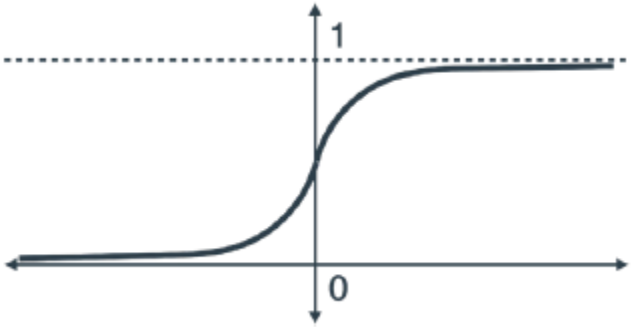
COMP-2704:
Supervised Machine Learning



Assignment 4: Logistic Classifiers

Setup

Complete the following steps to setup this assignment.

a) Open Jupyter Notebook in a web browser.
b) Within the "SupervisedML" folder, create a folder named "Assignment4" and navigate into this folder.
c) Open the link: https://github.com/luisguiserrano/manning/blob/master/Chapter_6_Logistic_Regression
d) Download the files:
   - *Sentiment_Analysis_IMDB.ipynb*
   - *IMBD_Dataset.csv*
   - *utils.py*
f) Upload all three files to the folder you created "~/SupervisedML/Assignment4".
g) Open the *Sentiment_Analysis_IMDB.ipynb* notebook and run all code cells. Fix any errors that occur.

Problem

Create a notebook with filename *SML_a4.ipynb* within the folder "~/SupervisedML/Assignment4". Add code and markdown cells to complete the following steps. You may copy relevant lines of code from *Sentiment_Analysis_IMDB.ipynb*.

1) *[2 marks]* Imagine a movie production company wants to use a sentiment analysis model to identify positive/negative reviews of their movies. Which is worse for this use case, a false positive or a false negative, or are they equally bad? What value of β would be suitable for an F_β score?

2) *[4 marks]* Load the original dataset into a dataframe and use the regex Python library to clean the text data so that it is better suited for sentiment analysis. Add a markdown cell to explain what you are doing.

3) *[1 mark]* Load the cleaned data and labels into an SFrame. Add a column named 'words' to the SFrame that stores the count of each word used in each review. Print the SFrame.

4) *[1 mark]* Split the data into training/validation/testing sets using 80%/10%/10% respectively.

5) *[3 marks]* Use Turicreate to create logistic classifiers for sentiment analysis. Experiment with different values of hyperparameters to develop two different models.

6) *[4 marks]* For each model:
   a) display the training and validation accuracies;

b) display the confusion matrix on the validation set;
c) calculate and display recall, precision, and F_β score (using the value of β you chose above) on the validation set.
d) plot the ROC curve and find the AUC for the validation set.

7) *[1 mark]* Select which of your two models is the best (or declare a tie) and justify your choice by commenting on metrics and the confusion matrix.

8) *[2 marks]* Using the test set:
a) calculate and display the accuracy;
b) display the confusion matrix;
c) calculate and display recall, precision, and F_β score.
d) plot the ROC curve and find the AUC.

9) *[2 marks]* With the exception of import and print statements, add a comment before each line of code in *SML_a4.ipynb* to explain what it does.

Contributions

Using markdown at the end of the notebook, list the contribution of each student to the assignment, referencing specific question numbers and other tasks such as formatting and submitting. Ideally, both students will contribute to all questions.

Submission

Upload your notebook to the Assignment 4 dropbox on the course website. Late submissions will lose 10%.

Total marks = 20