Assignment 1
Python Basics and Logistic Regression

STAT3612: Statistical Machine Learningn

DUE: **Feb 16, 2025, Sunday, 11:59 PM**

Goal

The first part of this assignment is to make you familiar with basic concepts in classification and linear and logistic regression models. The second and third part is an application of logistic regression in a practical machine-learning task. You will learn basic data processing skills and how to develop a logistic regression model for classification problem by completing this assignment. **Note that you should add necessary textual descriptions to your code when required in specific questions.**

Submission

Please submit the following **two files** in Moodle for grading:

- A PDF/HTML report of your answers to all questions. You are recommended to convert the Jupyter notebook file into PDF/HTML format and submit it as the report.
- The completed Jupyter notebook file `assign1.ipynb`.

Suggestion

It is highly recommended to go over the corresponding tutorial materials thoroughly before working on this assignment.

Part 1: Conceptual Questions

Q1 Please classify these statements as either **TRUE** or **FALSE** and give necessary explanations if it is **FALSE**. **[TOTAL: 12 points]**

(a) A marketing company wants to build a model for predicting the number of customers to a shoe shop every day. Since the number of customers is an integer, this is best viewed as a classification problem. **[3 points]**

(b) A classifier is called linear if it has a linear decision boundary. **[3 points]**

(c) A model with lower bias always performs better than a model with higher bias in terms of the mean squared error on test data. **[3 points]**

(d) A softmax operator maps real-valued model scores to probabilities (0 to 1). **[3 points]**

Q2 Consider the following training data,

i	x	y
1	1.0	2.8
2	2.0	5.0
3	3.0	7.5
4	4.0	8.7
5	5.0	11.1
6	6.0	12.9

where i is and index, x is the one-dimensional input variable and y is the output.

[TOTAL: 10 points]

(a) Fit a straight line to the data $y = \theta_0 + \theta_1 x$ using linear regression. Illustrate the training data in a graph with x and y on the two axes together with your estimated line. **[5 points]**

(b) If you try to fit a 5-th order polynomial to the training data.

$$\hat{y}_\theta(x) = \sum_{l=0}^{5} \theta_l x^l = \boldsymbol{\theta}^\top \mathbf{x}, \quad \mathbf{x} = \begin{bmatrix} 1 & x & x^2 & x^3 & x^4 & x^5 \end{bmatrix}^\top$$

using the cost function $J(\theta) = \sum_{i=1}^{6} \left(y_i - \hat{y}_\theta\left(x_i\right)\right)^2$. What training error will you attain? **[5 points]**

Part 2: Python Basics and Data Visualization

Companies often face high rates of employee attrition, with staff leaving due to various reasons such as job dissatisfaction, lack of career growth opportunities, or better offers elsewhere. Manually analyzing these patterns to predict which employees are at risk of leaving can be a tedious and error-prone process, requiring significant time and resources. Fortunately, machine learning provides an efficient solution to this challenge, enabling organizations to automate the classification of employees likely to churn. By leveraging data-driven insights, businesses can proactively address employee retention and improve workplace satisfaction, just as many leading companies are doing today.

Q3 In this question, you are required to do basic data processing on the employee attrition dataset step by step.

[TOTAL: 30 points]

(a) Use Pandas library to read the file `Employee-Attrition-Classification.csv`. We will use the following 7 attributes and the attrition label: "Age", "Years at Company", "Monthly Income", "Distance from Home", "Company Tenure", "Number of Promotions", "Number of Dependents", "Attrition". Remove the attributes we do not use. And convert "Attrition" labels using "Stayed=1, Left=0". **[6 points]**

(b) Use NumPy or Pandas to compute the statistics (i.e., mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, maximum) of 7 attributes (the same as Q3 (a)). **[6 points]**

(c) Compute the Pearson correlation coefficient matrix of the same 7 attributes. You are required to use NumPy or Pandas library to implement this. **[6 points]**

(d) Standardize 7 attributes (the same attributes as Q3 (a)). The standardization means to rescale the feature such that its mean is 0 and its standard deviation is 1. You are required to use NumPy or Pandas package to implement it. **[6 points]**

(e) Divide the dataset randomly into training (80%) and testing (20%) sets, and perform the standardization as Q3 (d) again, explain in words why you should not directly use the results in Q3 (d). **[6 points]**

Q4 You are required to complete the following data visualization using matplotlib or seaborn library.

[TOTAL: 18 points]

(a) Show the boxplot of 7 attributes in Q3 (b). Note that you should plot the original attributes, which are before standardization. **[6 points]**

(b) Plot the correlation matrix you have computed in Q3 (c). You are required to set the attributes as the labels of x and y axis, and use the color to represent the correlation coefficient. **[6 points]**

(c) Plot figures to compare the distribution of 7 attributes in training and testing set (the same attributes in Q3(e)). Please plot one figure (includes two sub-figures for training and testing set) for each attribute. Note that you are recommended to use subplots function in Matplotlib to implement this. **[6 points]**

Part 3: Logistic Regression

Q5 After initially exploring the data, we want to fit a logistic regression model with "Attrition" as the output and all other variables in Q3 (e) as the input. You are recommended to use sklearn library to answer this question.

[TOTAL: 30 points]

(a) Fit the logistic regression model and print the coefficients of your logistic regression model. **[8 points]**

(b) According to your model, please briefly discuss the relative importance of input features. Which features are positively affected your output, while which features are negatively affect your output? Do you think if it is reasonable according to common sense? **[7 points]**

(c) Please report the precision, recall and F score on training and testing sets, respectively. **[7 points]**

(d) The performance might not be satisfying, could you figure out how to improve the performance of this task? For example, you can try to use more attributes or interactions between attributes in your logistic regression model. Briefly explain your modifications in words and report the precision, recall, and F score on training and testing sets, respectively. **[8 points]**