Atefe Rajabi – 40230563

Problem Definition

The chosen dataset for this project is the Boston Housing Dataset, which involves predicting the median value of owner-occupied homes (MEDV) in $1000s based on selected housing-related features. This project aims to explore a regression problem where the objective is to develop a model that accurately estimates housing prices using a subset of features.

Features Selected:

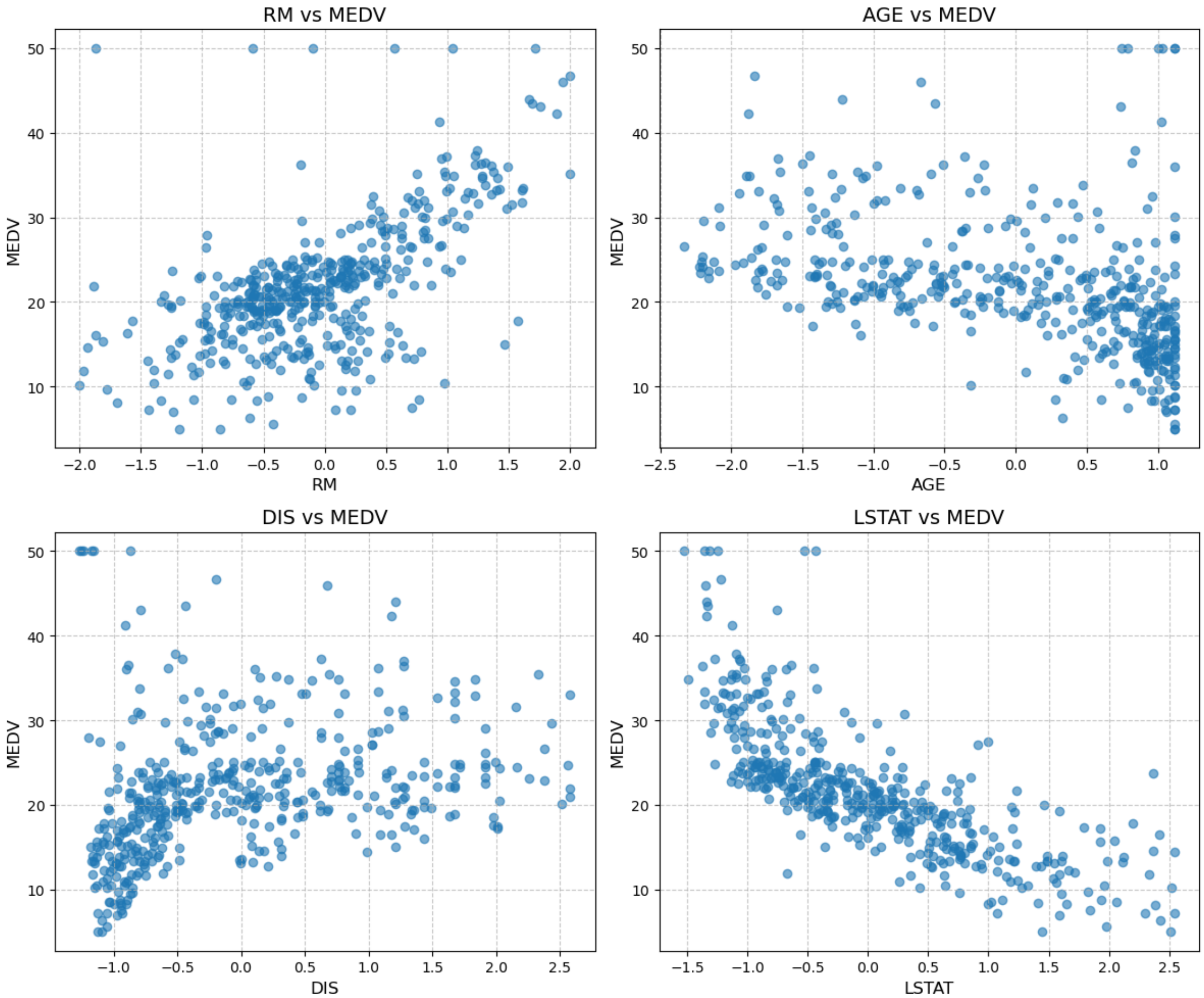
1. RM: Average number of rooms per dwelling.
2. AGE: Proportion of owner-occupied units built before 1940.
3. DIS: Weighted distances to five Boston employment centers.
4. LSTAT: Percentage of the population considered lower status.

Problem Definition:

Using the selected features (RM, AGE, DIS, and LSTAT), the goal is to:

1. Build a regression model that predicts the target variable MEDV (Median value of owner-occupied homes in $1000s).
2. Determine the optimal coefficients and offsets for the model based on the training set (80% of the dataset).
3. Analyze the contribution and impact of each feature separately by assessing the individual errors associated with them.

Dataset Visualization



preprocessing

The preprocessing of the dataset involves two key steps to prepare the data for analysis. First, the selected features are standardized to have a mean of zero and a standard deviation of one. This ensures that all features are on the same scale, which is particularly important for models sensitive to the magnitude of the data. Second, an outlier removal process is applied using the interquartile range (IQR) method. This method identifies and removes data points that fall outside 1.5 times the IQR from the first and third quartiles for each feature, reducing the influence of extreme values.

