Machine Learning: Assignment 8: Web Page Classification

Problem Statement:

The goal of this assignment is to build a classifier to detect phising/malicious web pages.

Data Set Description:

The phishing problem is considered a vital issue in industry especially e-banking and e-commerce for the online transactions involving payments. A group of researchers have identified different features related to legitimate and phishy websites and collected 1353 different websites from difference sources. Phishing websites were collected from Phishtank data archive (www.phishtank.com), which is a free community site where users can submit, verify, track and share phishing data. The legitimate websites were collected from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser and they collected 548 legitimate websites out of 1353 websites. There is 702 phishing URLs, and 103 suspicious URLs.

The data set is available in this link (data folder):

https://archive.ics.uci.edu/ml/datasets/Website+Phishing

The data format is also described in the above link.

Assignment Tasks:

In this assignment you can use any preprocessing technique, and any (more than one maybe) machine learning algorithm to classify the above data set. You should also evaluate the techniques using any suitable evaluation methodology. *You have to submit a report (about 2-3 pages) in pdf format. The report should contain the following sections:*

1. Problem statement

2. Methodology: Brief description of the (i) preprocessing, (ii) learning algorithms, (iii) evaluation strategy used.

3. Experimental Results: Report on performance of the algorithms in terms of the evaluation measures preferably as tables or graphs.

4. Discussion of the results.

Submission Guidelines:

You should name your report file as <rollnumber_8.pdf> (e.g., 14CS10001_8.pdf). The submitted report file *should* be in pdf and have the following header comments. No need to submit any program.

Roll # Name # Assignment number

*Please submit the program in moodle by **November 14, 2018 midnight** (hard deadline). Copying from friends/web will lead to strict penalties.*