School of Computer Science Engineering and Technology

Course- BTech	Type- Specialization Core
Course Code: CSET 301	Course Name- AIML
Year- 2022-23	Semester- 4th
Date-	Batch-

Lab Assignment W2L1

Logistic Regression Modeling for Early Stage Diabetes Risk Prediction

Part 2.1: Getting familiar with linear algebraic functions

Tasks

Create matrix of size 10*10 with random integer numbers

Compute the following linear algebric operations on the matrix using built in functions supported in Numpy, Scipy etc.

Find inverse of the matrix and print it

Calculate dot product of the matrix with same matrix in transpose A.AT

Decompose the original matrix using eigen decomposition print the eigen values and eigen vectors

Calculate jacobian matrix

Calculate hessian matrix

Part 2.2: Logistic Regression using newton method

Logistic regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as W) to predict an output value (y). A key

difference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a continuous value.

$$\hat{y}(w,x) = \frac{1}{1+exp^{-(w_0+w_1*x_1+...+w_p*x_p)}}$$

Dataset

Original Source: http://archive.ics.uci.edu/ml/machine-learning-databases/00529/diabetes_data_upload.csv

The dataset just got released in July 2020.

Features (X)

1. Age - Values ranging from 16-90

2. Gender - Binary value (Male/Female)

3. Polyuria - Binary value (Yes/No)

4. Polydipsia - Binary value (Yes/No)

5. sudden weight loss - Binary value (Yes/No)

6. weakness - Binary value (Yes/No)

7. Polyphagia - Binary value (Yes/No)

8. Genital thrush - Binary value (Yes/No)

9. visual blurring - Binary value (Yes/No)

10. Itching - Binary value (Yes/No)

11. Irritability - Binary value (Yes/No)

12. delayed healing - Binary value (Yes/No)

13. partial paresis - Binary value (Yes/No)

14. muscle stiffness - Binary value (Yes/No)

15. Alopecia - Binary value (Yes/No)

16. Obesity - Binary value (Yes/No)

Output/Target target (Y)

1. class - Binary class (Positive/Negative)

Objective

To learn logistic regression and practice handling of both numerical and categorical features

Tasks

Download, load the data and print first 5 and last 5 rows

Transform categorical features into numerical features. Use label encoding or any other suitable preprocessing technique

Since the age feature is in larger range, age column can be normalized into smaller scale (like 0 to 1) using different methods such

as scaling, standardizing or any other suitable preprocessing technique (Example - sklearn.preprocessing.MinMaxScaler class)

Define X matrix (independent features) and y vector (target feature)

Split the dataset into 60% for training and rest 40% for testing (sklearn.model_selection.train_test_split function)

Train Logistic Regression Model on the training set (sklearn.linear_model.LogisticRegression class)

Use the trained model to predict on testing set

Print 'Accuracy' obtained on the testing dataset i.e. (sklearn.metrics.accuracy_score function)

Further fun (will not be evaluated)

Plot loss curve (Loss vs number of iterations)

Pre-process data with different feature scaling methods (i.e. scaling, normalization, standardization, etc) and observe accuracies on

both X_train and X_test

Training model on different train-test splits such as 60-40, 50-50, 70-30, 80-20, 90-10, 95-5 etc. and observe accuracies on both

X_train and X_test

Shuffling of training samples with different random seed values in the train_test_split function. Check the model error for the testing

data for each setup.

Print other classification metrics such as:

classification report (sklearn.metrics.classification_report),

confusion matrix (sklearn.metrics.confusion_matrix),

precision, recall and f1 scores (sklearn.metrics.precision_recall_fscore_support)

Helpful links

Scikit-learn documentation for logistic regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

How Logistic Regression works: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Feature Scaling: https://scikit-learn.org/stable/modules/preprocessing.html

Classification metrics in sklearn: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics