Exercises and Questions for Session 4: Classification and Logistic Regression

Here are some questions to self-evaluate your knowledge about the lesson. The answers are provided at the end of the document, but I recommend trying to answer them first without looking at the answers.

Knowledge Check Questions

1. What is supervised classification in machine learning?
2. What is the difference between binary and multi-class classification?
3. What is a confusion matrix, and how is it calculated in i) the binary case, and ii) the multiclass case?
4. Define precision and recall.
5. What is logistic regression?
6. If you solve a classification problem using logistic regression, are you certain that the accuracy is maximized in the training set? Explain why.
7. How is logistic regression extended to handle multi-class problems?
8. Explain the term 'cross-entropy loss' in the context of logistic regression.
9. In a binary classification case, can you obtain a nonlinear boundary region using logistic regression? Why?
10. Describe how a decision boundary is defined in logistic regression.

Practice.

Getting the data.

The data used for these exercises will be collected from hrefhttps://www.kaggle.com/Kaggle, a platform that hosts datasets for data science projects. To obtain and understand the data:

- Visit `Kaggle.com` and create an account or log in.
- Navigate to the 'Datasets' section and search for relevant datasets, such as "Titanic" or "Iris".

- Download the dataset and read the accompanying documentation to understand the features and target variables.
- Utilize tools like Pandas in Python to explore and preprocess the data.

Exercise 1: Analyzing Socio-economic Factors in Titanic Survival

Context: The Titanic, a luxury passenger liner, tragically sank on its maiden voyage in 1912, leading to the deaths of over 1,500 passengers and crew. A dataset has been compiled that captures details of the travelers, including socio-demographic information, ticket details, and whether they survived the disaster. This dataset serves as a foundation for predictive modeling, where the goal is to link the likelihood of survival to various characteristics of the passengers.

Main Question: Were socio-economic or demographic factors significant predictors of survival on the Titanic?

Objective:

The objective of this exercise is to use logistic regression to analyze the Titanic dataset and identify key predictors of survival. By modeling these relationships, we aim to quantify the impact of factors such as class, sex, age, and fare on survival odds. To accomplish this, please follow the steps outlined below:

1. Download the Titanic dataset from Kaggle.
2. Load the data into a Python environment using Pandas.
3. Perform data preprocessing:
   - Handle missing values.
   - Convert categorical variables into numeric oif needed.
   - Identify the label to predict. What is considered the positive (1) and negative (0) class in this scenario?
4. Split the dataset into training and testing sets.
5. Train a logistic regression model on the training data.
6. Evaluate the model's performance using the following metrics:
   - Accuracy
   - Precision
   - Recall
   - F1-score
7. Discuss the implications of the findings:

   - Interpret the model coefficients to understand the influence of factors like socio-economic status, gender, and age on survival rates.
   - Analyze the precision and recall to determine if the model is reliable in predicting survival accurately.
   - Consider the balance between precision and recall, represented by the F1-score, to assess the model's overall predictive power.

8. Reflect on what the learning outcomes suggest about the historical context of the Titanic disaster.

Exercise 2: Using Patient Data to Predict Heart Disease

Context: Heart disease remains one of the leading causes of mortality worldwide, making it a critical area of medical research. The ”Heart Disease UCI” dataset on Kaggle comprises a number of attributes that can be used to predict the presence of heart disease in individuals. Attributes include age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, rest electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and more. This exercise aims to apply logistic regression to identify the most significant predictors of heart disease, thereby contributing to preventative health measures.

> **Main Question:** Can we predict the presence of heart disease based on a combination of medical, lifestyle, and demographic factors?

The objective of this exercise is to employ logistic regression to analyze the Heart Disease UCI dataset and identify key predictors of heart disease. To accomplish this, please follow the steps outlined below:

1. Access and download the ”Heart Disease UCI” dataset from Kaggle.
2. Import the dataset into a data analysis environment using Python and Pandas.
3. Perform exploratory data analysis to understand the distribution and relationships of the features:
   - Visualize the data to identify patterns and outliers.
   - Calculate summary statistics to get insights into the data's structure.
4. Conduct data preprocessing:
   - Clean the data by handling missing values and outliers if necessary.
   - Transform categorical variables using encoding techniques.
   - Normalize or standardize numerical variables to improve model performance.
5. Split the data into training and testing subsets.

6. Train a logistic regression model to predict the presence of heart disease.

7. Evaluate the model using appropriate metrics such as accuracy, precision, recall, F1-score, and the ROC curve.

8. Interpret the model results:

   - Discuss the impact of various predictors on heart disease likelihood.
   - Evaluate the model's sensitivity and specificity from the ROC curve analysis.

9. Reflect on the potential for using such models in real-world medical diagnostics and preventive healthcare.

Exercise 3: Predicting Wine Quality.

Context: The Wine Quality Dataset available on Kaggle contains data on various wines, featuring physicochemical properties and a sensory quality rating from experts. Each wine is graded with a quality score, making this dataset suitable for a multiclass classification problem. This exercise aims to predict the quality category of wines based on their physicochemical properties using a suitable classification algorithm.

Main Question: Can we accurately predict the quality category of wines based on their physicochemical properties?

Objective: The goal is to apply a multiclass classification technique to determine the wine quality category. This will involve using methods suited for handling multiple classes in the dataset and evaluating the model's ability to classify wines into the correct quality categories.

1. Download the Wine Quality Dataset from Kaggle.

2. Import the dataset into your preferred data analysis software.

3. Perform exploratory data analysis:

   - Analyze the distribution of quality ratings.
   - Visualize relationships between features and the target variable.

4. Conduct data preprocessing:

   - Handle missing values, if any.
   - Normalize or scale the features as necessary.

5. Encode the quality ratings into categorical class labels suitable for multiclass classification.

6. Split the dataset into training and testing subsets.

7. Train a classifier that is capable of handling multiple classes (e.g., Random Forest, Gradient Boosting Machines, or Multinomial Logistic Regression).

8. Evaluate the model using appropriate metrics such as accuracy, precision, recall, and F1-score for each class. Consider using a confusion matrix to visualize the performance.

9. Interpret the results and discuss the implications of the model findings in terms of wine production and quality assessment.