HW1: Auto MPG Analysis

Yixu Huang
ECS 171 Machine Learning, UC Davis

***Abstract*—Thid report contains the results and analysis for ECS 171 Homework Set 1, focusing on predicting and classifying automobile fuel efficiency (MPG) using the Auto MPG dataset from the UCI Machine Learning Repository.**

I. Question 1

I use quartile-based equal binning to categorize MPG values into four classes. The thresholds are determined by calculating the 25th (Q1), 50th (Q2), and 75th (Q3) percentiles.

Here are the thresholds I found:

Table I
MPG Category Thresholds

Category	Threshold Range
Low	MPG $\leq$ 17.00
Medium	17.00 $<$ MPG $\leq$ 22.75
High	22.75 $<$ MPG $\leq$ 29.00
Very High	MPG $>$ 29.00

II. Question 2

In order to determine the MPG category characteristic most distinctly, I began by creating a detailed 7×7 scatterplot matrix (Figure 2), which shows all 49 pairwise feature relationships. Every subplot illustrates the distribution of the 392 samples, which are color-coded according to their MPG category (blue: Low, green: Medium, orange: High, red: Very High). Through this visualization, we can easily observe the clustering patterns and class separability across various combinations of features.

To rank feature pairs, I calculated a separability score defined as the ratio of between-category variance to within-category variance. It shows how clearly the categories are separated compared to how spread out the data within each category is. Higher scores mean categories are more distinct and form tighter groups. Table III lists the top 10 most informative feature pairs. The pair **cylinders vs weight** had the highest score of 0.820020, meaning it shows the clearest separation among the four MPG categories. Figure 1 points out this best pair, showing little overlap between categories and clear grouping patterns.

To rank feature pairs, I calculated a separability score defined as the ratio of between-category variance to within-category variance. It measures how well-separated the category centroids are relative to the spread within each category—higher scores indicate clearer boundaries and tighter clusters. Table III presents the top 10 most informative feature pairs. The pair **cylinders vs weight** achieved the highest separability score of 0.820020, demonstrating the most distinct visual separation among the four MPG categories. Figure 1 highlights this optimal pair, showing minimal overlap between categories and well-defined clustering patterns.

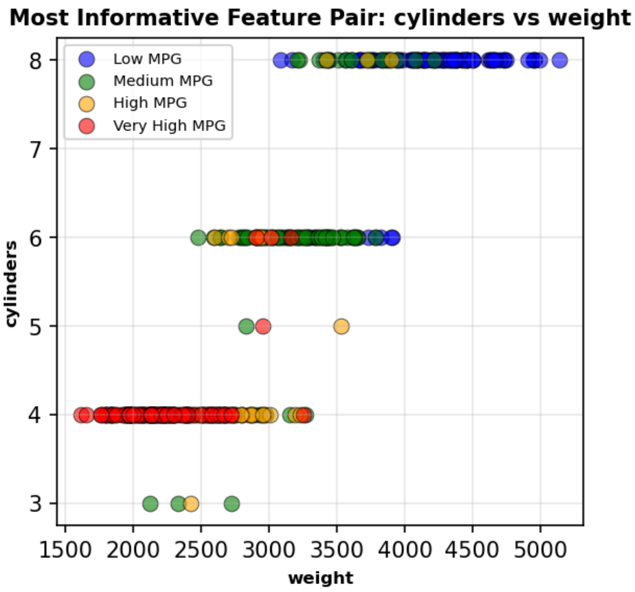

Figure 1. Most Informative Feature Pair: cylinders vs weight. The four MPG categories show clear separation with minimal overlap, validating the highest separability score.

Table II
Top 10 Most Informative Feature Pairs Ranked by Separability Score

Rank	Feature 1	Feature 2	Separability Score
1	cylinders	weight	0.820020
2	weight	origin	0.820019
3	weight	acceleration	0.819994
4	weight	model year	0.819982
5	displacement	weight	0.819425
6	horsepower	weight	0.819321
7	cylinders	displacement	0.781493
8	displacement	origin	0.781410
9	displacement	acceleration	0.779918
10	displacement	model year	0.779179

