CSE343/ECE343: Machine Learning
Assignment-1 Linear and Logistic Regression, ML in Practice
Name: Tushar Chandra
Roll no.: 2021211

1.)

a. No, if two variables exhibit a strong correlation, that doesn't necessarily mean that they will also possess a high correlation with each other. According to Simpson's Paradox, the direction or strength of a correlation can change when a third variable is considered.

Suppose two colleges, A and B, offer CS and Humanities majors with data as provided below:

College	Majors	Applications	Admitted	Admission Rate (%)
A	CS	100	40	40
A	Humanities	200	100	50
B	CS	50	30	60
B	Humanities	150	100	66.7

College A has a lower admission rate compared to College B. However, after aggregating the data:

CS: (40 + 30) / (100 + 50) = 58.3%
Humanities: (100 + 100) / (200 + 150) = 57.1%

Now, the overall admission rate is slightly higher for the CS major, even though college A had lower admission for both majors in each individual college. Combining data leads to the reversal of the original trend. This is known as Simpon's Paradox.

b. A mathematical function is categorised as a logistic function if it satisfies the following criteria:

   i. **S-shaped curve**: The function should have a S-shaped Curve.
   ii. **Bounded Range**: The functions' range should be between 0 and 1.

- **sinh(x):** It is a hyperbolic sine function and ranges from -infinity to infinity, so it's **not** a valid logistic function.
- **cosh(x):** Like sinh(x), it is also a hyperbolic cosine function and doesn't satisfy the bounded range hence **not** a logistic function.

- **tanh(x):** It satisfies the S-shape curve but ranges from -1 to 1, which doesn’t meet the bounded range criteria, but it can be rescaled at shifted to fit ranges between [0,1]. Hence, **it is** a logistic function.
- **signum(x):** It does not exhibit an S-shaped curve and is not bounded between 0 and 1, so it’s **not** a valid logistic function.

c. For very sparse datasets, the **Leave-One-Out-Cross-Validation (LOOCV)** technique is used. In LOOCV, each data point is used as the validation set, while the rest of the data is used for training. This process is repeated for every data point. Further, the results are averaged to evaluate the model’s performance.

Some benefits of LOOCV:

1. **Maximizes Data Utilization**
2. **Reduces Bias**

<u>Difference from K-Fold Cross Validation</u>:

1. **Computational Complexity:** The complexity of LOOCV is higher because of multiple iterations, whereas K-Fold CV is more efficient.
2. **Bias:** LOOCV provides an unbiased estimate of the model's performance as each data point is treated as a separate validation set. Whereas K-Fold CV can have slight bias since some of the data points might not serve as validation.
3. **Data Usage:** LOOCV uses almost all the data points for training in each iteration, whereas K-Fold CV divides the data into K subsets, allowing a balanced used of data for both training and evaluation.

d. The Least Square Regression line aims to find the best-fitting line that minimises the sum of squared differences between the observed points (yi) and predicted values (y_predicted) for the given set of x-values (xi). The slope-intercept form of the regression equation is given by:

$$\mathbf{y = mx + c}$$

m can be calculated from the formula given below:

$$\mathbf{m = \sum (xi - x^*) \times (yi - y^*) / \sum (xi - x^*)^2}$$
$$\mathbf{c = y^* - m^*x^*}$$

- x* refers to the mean of xi values
- y* refers to the mean of yi values

e. (a.) α, β, σ

The error term cannot be estimated as it represents the noise in the data that cannot be directly predicted.

f. (d.) $Y=\alpha+\beta_1x+\beta_2x_2+\varepsilon\beta_2 >0$

As temperature increases, the electric bill follows a non-linear trend, initially decreasing before rising, suggesting a quadratic relationship with a positive coefficient for the x^2 term.