Machine Learning — Programming Assignment

June 2023

1 Problem definition and data

Clickbait headlines are designed to attract attention and to entice users to read the linked piece of online content. They are typically deceptive, sensationalized, or otherwise misleading. More importantly, they do not accurately reflect the content being delivered.

A software house wants to build a browser capable of distinguishing clickbaits from regular headlines. To do so a dataset has been collected, including 32 000 headlines, equally divided in the 'clickbait' and 'non-clickbait' classes. The dataset includes a training, validation, and test sets consisting of 24 000, 4000 and 4000 samples, resepctively. The data is stored in text files, with one headline for each line.

2 Assignment

We want to build a classifier that is able to predict if a given headline text is actually a clickbait. For the programming assignment you are expected to:

1. analyze and comment the data;
2. design and implement a suitable data pre-processing procedure;
3. implement, train and evaluate one or more classification models;
4. use suitable data processing and visualization techniques to analyze the behavior of the trained models.

Two scenarios should be considered: a generic one, in which all errors are equally important, and a 'precision-oriented' scenario, in which we would like to keep as small as possible che chance of false positives.

All the above should be implemented as scripts in the Python programming language. Any machine learning library (included `pvml`) can be used. Data and code used during the course can be used for the assignment if needed.

3 Report

Prepare a report of three to five pages documenting all your work. Provide detailed instructions on how to reproduce the results. The report must be in the PDF format. Include your name in the report and conclude the document with the following statement: "I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher."

Make a ZIP archive with the report and the Python scripts, and submit it from the course web page. To keep the size of the submission manageable, do not include files containing the original data, the features etc.