Logistic Regression

Due Date: 10th May 2019-11.00am

Implement logistic regression in a python notebook using the titanic data provided as described in the link below:

http://hamelg.blogspot.com/2015/11/python-for-data-analysis-part-28.html

Required:

- The Python notebook of your work
- Discussion of the following questions based on the tutorial. (To be done at the very bottom of your notebook)

Questions:

1. Describe what the sigmoid function defined in Cell number 3 in the tutorial does.
2. When does a sigmoid function output a probability greater than 0.5?
3. Age is a feature in the titanic data. It has some missing values:
    a. Explain how the missing values are handled in the tutorial
    b. Use a numpy function to imputing the missing values in age, to obtain the same results as in the tutorial
4. Do sklearn's machine learning functions allow for the features to be categorical?
5. In the tutorial, Label Encoding is used
    a. Explain the importance of label encoding
    b. Describe how it has been implemented in the tutorial
    c. Describe other ways in which encoding could have been done
6. What is the logistic regression model learned?
7. Discuss the difference between the methods predict and predict_proba of a logistic regression model
8. In regards to metrics:
    a. What is a confusion matrix?
    b. Discuss the confusion matrix of the model learned