III. Question 3 & 4

IV. Questions 3 & 4: Single-Variable Polynomial Regression

I implemented a custom polynomial regression solver using the Ordinary Least Squares (OLS) estimator. The `SinglePolyRegression` class creates polynomial feature matrices $[1, x, x^2, ..., x^d]$ for degree d and computes coeffi-

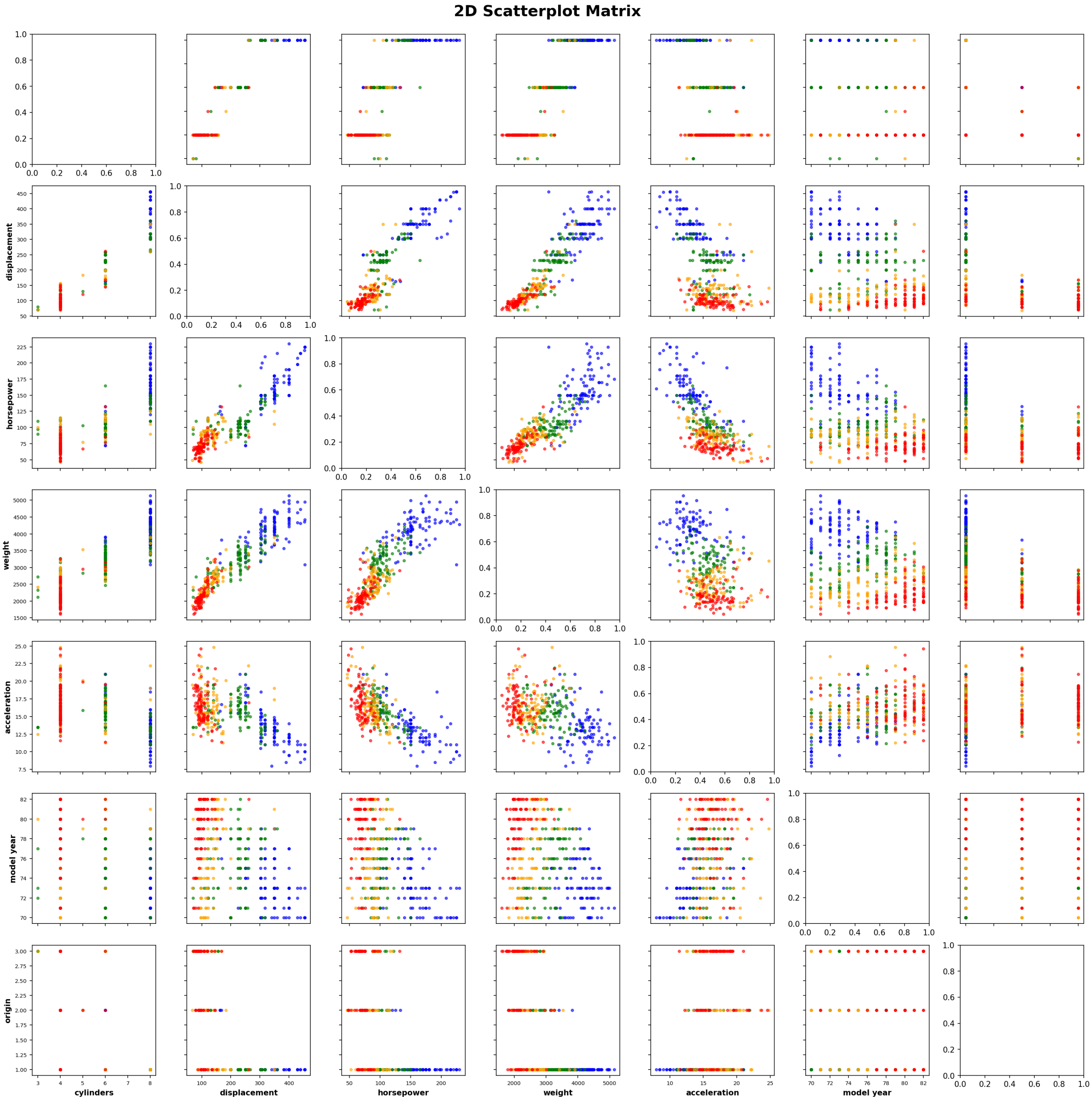

Figure 2. 2D Scatterplot Matrix of all feature pairs. Each point is colored by its MPG category, revealing varying degrees of separability across different feature combinations.

cients using the closed-form solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

The dataset was split into 292 training samples and 100 testing samples. For each of the 7 features (cylinders, displacement, horsepower, weight, acceleration, model year, origin), I trained polynomial models of degrees 0-3 to predict MPG. Tables IV and V present the mean squared errors for all feature-degree combinations.

