DATA 220 Lab2 (Python)

Q1: Recommender System (20 Points)

1. Download m1.zip file from the link (https://grouplens.org/datasets/movielens/1m/) (0 points)
2. Load the movies and ratings data. (2 points)
3. What do you mean by Singular Value Decomposition (SVD)(2 points)?
4. Explain content-based vs collaborative recommendation. (2 points)
5. Create m x u matrix with movies as row and users as column. Normalize the matrix. (2 points)
6. Perform SVD to get U, S and V. (4 points)
7. Select top 50 components from S. (2 point)
8. Get the top 50 eigenvectors using eigenvalues. (2 point)
9. Using cosine similarity, find 10 closest movies using the 50 components from SVD. (2 points)
10. Discuss results of above SVD methods. (2 point)

Q2: House Prices Prediction (50 points)

Data Exploration (10 points):

1. Start by importing the dataset and exploring its structure.
2. What are the features and the target variable? (1 point)
3. How many samples are in the dataset? Are there any missing values? (2 points)
4. Summarize the dataset. Min, max, avg, std dev, etc. stats for continuous features. (2 points)
5. Visualize the distribution of each feature (sqft_living, sqft_lot, floors, SalesPrice)(5 marks)

Linear Regression (Single Variable) (10 points):

6. Implement your own linear regression model using the "sqft_lot" feature as the independent variable and "SalePrice" as the target variable. Print coef and intercept. (5 points)
7. Calculate the sum of squared errors for your model. (1 point)
8. Plot the regression line along with the actual data points. (1 point)
9. Use the LinearRegression function from sklearn.linear_model library and compare the coef and intercept with your model. (3 points)

Linear Regression (Multivariate) (6 points):

10. Use the LinearRegression function from sklearn.linear_model library to include multiple features sqft_living, sqft_lot and print the coef and intercept. (3 points)
11. Print R-squared (R^2) score. (1 point)
12. Visualize the relationships between the selected features and SalePrice. (2 points)

Polynomial Regression (10 points):

13. Use a polynomial feature's function and implement a polynomial regression model of degree 2 for the features sqft_lot and the target variable. (4 points)
14. Print R-squared (R^2) score. (1 point)
15. Experiment with different polynomial degrees and find the best fit as per your perspective. (3 points)
16. Plot the polynomial regression curve along with the actual data points. (2 points)

RANSAC (Robust Regression) (10 points):

19. Apply RANSAC (Random Sample Consensus) to fit a robust linear regression model to the features sqft_lot and the target variable. (4 point)

20. Print coef and intercept. Visualize plot wrt inliers and outliers. (4 point)

21. Print R-squared (R^2) score with and without inliers. (2 point)

Model Evaluation (4 points):

22. Compare the results and discuss which model(s) best-predicted housing prices. (4 points)

Q3: Life Expectancy prediction (40 points)

1. Load the dataset and present the statistics of data.(1 point)
2. Identify and specify the target variable from the dataset.(1 point)
3. Categorize the columns into categorical and continuous.(1 point)
4. Identify the unique values from each column.(1 point)
5. Identify the Missing values and compute the missing values with mean, median or mode based on their categories. Also explain why and how you performed each imputation. (2 points)
6. Check for the outliers in each column using the IQR method.(1 point)
7. Impute the outliers and impute the outlier values with mean, median or mode based on their categories.(2 points)
8. Calculate summary statistics for numerical columns, such as mean, median, standard deviation, etc.(1 point)
9. Identify and perform label encoding on certain columns:(2 points)
    (a) Specify and explain on which columns you perform and why.
    (b) Explain what is label encoding and how it changes the dataset.

10. Perform data normalization on ‘Adult Mortality', 'BMI', 'GDP’ numerical columns using StandardScaler() (2 points)

11. Compute a correlation matrix and plot the correlation using a heat map and answer the following questions: (2 points)

(a) The Features which are Most Positively Correlated with target variable.
(b) The Features which are Most Negatively Correlated with target variable.

12. Drop the column ‘country’ from the dataset and split the dataset into training and testing in a 70:30 split. (2 points)

13. Build a linear regression model using the training and testing datasets and compute mean absolute error. (4 points)

14. Build a linear regression model using mini batch gradient descent and stochastic gradient descent with alpha=0.0001, learning rate='invscaling', maximum iterations =1000, batch size=32 and compute mean absolute error. (6 points)

15. Build a linear regression model using mini batch gradient descent with learning rate = 0.0001, maximum iterations =1000 and batch size=32. Manually without using any scikit learn libraries.(10 points)

16. Compare the results from each approach and also explain the difference between mini batch gradient descent and stochastic gradient descent.. (2 points)