Capita Selecta Computer Science
Artificial Intelligence [H05N0a]
PU Learning

Homework 2021-2022

Overview

The homework is about learning from positive and unlabeled (PU) data, which is a special case of binary classification where a learner only has access to labeled positive examples and unlabeled examples. The unlabeled data contains both positive and negative examples.
The goal of the homework is to get familiar with the typical PU learning setting as well as with classic algorithms capable of dealing with the absence of negative labels. In particular, you will need to fully comprehend (at least) two PU learning models: S-EM [2] and Modified Logistic Regression [3]. We strongly recommend to, first, get familiar with the literature (PU learning survey) [1] and, second, to study in details the papers related to the methods [2,3].

Labeling mechanism

In order to show the practical differences between the performance of the two methods above, two datasets are provided together with their link to the original source. A particular and different labeling mechanism has been applied to each of them.
In the first dataset, namely *"Dataset1_train"* and *"Dataset1_test"*, a biased technique for sampling positive labels has been used. We followed this procedure: first, we clustered the data into two clusters, second we assigned a random weight to each of them, and finally we drew the examples from the positive population according to their weight.
In the second dataset, namely *"Dataset2_train"* and *"Dataset2_test"*, the Selected Completely at Random (SCAR) assumption holds. Therefore, we simply randomly selected the examples from the positive population.

Assignment

The successful completion of this homework consists of **three** parts.

First, you will have to correctly implement the two PU learning methods, namely *S-EM* [2] and *Modified Logistic Regression* [3]. The paper that proposes S-EM assumes a text classification setting, for which they use a Bernoulli Naive Bayes classifier, which does not apply to our datasets. Therefore, you will have to generalize the method to work with any classifier, meaning that you will need to generalize updating the Naive Bayes probabilities to refitting the classifier parameters. The Modified Logistic Regression method, mentions that they optimize the log likelihood by taking its gradient. This gradient is given further in this document.[1] For the implementation, we provide you the empty classes with the methods in python file, namely *"PU_Learning.py"*. You will have to fill in the empty spaces with your code. Tip: the code should work in general, not just the inputs used in the notebook. Do not forget about edge cases.

Second, you will have to show that the implemented methods work correctly. For this part, the notebook named *"Homework.ipynb"* is provided. Fill in the missing parts and add the additional experiments at the bottom, if any. Please, provide comments for better readability.

Third, you will have to write a report of **maximum two pages**. In the report, you should compare the different methods on the different datasets and interpret the results. To get started, answer the following questions:

1. According to your knowledge, which method should perform better on each of the datasets before testing it? Why?
2. Which method is in practice performing better? Does it follow your intuition? If not, why?
3. How do the predicted probabilities of the PU classifier compare to the probabilities predicted by a classifier trained with all the labels? (Their difference is measured using MSE in the notebook)
4. What is the impact of the dataset's labeling mechanism on the performance of each of the methods?

[1] If you need to refresh your general knowledge about gradient descent in logistic regression, you can start from this page: https://www.baeldung.com/cs/gradient-descent-logistic-regression.

After this initial analysis, formulate 3 interesting research questions to gain better understanding of the problem and the methods. Focus on analyzing them in depth and communicating the gained insights well. Clearly state each of the questions before answering them.

Likelihood Gradients for Modified Logistic Regression

The likelihood for the modified logistic regression with weights w and variable b is:

$$\text{loglik}(w,b) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{1}{1+b^2+e^{-wx_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+b^2+e^{-wx_i}}\right)\right]. \tag{1}$$

The parameters are updated using their gradient, with learning rate η, as follows:

$$\begin{aligned} \text{w\_new} &= w + \eta \cdot \frac{\partial \text{loglik}(w,b)}{\partial w} \\ \text{b\_new} &= b + \eta \cdot \frac{\partial \text{loglik}(w,b)}{\partial b} \end{aligned} \tag{2}$$

,

with their gradients:

$$\begin{aligned} \frac{\partial \text{loglik}(w,b)}{\partial w} &= \sum_{i=1}^{N} x_i e^{-wx_i}\left[\frac{y_i}{b^2+e^{-wx_i}} - \frac{1}{(1+b^2+e^{-wx_i})\,(b^2+e^{-wx_i})}\right] \\ \frac{\partial \text{loglik}(w,b)}{\partial b} &= \sum_{i=1}^{N} 2b\left[\frac{1-y_i\,(1+b^2+e^{-wx_i})}{(1+b^2+e^{-wx_i})\,(b^2+e^{-wx_i})}\right]. \end{aligned} \tag{3}$$

Submission

Submit your solution on Toledo → Capita Selecta → PU Learning → Homework (deadline: 17 dec). The submission consists of a zip file, named *LastnameFirstname.zip*, which includes the following files:

- PU_learning.py
- Homework.ipynb
- report.pdf

The deadline is Friday, December 17, 2021, 23:59.

Contact

For any issues, contact Lorenzo Perini at lorenzo.perini@kuleuven.be.

References

1 Learning from positive and unlabeled data: a survey.

2 Partially Supervised Classification of Text Documents.

3 A Modified Logistic Regression for Positive and Unlabeled Learning.