A. Key Findings

The analysis reveals that **horsepower with degree 3 polynomial** achieved the lowest testing MSE of 59.8984, making it the most informative single feature for MPG prediction. Fig. 3 illustrates the polynomial fits for horsepower, showing progressive improvement from linear to cubic models.

Table III
TOP 10 MOST INFORMATIVE FEATURE PAIRS RANKED BY SEPARABILITY SCORE

Rank	Feature 1	Feature 2	Separability Score
1	cylinders	weight	0.820020
2	weight	origin	0.820019
3	weight	acceleration	0.819994
4	weight	model year	0.819982
5	displacement	weight	0.819425
6	horsepower	weight	0.819321
7	cylinders	displacement	0.781493
8	displacement	origin	0.781410
9	displacement	acceleration	0.779918
10	displacement	model year	0.779179

Table IV
TRAINING MSE FOR POLYNOMIAL REGRESSION (DEGREES 0-3)

Feature	Deg 0	Deg 1	Deg 2	Deg 3
cylinders	38.6153	12.4484	12.2720	10.9563
displacement	38.6153	10.7575	8.9300	8.7831
horsepower	38.6153	13.8179	10.3748	10.3495
weight	38.6153	8.2439	6.5901	6.3667
acceleration	38.6153	30.0352	29.3196	29.0590
model year	38.6153	36.0900	36.0899	35.6222
origin	38.6153	24.2852	23.2419	186000.75*

*Severe overfitting for degree 3

Table V
TESTING MSE FOR POLYNOMIAL REGRESSION (DEGREES 0-3)

Feature	Deg 0	Deg 1	Deg 2	Deg 3
cylinders	156.7484	74.6455	74.0110	68.8104
displacement	156.7484	70.5536	65.0830	65.6795
horsepower	156.7484	73.4363	60.1077	**59.8984**
weight	156.7484	67.4849	65.8921	67.9634
acceleration	156.7484	131.5734	131.4460	136.8197
model year	156.7484	88.5331	88.1441	288.1421
origin	156.7484	112.0918	113.8325	141000.38*

*Severe overfitting for degree 3

V. QUESTION 5

I extended the single-variable polynomial regression to handle all 7 features simultaneously. The key modification to the `MultiPolyRegression` class is in the `_create_poly_features` method, which can construct feature matrices for multivariate inputs:

- **Degree 0**: Constant term only (1 coefficient)
- **Degree 1**: Constant + all linear terms $[1, x_1, x_2, ..., x_7]$ (8 coefficients)
- **Degree 2**: Constant + linear terms + quadratic terms $[1, x_1, ..., x_7, x_1^2, ..., x_7^2]$ (15 coefficients)

I have to emphasis that our degree 2 implementation includes only pure quadratic terms (x_i^2), not interaction terms ($x_i x_j$), resulting in 15 total coefficients rather than the full 36 terms of a complete second-order polynomial.

Using the same 292/100 train/test split as Question 4, we trained models with degrees 0, 1, and 2 on all 7 features simultaneously. Table VI presents the results.

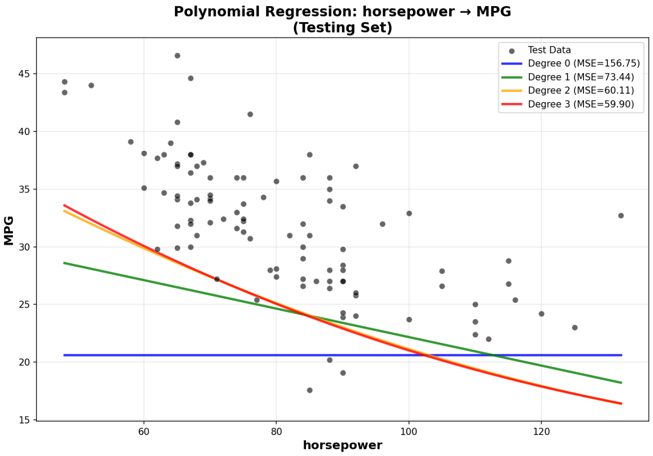

Figure 3. Polynomial regression fits (degrees 0-3) for horsepower vs MPG on the testing set. The degree 3 polynomial achieves the best fit with MSE = 59.90.

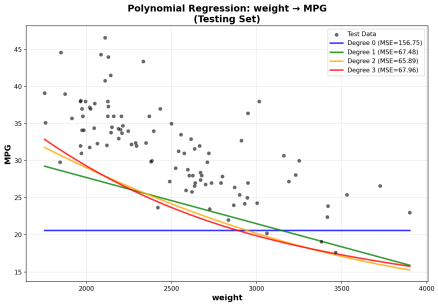

Figure 4. Weight vs MPG polynomial fits (Testing MSE: 65.89 for degree 2).

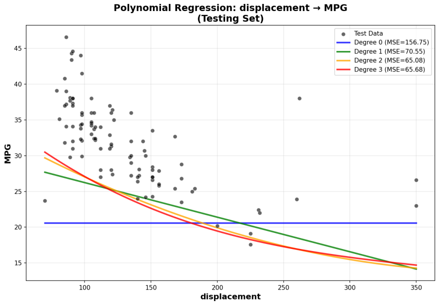

Figure 5. Displacement vs MPG polynomial fits (Testing MSE: 65.08 for degree 2).

Table VI
MULTIVARIATE POLYNOMIAL REGRESSION RESULTS

Degree	Coefficients	Train MSE	Test MSE
0	1	38.6153	156.7484
1	8	6.7502	36.9744
2	15	4.2448	**19.5793**

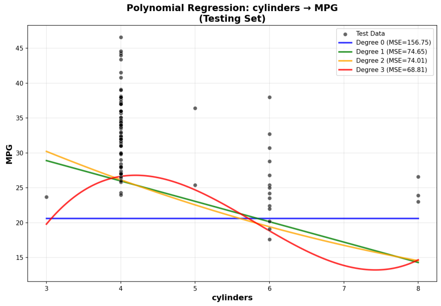

Figure 6. Cylinders vs MPG polynomial fits (Testing MSE: 68.81 for degree 3).

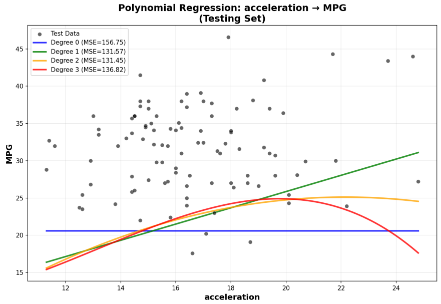

Figure 7. Acceleration vs MPG polynomial fits (Testing MSE: 131.45 for degree 2).

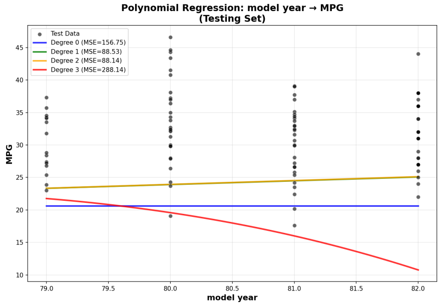

Figure 8. Model year vs MPG polynomial fits (Testing MSE: 88.14 for degree 2).

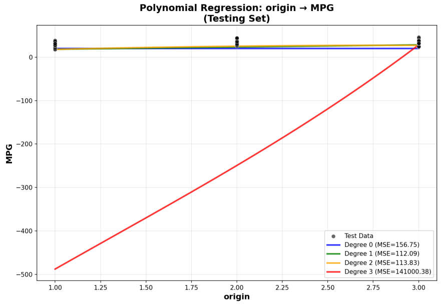

Figure 9. Origin vs MPG polynomial fits showing severe overfitting at degree 3.

A. Analysis

The second-order multivariate polynomial achieved the best performance with a testing MSE of 19.5793, representing a **67% reduction** compared to the best single-feature model (horsepower degree 3: MSE = 59.8984). This substantial improvement demonstrates that combining multiple features simultaneously instead of relying on individual features is more applicable and accurate.

