Texas Tech University
Department of Computer Science

Course Name: Introduction to Artificial Intelligence **Number:** CS3368 **Period:** Fall 2025
Instructor: Dr. Juan Carlos Rojas **Email:** Juan-Carlos.Rojas@ttu.edu

Homework 3

Submit your programs packed together into a single ZIP file, and a single report document in Word or PDF that shows your programs working. For the report include screen shots, or copy-paste of the terminal shell that shows the commands used and the output produced.

Problem 1

This problem focuses on polynomial regression, a linear regression approach that adds second-order terms to help capture non-linear relationships between the input features and the target.

General Preparation

Load the vehicle dataset (vehicle_clean2.csv), expand the categorial columns and trim it to keep only the following 3 features: “price”, “year”, “odometer”. You can drop all remaining columns.

Use “price” as labels and the remaining 2 columns as data. Shuffle and split the data using a random state of 2025.

Standardize the scale of the training and test data.

Part 1: Baseline

Train a multivariate linear regression model using both input columns. Compute the RMSE against both the training and test subsets.

Comment on the relative importance of each of the two features, as measured by the absolute value of their coefficients.

Comment on the level of fitting (underfitting, overfitting, or balanced fit).

Part 2: Add Square Terms

Add new columns derived from the existing ones corresponding to their values squared (“year_squared”, “odometer_squared”). This needs to be done to both the training and test data.

Re-standardize all four input columns before training.

Train a new multivariate linear regression model using all 4 input features. Compute the RMSEs against both the training and test data.

Comment on the relative importance of each of the four features, as measured by the absolute value of their coefficients.

Comment on the level of fitting (underfitting, overfitting, or balanced fit).

Comment on the effect of adding square terms to the prediction.

Explain: what did these terms express in the model?

Part 3: Add Cross Terms

Add one column corresponding to the product of year*odometer. You should now have 5 input features.

Re-standardize all five input columns before training.

Train a new multivariate linear regression model using all 5 input features. Compute the RMSEs.

Comment on the relative importance of each of the five features, as measured by the absolute value of their coefficients.

Comment on the level of fitting (underfitting, overfitting, or balanced fit).

Comment on the effect of adding cross terms to the prediction. Does the effect of year on price depend on mileage (or vice versa)?

Explain: what did this term express in the model?

Part 4: Conclusions

Compare the baseline, squared, and cross-term models. Which one fits best on training and test data? Which one would you trust most for predicting unseen vehicles? Why?

Would it make sense to add square and cross terms of one-hot encoded categorical columns? Why?