Linear regression

To apply linear regression, it is necessary to minimize an error function, which can be based on different norms such as L0, L1, L2, or L∞. For the L2 norm, the error function is differentiable, allowing us to derive it with respect to the model parameters. Using gradient descent, we can iteratively optimize the parameters to find the best fit. However, for L0, L1, and L∞ norms, the error functions are not differentiable, requiring alternative optimization approaches. In this context, methods such as simulated annealing or alternate search strategies are applied to optimize the model parameters effectively.

Derivation-Norm2

Feature	Slope	Bias
RM	6.528021410116338	22.069036130734332
AGE	-3.4642306183938376	21.93099080438914
DIS	2.401859143298355	21.855126476651453
LSTAT	-6.555160631236191	21.927710080734926

Feature	Train Error (Norm-2)	Test Error (Norm-2)
RM	117.817588779328	63.490327140929196
AGE	137.47858352210596	62.81559053413142
DIS	145.61392663244925	65.95703763869105
LSTAT	99.91472888261491	46.54250684479068
Average Prediction	113.68419406608756	53.630187052981555

SKLearn Library

Applying linear regression on whole features once.

Mean Squared Error: 46.174958773873314 on test data
Coefficients: [2.94198891 -0.48296217 -0.96901766 -5.27106164]
Intercept: 22.02022636917157

Simulated Annealing

Result (norm0_residuals_objective)

Feature	Slope	Bias	Error
RM	0	0	467
AGE	0	0	467
DIS	0	0	467
LSTAT	0	0	467

Result (norm1_residuals_objective)

Feature	Slope	Bias	Error
RM	6.5051656134616165	22.48537167776916	1948.0868465042695
AGE	-3.1325042271408576	20.417808199392947	2235.9301414428865
DIS	2.547669891664618	20.545870351344792	2424.6394023438015
LSTAT	-5.8940133459570525	20.676839853903715	1663.4834308264105

Result (norm2_residuals_objective)

Feature	Slope	Bias	Error
RM	6.291629576915702	22.05212087319519	133.7794032546542
AGE	-3.438411065390734	21.65921297166733	151.03484714001212
DIS	2.5551545781676497	21.678789870507693	159.77889985929806
LSTAT	-6.471339358234261	21.722120376842042	110.12105862612323

Result (norm_infinity_residuals_objective)

Feature	Slope	Bias	Error
RM	0.06674438316128617	27.59067654313202	22.534329364891306
AGE	-0.019176524574052944	27.521685752898595	22.500256092442566
DIS	-1.7586105770227418	25.74525359989529	22.72978282229613
LSTAT	-8.322477557393837	26.342069517917935	20.014586247213654

Alternate Search

Results (L0_norm)

Feature	Slope	Bias	Error
RM	-25.0	-25.0	467
AGE	-25.0	-25.0	467
DIS	-25.0	-25.0	467
LSTAT	-25.0	-25.0	467

Results (L1_norm)

Feature	Slope	Bias	Error
RM	6.313131313131315	22.47474747474748	1949.132044043684
AGE	-3.28282828282828	20.45454545454546	2237.40297752675
DIS	2.7777777777777786	20.45454545454546	2428.0044009249723
LSTAT	-5.808080808080806	20.45454545454546	1665.3233099595159

Results (L2_norm)

Feature	Slope	Bias	Error
RM	6.313131313131315	21.969696969696976	133.79249035553278
AGE	-3.28282828282828	21.46464646464647	151.13119749132878
DIS	2.7777777777777786	21.46464646464647	159.91037042133277
LSTAT	-6.313131313131311	21.969696969696976	110.28954982344055

Results (infinity_norm)

Feature	Slope	Bias	Error
RM	-0.2525252525252526	25.0	25.43328533314684
AGE	1.262626262626263	25.0	24.05597911915741
DIS	-2.27272727272727	25.0	23.029224685394066
LSTAT	-9.343434343434343	25.0	20.909709889361

Discussion

Considering features separately and then averaging the predictions is less effective than applying linear regression with all features at once because it disregards the relationships and interactions among the features.

1. Lack of Feature Interaction:
   Linear regression optimizes all coefficients simultaneously, capturing the combined effect of multiple features on the target variable. When features are considered independently, their interactions are ignored, leading to suboptimal predictions.
2. Bias in Coefficients:
   By training models on individual features separately, the coefficients are determined without considering the contributions of other features. This can result in biased estimates since real-world relationships between features and the target variable are often interconnected.
3. Averaging Dilutes Accuracy:
   When predictions from individual features are averaged, the model essentially assumes equal importance for all features, which is not always the case. In contrast, linear regression assigns weights to each feature based on their contribution to minimizing the overall error.
4. Redundancy:
   Some features may provide overlapping or redundant information. Training on all features together allows the model to appropriately allocate weights, reducing overemphasis on any one feature.
5. Minimization of Error:
   Linear regression works by minimizing a single global error function (e.g., norm-2 or MSE) across all features simultaneously. When features are treated separately, errors are minimized independently, but the combined predictions may not achieve the same level of overall optimization.

In the provided results, the train and test errors for the averaged predictions are higher (53.63) compared to the test error of 46.17 achieved when all features were used together. This highlights that applying linear regression on the full set of features yields better accuracy because it takes into account their combined effects and optimizes the error function globally.

The error in the infinity norm is often less than the error in other norms, due to the way each norm measures error. This norm aims to minimize the worst-case error rather than the sum or average of errors. As a result, it ignores smaller deviations in favor of controlling the largest error. By concentrating on the maximum error, the norm can lead to lower values compared to other norms because it sacrifices performance on smaller errors to minimize the peak error.