BAX 452 HW 1

Course Instructor : Dr. Rahul Makhijani
Due Date : 17^{th} January 2025

Please note: You are supposed to type up the solutions and submit it on canvas by **17th January 11:59 pm PST**. Refer to the course syllabus regarding late HW guidelines.

Concepts to be gained from this HW

This homework will revisit linear algebra concepts such as positive definiteness, rank, and eigenvalues. Additionally, it aims to introduce students to the exploration and preprocessing of machine learning datasets and review concepts from linear regression. you will also use gradient descent to find the minima of a function

Linear Algebra Review 10 pts

A positive semidefinite (PSD) matrix is a symmetric matrix s.t $v^T Mv \geq 0$ for all $v \in R^n$.

1. Prove that a PSD matrix should have non negative eigenvalues
2. True or False **Note:** If the answer is true provide a proof. For false statement a counterexample is sufficient

   Correlation Matrix - A correlation matrix C for a random vector $X \in R^n$ can be defined as having each entry $C_{ij} = \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$

   Hint 1 A correlation matrix is a positive semidefinite matrix.

   (a) Can correlation be used for non-linear relationships ? if your answer is false provide a counterexample where two dependedent variables have zero correlation .

   (b) Does correlation imply causation ? If False provide a counter example. in this case, you can even provide a qualitative answer in this case,

   (c) A correlation matrix is unaffected by centering or sclaing the original random vector X.

Vector Calculus 10 pts

1. Let $f(x) = \frac{1}{2}x^T Ax + b^T x$ where A is a symmetric matrix and b is a vector $\in R^n$. What is $\nabla f(x)$ What is $\nabla^2 f(x)$?

2. Consider the function $f(u, \lambda) = u^T X^T X u - \lambda u^T u$. What is the property of u and λ with respect to X at which the $\frac{\partial f}{\partial u} = 0$

Linear Invariance of Regression Coefficients 30 pts

Assume a linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon$, i = 1, ..., n, and assume the x_i are fixed (non-random), and that the ϵ_i are uncorrelated, gaussian iid variables each having mean 0 and variance σ^2. Let $\tilde{x_i} = \frac{x_i - \bar{x}}{s_x}$, i = 1, ..., n be standardized versions of the x_i , where $\bar{x}$ is the sample mean, and s_x is the sample standard deviation, with n in the denominator. Suppose that we fit a simple linear regression model of y_i on $\tilde{x_i}$ by least squares.

1. What are the least squares estimates α_0 and α_1 (intercept and slope) (in their simplest form) for these transformed $\tilde{x_i}$?

2. Derive the relationship between this slope estimate and the sample correlation coefficient between y_i and x_i

3. What are the sampling variances of each of these estimates, and their sampling covariance?

4. Can you use these estimates to obtain LS estimates for the linear regression model with x_i and y_i? How?

5. What if each x_i was multiplied by 100 before computing $\hat{\alpha_0}$ and $\hat{\alpha_1}$. Would the estimates change? What are the practical implications?

6. Provide some examples where the assumption of independence and correlation of ϵ_i and x_i is not valid.

Probability Review 10 points

Consider a medical test for a rare disease which affects 2% of the population. The test has a sensitivity of 98% (probability of a positive test given the disease) and a specificity of 95% (probability of a negative test given no disease).

- Using Bayes' Theorem, calculate the probability that a person has the disease given that they tested positive.
- Perform a simulation to estimate the same probability by generating a large random sample of the population. Simulate whether each individual has the disease and the corresponding test result, then calculate the proportion of people who actually have the disease among those who tested positive.
- Compare the results from your calculation and simulation. Discuss any discrepancies and explore how changing the prevalence of the disease impacts the posterior probability, both analytically and through simulation

Multiple Linear Regression 30 pts

We will apply multiple linear regression to gain insights into California housing prices.

For Python users, s the California Housing dataset can be obtained from the following resources : -

1. **scikit-learn:** You can load it directly using the `datasets` module in scikit-learn.

   ```
   sklearn.datasets import fetch_california_housing
   housing = fetch_california_housing()
   ```

2. UCI Machine Learning Repository
3. Kaggle website

Instructions

1. **Data Exploration: 5 pts**
   - Load the California Housing dataset. Provide a brief description of the dataset, including the number of observations, features, and their types.
   - Perform exploratory data analysis (EDA) to understand the distribution of each feature and identify any missing values or outliers.
2. **Preprocessing: 5 pts**
   - Handle any missing data appropriately. Justify your method of imputing or removing missing values.
   - Normalize or standardize the features if necessary. Explain your choice.
3. **Model Building: 5 pts**
   - Construct a multiple linear regression model using `MEDV` (median house value in $1000s) as the dependent variable and the other features as independent variables.
   - Split the dataset into training and test sets (e.g., 70-30 split) and justify your choice of splitting ratio.
4. **Model Evaluation: 5 pts**
   - Fit the model on the training data and evaluate its performance on both the training and test datasets using metrics such as Mean Squared Error (MSE) and R-squared.
   - Analyze the statistical significance of each coefficient and discuss which features seem to be the most important predictors of housing prices.
5. **Assumption Checking: 5 pts**
   - Check the assumptions of linear regression, including linearity, independence, homoscedasticity, and normality of residuals. Provide plots and interpretations to support your analysis.

6. **Improving the Model: 5 pts**
    - Based on your findings, suggest and apply methods to improve the model, such as feature engineering, interaction terms, or polynomial regression.
    - Re-evaluate the improved model and compare its performance with the initial model.

7. **Conclusion:**
    - Summarize your findings and discuss the practical implications of your analysis in understanding the factors affecting housing prices in the California area.

Submission

Submit your code, analysis, and results in a well-documented Jupyter Notebook or a similar format. Ensure that your submission includes visualizations, explanations, and interpretations for each step of the process.

Gradient Descent 10 pts

Consider the function:

$$f(x, y) = x^2y^2 + x^2 + 2x + 2 + y^2 + 2y$$

1. Compute the gradient of the function, denoted as $\nabla f(x, y)$.
2. Implement the gradient descent algorithm to find the local minima of this function.
3. Use an appropriate learning rate and number of iterations to ensure convergence.
4. Report the values of x and y at the local minima and the corresponding value of the function $f(x, y)$.
5. Plot the function values over different iterations of the minima value.
6. Discuss the choice of learning rate and any challenges faced during implementation.