Introduction to ML
Assignment 10

Exercise on paper: Logistic regression

In this exercise, you will explore the fascinating world of logistic regression by classifying observations as **oranges** or **apples**. You and your friend both fit logistic regression models to the same data, but with different formulations. Your goal is to compare the results and interpret the coefficients.

1. Suppose you fit a logistic regression model and find that:

$$\widehat{\Pr}(Y = \text{orange} \mid X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

   Your friend, however, uses the `softmax` formulation and finds:

$$\widehat{\Pr}(Y = \text{orange} \mid X = x) = \frac{\exp(\hat{\alpha}_{\text{orange0}} + \hat{\alpha}_{\text{orange1}} x)}{\exp(\hat{\alpha}_{\text{orange0}} + \hat{\alpha}_{\text{orange1}} x) + \exp(\hat{\alpha}_{\text{apple0}} + \hat{\alpha}_{\text{apple1}} x)}$$

   (a) What is the log odds of **orange** versus **apple** in your model?

   (b) What is the log odds of **orange** versus **apple** in your friend's model?

2. Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.

3. Now, suppose that you and your friend fit the same two models on a different dataset. This time, your friend gets the coefficient estimates:

$$\hat{\alpha}_{\text{orange0}} = 1.2, \quad \hat{\alpha}_{\text{orange1}} = -2, \quad \hat{\alpha}_{\text{apple0}} = 3, \quad \hat{\alpha}_{\text{apple1}} = 0.6$$

   What are the coefficient estimates in your model?

4. Finally, suppose you apply both models from (d) to a dataset with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Note: No coding is required for this exercise. Use your knowledge of logistic regression and the relationship between the two formulations to answer the questions.

Coding exercise: Student Depression Survey

In this task, you will use Logistic Regression to predict whether a student is experiencing depression based on different factors such as academic pressure, sleep habits, and financial stress. You will go through the full machine learning process, from exploring the data to training and evaluating a model. You are also encouraged to create visualizations to better understand the data and support your analysis. This exercise is made to guide you in different analysis. Feel free to use your imagination and understanding of classification to bring interesting analysis/insights.

The dataset contains the following features:

- **Gender**: Male/Female
- **Age**: Student's age
- **City**: City where the student lives
- **Academic Pressure**: Level of academic stress (1-10)
- **Work Pressure**: Level of work-related stress (1-10)
- **CGPA**: Student's academic performance (0.0-4.0)
- **Study Satisfaction**: How satisfied the student is with their studies (1-5)
- **Job Satisfaction**: Satisfaction with a job (if applicable) (1-5)
- **Sleep Duration**: Sleep hours category (e.g., "Less than 5 hours", "5-6 hours")
- **Dietary Habits**: Eating habits (e.g., Healthy, Moderate)
- **Degree**: The degree the student is pursuing
- **Suicidal Thoughts**: Whether the student has had suicidal thoughts (Yes/No)
- **Work/Study Hours**: Hours spent working or studying daily
- **Financial Stress**: Level of financial pressure (1-10)
- **Family History of Mental Illness**: Yes/No
- **Depression**: Target variable (1 = Has depression, 0 = No depression)

Task 1 – Explore the Data

Understand the dataset and find interesting patterns :

1. Use basic pandas functions to check the data.
2. Look for missing values, outliers, and patterns in the features.
3. Create visualizations such as heatmap, histograms, bar charts, scatter plots etc. to explore relationships between different features and depression. Please, spend some time in this part. Your plots should be "insightful". Don't stop at simply creating graphs. You should also understand them and extract value from them.

Task 2 – Clean and prepare the data for modeling

1. Handle missing values (e.g., filling or removing them).
2. Convert categorical variables into numbers using encoding.
3. Normalize or standardize numerical features if needed.
4. Explain why you made certain preprocessing choices.

Task 3 – Train the Model

Train a Logistic Regression model to predict depression :

1. Create a python class to fit a logistic regression with gradient descent. Make your class modular ; each method should have a precise function (i.e. `fit`, `predict`, `errors`, `coeff` etc...).
2. Split the data into training and testing sets
3. Train your Logistic Regression model.
4. Make predictions on the test data

Task 4 – Evaluate the Model

Measure how well the model performs :

1. Calculate accuracy, precision, recall, and F1-score.
2. Create a confusion matrix to see how often the model makes correct and incorrect predictions.
3. Plot an ROC curve to analyze model performance.
4. Try to find a method to analyze predicted probabilities