Machine Learning

Shahid Beheshti University
Spring 2025
Linear Regression with Uber and Lyft Ride Data

Assignment 2

1 Theoretical

Answer the following questions to deepen your understanding of regression techniques:

1. Why is linear regression called a linear model, and how can non-linearity be introduced to capture more complex patterns?
(Think about what "linear in parameters" means and why it's limiting.)
(Give examples like polynomial features, $\sin(x)$*, or activation functions in neural networks.)*

2. If multiple independent variables are highly correlated in a regression model, what problems might arise in output interpretation and model accuracy during training process? How can these problems be detected and resolved? (how can it effect on analyze feature importance)

3. Why is feature normalization important before applying gradient descent? Mathematically justify your answer.

4. Discuss the limitations of using Mean Squared Error (MSE) as a cost function in the presence of outliers. Propose and justify an alternative.

5. What are the conditions under which Stochastic Gradient Descent can converge to a global minimum even when Batch Gradient Descent gets stuck in a local minimum?

6. Why might we apply a logarithmic transformation to a predictor variable in a regression model? Discuss the advantages of this approach, and propose alternative transformation methods (e.g., square root, Box-Cox, inverse) along with scenarios where each method would be preferred.

7. Explain the intuition behind Ridge Regression and Lasso Regression. How do these models modify the cost function compared to standard Linear Regression, and what are the benefits of each?

8. Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter α or reduce it?

9. What is Locally Weighted Linear Regression (LWR)? How does it differ from traditional global linear regression? Discuss its benefits and limitations, especially in terms of complexity and interpretability.

10. Search and describe about ElasticNet Regression. Explain how it combines ideas from other models you've studied and when it might be a better choice than Ridge or Lasso alone.

2 Practical

In this assignment, you will explore and apply linear regression techniques to model and predict ride prices using a real-world dataset from Uber and Lyft services in Boston, MA. The dataset contains hundreds of thousands of ride entries with various contextual features. Your goal is to build and evaluate regression models that estimate ride prices accurately and to investigate which types of linear regression approaches yield the best results.

Dataset Overview

This dataset contains a large number of ride records, each with a diverse set of attributes. Rather than relying on predefined features, you are expected to explore the dataset thoroughly to determine which inputs are most informative for predicting ride prices. Use both statistical techniques and domain intuition to guide your feature selection. This step is critical to building effective models and should be justified clearly in your report.

Assignment Tasks

1. Data Preprocessing and Exploratory Analysis

Begin by cleaning the dataset. This includes handling missing values, correcting or removing outliers, and converting categorical variables into suitable numerical representations. Perform exploratory data analysis (EDA) to understand distributions, uncover patterns, and identify any anomalies.

Focus on discovering which features have the strongest relationship with ride prices. You should use statistical methods and visual tools to support your insights. Consider transformations or feature combinations if they help uncover linear or polynomial trends. Also you should consider normalizing or standardizing features

2. Baseline Linear Regression

Split your dataset into training and testing sets (e.g., 80% training, 20% testing). Train a standard Linear Regression model to predict the ride price based on your selected features.

Evaluate the model using the following metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Coefficient of Determination (R^2)

Provide visualizations such as residual plots and scatter plots comparing predicted and actual prices. Analyze the coefficients to interpret which variables have the most significant impact on price.

3. Model Variants and Nonlinear Extensions

Apply regularized linear models such as Ridge (L2) and Lasso (L1) regression to improve model generalization and reduce overfitting. Use cross-validation to find optimal regularization parameters and compare results with the baseline model.

In addition, explore polynomial regression by introducing higher-degree terms (e.g., degree 2 or 3) to capture nonlinear relationships in the data. Use this transformation globally or selectively for specific features where it proves beneficial.

Discuss trade-offs in terms of model complexity, performance, and interpretability. Support your findings with performance metrics and appropriate plots.

4. Model Enhancement through Data Preparation

Once your models are built, explore additional techniques to improve performance using feature engineering and data transformations. Examples may include:

- Creating interaction features between variables
- Binning or encoding categorical values differently
- Identifying and filtering noisy or low-quality samples

Explain how these changes impact model performance and whether they improve interpretability or generalization. You may apply them to your best-performing model or test them across multiple models for comparison.

Deliverables

Your submission should include:

- A PDF file including your answers to theoretical questions
- Your notebook including all outputs
- A clean and well-documented PDF report including:

- A summary of preprocessing and EDA steps, with justifications for feature selection
- Training and evaluation results for Linear, Ridge, Lasso, and Polynomial Regression
- Visualizations to support your analysis (e.g., error curves, scatter plots, coefficient paths)
- A conclusion summarizing which approach performed best and why

Bonus (Optional)

As a bonus task, analyze how model performance changes over time by segmenting the data into different days or hours and retraining your models. You may also experiment with Applying log, square root, or Box-Cox transformations to skewed features.