Texas State University

Assignment 2: Logistic Regression

Chichi Christine

Intro to Machine Learning CS 4347

November 1, 2019

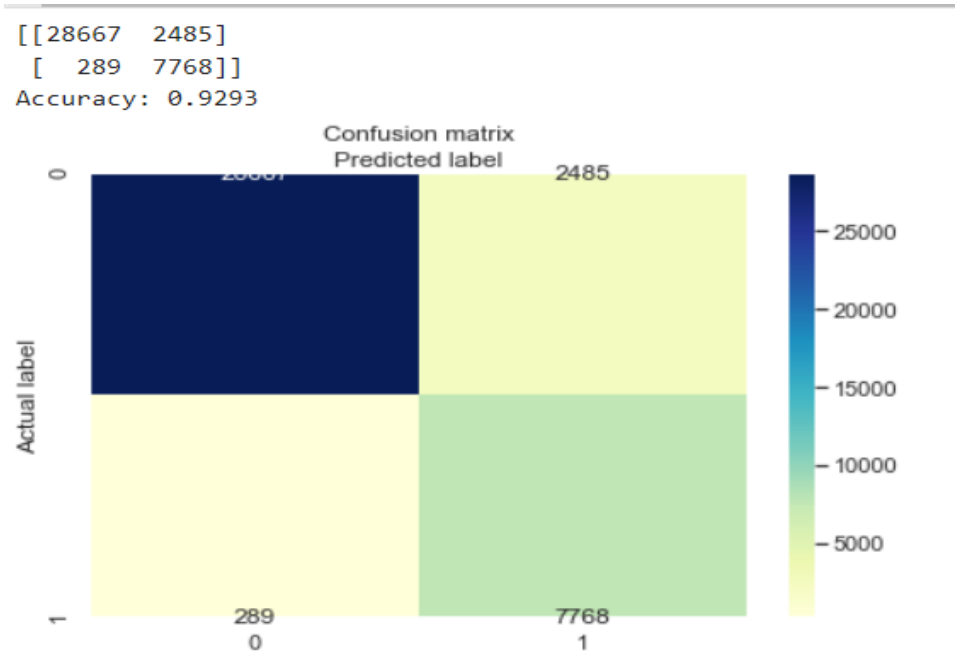
REPORT

Tools used and why

The skin dataset contains random samples of B,G,R values from face images of various people. Classes are skin or non-skin for tuples. Input values in x matrices are B, G or R and y values are 1 or 0.

The programs were written in a JupyterLab notebook. Import sklearn.linear_model LogisticRegression in order to implement logistic regression using the liblinear library. Fit the model using xtrain and ytrain data. Predict y values for xval and xtest data. Liblinear is a solver suitable for small datasets.

A confusion matrix evaluates the accuracy of a classification. Values in the matrix are C(I,j). Actual labels = I, predicted labels = j. C(0,0) = true negatives, C(1,1) = false positives. Sns.heatmap plots the rectangular confusion matrix as a color encoded matrix; cmap maps each value to a color space.



Fold 1 val data confusion matrix and heatmap

Accuracy of the model's predictions comes from metrics.accuracy_score(ytest, y_pred). It returns the decimal fraction of correctly classified samples. Using sklearn, find precision, recall and f-beta score. Calculate those values for the data.

Precision: is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.

Recall: is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is the ability of the classifier to find all the positive samples.

F-beta score can be interpreted as a weighted harmonic mean of the precision and recall. F-beta score reaches its best value at 1 and worst score at 0.

The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important. beta < 1 lends more weight to precision, while beta > 1 favors recall.

ACCURACY

FOLD	VAL	TEST
1	0.9293	0.9283
2	0.9304	0.9279
3	0.9313	0.9279
4	0.9266	0.9286
5	0.9303	0.9282
AVG	0.9296	0.9282

Sample calculations of precision, recall and f-beta score gave:

Fold 3 Test data

Accuracy: 0.9279

Precision: 0.9279

Recall: 0.9279

F-beta score: 0.9279

Fold 4 Test data

Accuracy: 0.9286

Precision: 0.9286

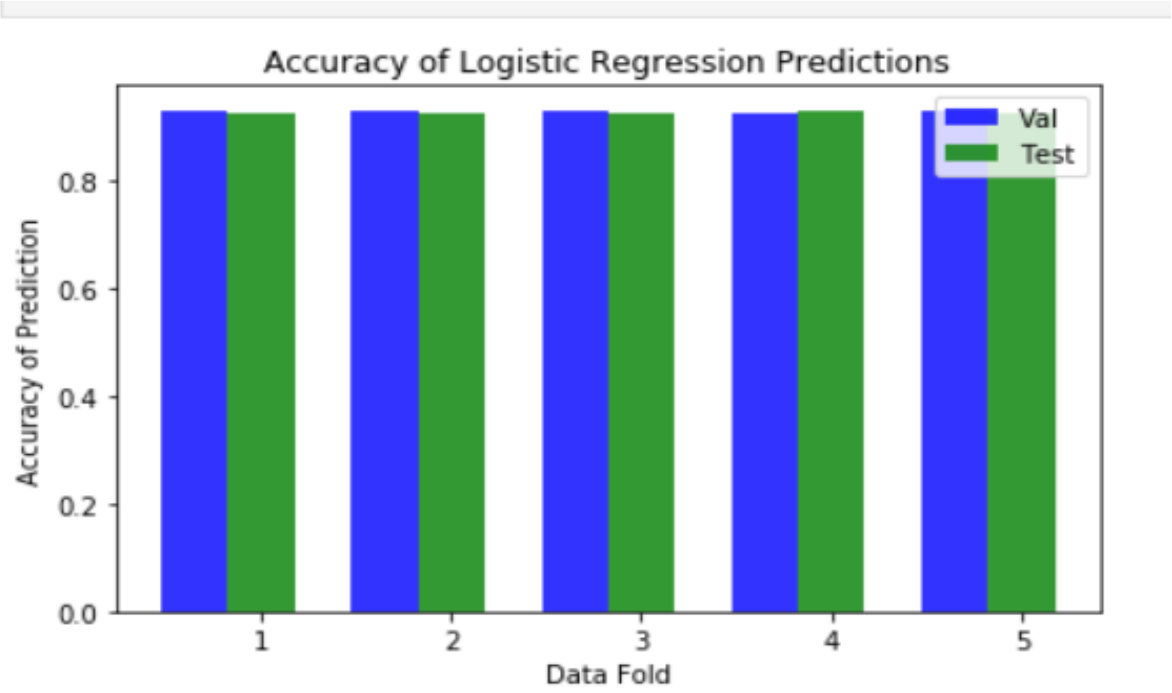
Recall: 0.9286

F-beta score: 0.9286

Precision, recall and f-beta score were similar values and like the accuracy.

Graph

Matplotlib as plt was used to plot a bar chart of accuracy for each fold.



Conclusion

Data in this project is not very varied in values, therefore we get similar values for accuracy and other metrics in each fold of the data. The model predicted y labels for data in all folds with an accuracy of about 0.928 – 0.929. If the data was varied, accuracy and other metrics would vary noticeably from one fold to another.

GitHub repo: https://github.com/mlp12/CS4347_Logistic_Regression