$)	0.436	0.929
Lasso (degree=5, $\lambda = 0.1$)	0.759	0.877
Polynomial (degree=6)	0.425	0.931
Ridge (degree=6, $\lambda = 0.1$)	0.457	0.926
Lasso (degree=6, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=7)	0.509	0.917
Ridge (degree=7, $\lambda = 0.1$)	0.461	0.925
Lasso (degree=7, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=8)	0.511	0.917
Ridge (degree=8, $\lambda = 0.1$)	0.453	0.926
Lasso (degree=8, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=9)	0.565	0.908
Ridge (degree=9, $\lambda = 0.1$)	0.467	0.924
Lasso (degree=9, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=10)	0.562	0.909
Ridge (degree=10, $\lambda = 0.1$)	0.456	0.926
Lasso (degree=10, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=11)	0.542	0.912
Ridge (degree=11, $\lambda = 0.1$)	0.475	0.923
Lasso (degree=11, $\lambda = 0.1$)	0.700	0.886
Polynomial (degree=12)	0.544	0.912
Ridge (degree=12, $\lambda = 0.1$)	0.469	0.924
Lasso (degree=12, $\lambda = 0.1$)	0.700	0.886

Table 2: Model Performance with MSE and R^2

Model	MSE	R2
Polynomial (degree=6)	1.770	0.682
Ridge (degree=6, alpha=0.01)	1.769	0.682
Polynomial (degree=7)	1.763	0.684
Ridge (degree=7, alpha=0.01)	1.765	0.683
KNN (k=15)	1.804	0.704
KNN (k=20)	1.837	0.700

Table 3: Tests

Model	MSE	R2
KNN (k=17)	1.127	0.807
Polynomial (degree=6)	1.146	0.803
Ridge (degree=6, $\lambda = 0.01$)	1.143	0.804