Machine Learning I (Final Project) (Due Sep 25'th, 2021)

Task #1:

Objective:

- We can use the load prediction dataset to detect the possible defaulters for Consumer Loans.
- In addition, you will use the regression analysis to predict the income of a person using various features provided in the dataset.

Dataset: Loan Prediction Based on Customer Behavior dataset

- Available at: https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior
- The dataset consists of 13 different attributes for the customer such as age, martial status, income, experience …etc and one target variable indicating the risk flag of being a loan defaulter.

You are requested to do the following **for both the regression and classification problems**:

Task 1-a: Classification

1- Load the data and perform all necessary data cleaning and scaling.

2- Data inspection. Use any relevant functions that can help you to understand the data. Use any necessary visualization techniques to inspect your data

3- Explore the selection of various feature variables for classification. You should include at least one categorical feature.

4- Classify the data using various classification methods explored in ML1 (logistic regression, SVC, Decision trees, KNN classifier). Explore using different model parameters in the built-in sklearn libraries.

5- Explore the use of your own implementations of each Model. Comment on your results.

6- For each model provide suitable quantitative metrics for assessing the performance of your model based on the required application.

Task 1-b: Regression Analysis

1- Load the data and perform all necessary data cleaning and scaling.

2- Data inspection. Use any relevant functions that can help you to understand the data. Use any necessary visualization techniques to inspect your data

3- Explore the selection of various feature variables for regression to estimate the income of each customer. You should include at least one categorical feature.

4- Perform various regression analysis using various methods explored in ML1 (Linear regression, Multiple regression, SVR regression, Polynomial regression). Explore using different model parameters in the built-in sklearn libraries.

5- Explore the use of your own implementations of each Model. Comment on your results.

6- For each model provide suitable quantitative metrics for assessing the performance of your model based on the required application.

Task #2:

You are asked to write a small survey 3 to 4 pages' maximum in a professional way to discuss the origins of Kernels in Machine Learning. Questions that should be answered in that survey are:

1. Who proposed the idea with SVM? History of SVM?
2. What is the motivation?
3. The usage of kernels in SVM?
4. What rules that should be fulfilled to implement a kernel function?
5. Comparison between kernel types
6. How to choose correct Kernel for an ML problem?

7-Extending SVC for multi-class classification

8. The Concept of VC dimension
9. The Curse of Dimensionality

Try to have a good technical report structure and the order of topics should be rational. The most important thing is learning to extract useful information from the resources given to you and get used to reading ML papers research papers. This task is designed to tell you to what extent the machine learning field is complicated and broad! You can Dig in the math up to the level of your interest.

The following resources are extremely important.

1.The following video with its other subsequent two parts:
Support Vector Machines Part 1 (of 3): https://www.youtube.com/watch?v=efR1C6CvhmE

2. Lecture 7 - Kernels https://www.youtube.com/watch?v=8NYoQiRANpg

Stanford CS229: Machine Learning (Autumn 2018) Andrew Ng lecture about SVM and kernels.
Slides here: read till section 6 Kernel Methods, live lecture notes:
http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes3.pdf,
full course if interested: http://cs229.stanford.edu/syllabus-spring2021.html

3. Learning from data: Learning from Data - Online Course (MOOC) by Prof Yasser Abu

Mostafa:
https://home.work.caltech.edu/telecourse.html
a. VC dimension: lecture 7
b. SVM and Kernels: lectures 14 & 15.

4. This is a complete course about kernels with slides and videos: Machine learning with kernel methods, 2021.
https://members.cbio.mines-paristech.fr/~jvert/svn/kernelcourse/course/2021mva/index.html
First 4 lectures

5. The concept of a hilbert space and inner products: Inner Products in Hilbert Space
https://www.youtube.com/watch?v=g-eNeXlZKAQ

6. VC dimension:
https://www.youtube.com/watch?v=puDzy2XmR5c

7. Papers and books:
a. CiteSeerX — A Training Algorithm for Optimal Margin Classifiers SVM first paper
https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3818

b. https://arxiv.org/pdf/math/0701907.pdf Just read the first two sections “intro and kernels” https://arxiv.org/pdf/math/0701907.pdf