Table VII
LOGISTIC REGRESSION PERFORMANCE (WITHOUT NORMALIZATION)

Dataset	Precision (Macro)	Accuracy
Training	0.7940	–
Testing	0.7724	0.76

Table VIII
CLASSIFICATION REPORT (WITHOUT NORMALIZATION, TESTING SET)

Category	Precision	Recall	F1-Score	Support
Low	0.92	0.92	0.92	25
Medium	0.58	0.79	0.67	19
High	0.59	0.67	0.63	24
Very High	1.00	0.69	0.81	32
Macro Avg	0.77	0.77	0.76	100
Weighted Avg	0.80	0.76	0.77	100

Table IX
PERFORMANCE COMPARISON: NORMALIZED VS UNNORMALIZED

Model	Train Precision	Test Precision	Converged
Unnormalized	0.7940	**0.7724**	No
Normalized	0.7730	0.7417	Yes

VI. QUESTIONS 6 & 7

I applied multiclass logistic regression to classify cars into the four MPG categories defined in Question 1. The dataset was shuffled (random_state=42) before splitting to improve category balance in the test set. I used scikit-learn's `LogisticRegression` with the default LBFGS solver and evaluated performance using macro-averaged precision.

A. Question 6: Without Normalization

Training logistic regression directly on the raw features with `max_iter=100000` still failed to converge, triggering a convergence warning from scikit-learn. Despite this, the model achieved reasonable classification performance as shown in Table VII.

B. Question 7: With MinMax Normalization

I applied MinMaxScaler to normalize all features to the [0,1] range before training. In practice, I found that this scaling resolved the convergence issue, allowing the model to converge with `max_iter=1000`. However, the normalized model showed **lower performance** compared to the unnormalized version.

C. Analysis

Counterintuitively, normalization **decreased** testing precision from 0.7724 to 0.7417, despite improving convergence behav-

Table X
CLASSIFICATION REPORT (WITH NORMALIZATION, TESTING SET)

Category	Precision	Recall	F1-Score	Support
Low	0.92	0.88	0.90	25
Medium	0.70	0.74	0.72	19
High	0.47	0.79	0.59	24
Very High	0.88	0.44	0.58	32
Macro Avg	0.74	0.71	0.70	100
Weighted Avg	0.76	0.69	0.69	100

ior. Key observations:

- The model without normalization achieves better precision on Low (0.92) and Very High (1.00) categories while normalized model shows more balanced but overall lower recall, particularly struggling with Very High (0.44 vs 0.69)
- The slow convergence in the case without normalization does not necessarily indicate poor performance—the model found a good local optimum despite non-convergence
- Feature scaling may have inadvertently reduced the discriminative power of naturally high-variance features like weight and displacement

VII. QUESTION 8

I evaluated the trained models by predicting the MPG and category for a hypothetical 1981 USA vehicle with the following specifications:

Table XI
HYPOTHETICAL VEHICLE SPECIFICATIONS (1981 USA MODEL)

Feature	Value
Cylinders	4
Displacement	400 cc
Horsepower	150 hp
Weight	3500 lb
Acceleration	8 m/s^2
Model Year	81
Origin	1 (USA)

I loaded the saved second-order multivariate polynomial regression model (Question 5) and the logistic regression classifier without normalization (Question 6) to make predictions.

A. Prediction Results

Table XII
MPG PREDICTIONS FOR HYPOTHETICAL VEHICLE

Method	Predicted MPG	Predicted Category
Polynomial Regression (Degree 2)	21.13	Medium
Logistic Regression (Direct)	–	Medium

Both approaches consistently predict the **Medium MPG** category for this vehicle. The polynomial regression model estimates a continuous MPG value of 21.13, which falls within the Medium range ($17.00 < \text{MPG} \leq 22.75$). The logistic regression classifier directly predicts the Medium category with high confidence.

Table XIII
CLASS PROBABILITY DISTRIBUTION FROM LOGISTIC REGRESSION

Category	Probability
Low	0.1974
Medium	**0.8013**
High	0.0013
Very High	0.0000

B. Analysis

The logistic regression model shows 80.13% confidence in the Medium category, with 19.74% probability assigned to Low MPG.

The predicted MPG of 21.13 is reasonable for a 1981 USA vehicle with these characteristics—moderately efficient but not high-performing due to its weight and engine power. The strong agreement between the regression-based and classification-based predictions validates both models' reliability.

REFERENCES