Programming Questions

Scenario

You are a member of a data science team at the Australian Bureau of Meteorology. Your task is to develop and analyze predictive models for daily rainfall (the Rainfall column) in a selected city using the weatherAUS dataset. You must apply advanced techniques in linear, probabilistic, and polynomial regression, perform in-depth data analysis, and justify your modeling choices both numerically and conceptually.

Data Preparation & Exploratory Analysis

1. Data Selection and Cleaning

- Select data for a single city (e.g., Sydney) from weatherAUS.csv.
- For each feature you intend to use for modeling, report the number of missing values.
- Propose and implement a smart strategy for handling missing values (such as deletion or imputation).
  If you use different strategies for different features, clearly explain your reasoning for each choice.
- Identify and visualize outliers in Rainfall and at least one major feature (e.g., MaxTemp) using boxplot and/or z-score.
  After removing outliers, compare and discuss summary statistics (mean, median, standard deviation) and show visualizations before and after removal.
- Feature Selection:Select at least 4–5 relevant features for prediction (e.g., MinTemp, MaxTemp, Humidity3pm, WindSpeed3pm, Pressure3pm). Justify your choices using correlation analysis, intuition, or data visualization.
- Randomly split your cleaned dataset into 80% train and 20% test.
  Use a random seed (and specify it in your report) to ensure reproducibility.

Part I: Linear Regression (Multivariate)

2. Model Building and Interpretation

- Implement multivariate linear regression using the closed-form solution (Normal Equation).
- Explicitly construct the design matrix (including a bias/intercept term if used) and calculate the regression weights.
- Report the estimated weights and interpret which features most strongly affect rainfall.
- Calculate and report training and test MSE.Plot predicted vs actual rainfall for the test set (scatter plot).
- Ablation Analysis:Remove one important feature (e.g., Humidity3pm) and repeat the training. Discuss and explain the observed changes in model performance and weight values.
- Challenge:Create a new feature with random values (e.g., drawn from a standard normal distribution) and include it in your model as an additional predictor. Analyze its impact on weights and test MSE.

Part II: Polynomial (Nonlinear) Regression

3. Nonlinear Modeling and Overfitting Analysis

- Select one major feature (e.g., MaxTemp) and fit polynomial regression models of degrees 2, 3, and 5 (using the closed-form solution for each). Explicitly show how you construct the design matrix for each degree.
- For each degree, plot the predicted curve together with the actual data points for both train and test sets.
- Compute and compare training and test MSE for all degrees.
- Outlier Robustness:Create an artificial data point with an unusually high or low Rainfall value in the test set and show its effect. Evaluate and visualize its impact on prediction curves and MSE for each polynomial degree.
- Based on your results, explain which models show underfitting or overfitting and why.

Part III: Probabilistic Linear Regression (MLE Approach)

4. Uncertainty Modeling and Parameter Stability

- Assume Gaussian noise for the labels (Rainfall). Derive and implement the maximum likelihood estimation (MLE) of regression parameters. Show both the mathematical derivation (demonstrating equivalence with least squares) and numerical implementation.
- Calculate and report log-likelihood for the train and test sets.
- Add artificial Gaussian noise to the labels (Rainfall) at three different variance levels.For each noise level, train the model 10 times (using different random seeds). Report and interpret the mean and variance of the regression parameters.
- Analyze and discuss:
    - How does increased noise affect prediction stability and confidence in parameter estimates?
    - Why is a probabilistic interpretation essential for real-world decision-making (e.g., flood warning or agricultural planning)?

Part IV: Synthesis & Advanced Discussion

5. Comparative and Practical Insights

- Compare the three approaches (linear, polynomial, probabilistic) for the following scenarios:
    - Clean data
    - High-noise data
    - Presence of outliers
- For a real application (e.g., deciding on irrigation or flood alert), which model would you choose and why?
  Support your answer with quantitative results and conceptual reasoning.
- Advanced Challenge:
  Propose and implement one method for improving robustness and model performance (e.g., outlier removal, simple regularization, or locally weighted