Homework 7

2-AIN-150, Winter 2022

Deadline: 8.1.2023, 23:59

Before you start solving the homework, please read the general instruction at the end of the document. Submitted solutions should be your own. Do not copy and do not try to find solution in literature or over the internet.

Leveraging unlabeled data during training

Consider binary classification scenario. Our training data has 200 samples with 2400 attributes. But we also have 10000 unlabeled samples from same distribution.

Your task is:

- Use to unlabeled data to fit some relevant transformation (e.g. one that lowers ammount of attributes).
- Train suitable classifier on transformed data.
- Cross-validate accuracy of your classifier (anything better than 75% is good enough).
- When you are sure about your methods, classify the new test data.

If your accuracy is very high (over 90%), or you use very interesting methods, you can get up to 5 bonus points.

Your program can use any libraries for basic math, matrix operation, matrix inversion, solving systems of linear equations, and calcualating numerical or symbolic derivatives. You are **allowed** to use scikit-learn library and its equivalents in other languages.

Python instructions There is `template.py` in package. You should fill out the function `magic(X, y, U, T)`, where (X, y) are labeled data, U are unlabeled data, and T are test data. Program can be runned using `python template.py <input file>`

General instructions

You should submit your code via Classroom. **Add some short commentary about your methods into source code.**

Ideally submit your homeworks in Python (fill out required functionality from assignment). You can use different language if you really want, but you need also to add auxiliary functionality like reading input and output. But your solution should be runnable under Linux using open source software.