Athens University of Economics and Business	**Department of Informatics**
M.Sc. Program in Data Science	
Optimization Techniques	
Instructors: G. Zois	
Project on Support Vector Machines	**Submission Deadline: Sunday, June 23, 2024**

The purpose of this project is to implement the Sequential Minimization Algorithm (SMO) for binary classification problems. You are allowed to work either on your own, or in groups of 2 people.

The way that the SMO algorithm works is that it tries to solve the dual rather than the primal problem. Suppose that we have a dataset of m training examples and n features, where the i-th example is in the form $(x^{(i)}, y^{(i)})$, with $y^{(i)} \in \{-1, +1\}$ being the label of the example. Then, if we look at the problem of linear separation of the data possibly with the existence of outliers, the primal problem is to identify a vector $w = (w_1, w_2, \ldots, w_n)$ and a constant term b, according to the following constraints:

$$
\begin{aligned}
&\min\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} s_i \\
&\text{s. t.:} \\
&\qquad y^{(i)} \cdot (w^T \cdot x^{(i)} + b) \geq 1 - s_i,\ i = 1, \ldots, m \\
&\qquad s_i \geq 0,\ \ i = 1, \ldots, m
\end{aligned}
$$

Here C is the penalization constant for the outliers. The corresponding dual problem is:

$$
\begin{aligned}
&\max\ W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\
&\text{s. t.:} \\
&\qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \\
&\qquad 0 \leq \alpha_i \leq C,\ \text{for } i = 1, \ldots, m
\end{aligned}
$$

If we solve the dual, we can compute the vector w and the term b as we saw in class.

Part A Implement a basic working version of the SMO algorithm. For this, you can hardwire a value for the constant C, which you will then try to optimize in Part B. In order to implement SMO, there are certain decisions you need to make during the execution of the algorithm, such as

- how to pick the pair of variables that you will examine in each iteration,
- how to pick an initial solution,
- how to set the termination criterion.

You are free to choose what you think is best for the algorithm. Please mention explicitly in your report what are the decisions you took for all the above issues.

There is no requirement for a graphic environment. Your training program should take as input the filename of the dataset and produce a file with the values for w and b. You are free to work with any of the standard programming languages (Java, Python, C, C++, ...).

Relevant References

1. Part V, Section 9 of the lecture notes from the course on Machine Learning by Andrew Ng. `https://sgfin.github.io/files/notes/CS229_Lecture_Notes.pdf`
2. Technical report by John Platt on the SMO algorithm (see at the course's eclass)

Part B The purpose of this part is to apply (and possibly optimize) your algorithm for a specific dataset.

- Go to `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`
- Download the gisette dataset.
- Use at least 4500-5000 points for training your algorithm.
- Try to optimize the choice of the constant C or any other choices that you make for the training.
- Optionally, you can also try to use kernels, however, for the purposes of the project, a linear separation is enough.
- It is recommended to try out other datasets as well from the above website. But your report should focus on gisette. In case you have difficulties with gisette or you have a strong preference for another dataset, you should contact me first.

Part C - Deliverables

The deliverables of your work should consist of:

1. All your code, both for training and for testing. Put everything that is needed for the code to run in a folder called src. Provide enough documentation so that it is clear what the algorithms do, and how they should be run.
2. A report (minimum 3 pages) where you will comment on the implementation of SMO, on the choices you made and on the results you obtained for the dataset. You are free to use plots to illustrate any aspects of the algorithm (plots are not mandatory though).