# Homework on machine learning

May 14, 2024

Electronic submission to Gradescope due on **5:00pm Wednesday 5/15**.

The required python scripts are formulated and described in the colab notebook `https://colab.research.google.com/drive/1XnRa3XcHKAz4hC9jFKgUpPzjyK_M1FHn?usp=sharing` and `https://colab.research.google.com/drive/1vFs6ECjiSBGdY3PCG6G9yO_RQVzfyup_?usp=sharing`..

Please do not directly edit the colab notebook. Make a copy of the notebook and then solve the problems on it. While submitting please attach the relevant python scripts and/or excel files to enable us to give partial credits. Please send the required excel files to us (on the staff mailing list `msande214-spr2324-staff@lists.stanford.edu`) on email and paste the required urls for the colab files on the submitted pdf.

## Problem 1: Linear Regression

We use the Boston housing dataset from Pedregosa et al. (2011), which is a well-known dataset for regression tasks. It contains information about various houses in Boston, including features like the number of rooms, property tax rate, pupil-teacher ratio, etc., with the target variable being the median value of owner-occupied homes in 1000s. In this homework we expect you to train a regression model to obtain the weight vector $\mathbf{w}$ (denoting the coefficients corresponding to the features) and bias $b$ (constant term) on python using optimization solvers. Train the model with and without regularization and report the difference. The files to download the dataset and test-train split are given in the colab file.

**Note**: This Boston housing data has an ethical issue `https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html` and will be discussed in the lecture.

## Problem 2: Pattern classification

Consider the following data from the attached file "data.csv" with four columns corresponding to the label, x-coordinate , the y-coordinate and z-coordinate. The file consists of 100 rows with each row corresponding to one point. Observe that we just have two labels (0 and 1) in the entire csv file.

Solve a linear program to check whether this data points corresponding to the first label and those corresponding to the second label can be separated from each other by a hyperplane. Also try plotting the data on python (using plt.scatter) or any other visualization software.

**Note**: Since, each data point has 3 dimensions, it is fine to give three plots with two dimensions plotted each time. An example is given in the colab file.

## Problem 3: Iris classification using logistic regression

Consider the iris data from Pedregosa et al. (2011) stored in file "iris-biodata-hw2.xlsx" which consists of two different types (versicolor and setosa) of irises' with four features for every data point. Let us denote every data point by $\mathbf{u}^i \in \mathbb{R}^4$ for every $i \in [N]$. We now consider a model with parameters $\mathbf{P} \in \mathbb{R}^4$ (corresponding to the coefficients for every feature) and constant term $\alpha \in \mathbb{R}$.

Recall that for pattern classification in class, we defined the penalty to be $\max(0, -\mathbf{P}^T\mathbf{u}^i + \alpha + 1)$ and $\max(0, \mathbf{P}^T\mathbf{u}^i - \alpha + 1)$ when the data-point $\mathbf{u}^i$ belongs to versicolor and setosa respectively. In this problem we shall consider a loss function based on the logistic regression model.

The penalty for the case when iris data-point $\mathbf{u}^i$ belongs to versicolor class is given by

$$L_1(\mathbf{z}^i) = -\log\left(\frac{1}{1+e^{-z}}\right) \text{ where } \mathbf{z}^i = \mathbf{P}^T u^i - \alpha \tag{1}$$

and when the iris-datapoint $\mathbf{u}^i$ belongs to setosa class is given by

$$L_2(\mathbf{z}^i) = -\log\left(\frac{e^{-z}}{1+e^{-z}}\right) \text{ where } \mathbf{z}^i = \mathbf{P}^T u^i - \alpha \tag{2}$$

Plot the function $L_1(z)$ as a function of $z \in (-2, 2)$.

Now, train the model on excel by solving an optimization problem on excel by summing over all the penalty functions for every data point. Train the model with and without a regularisation parameter $C$ and observe the difference in the trained weight vectors.

Solve and report the objective and the learnt weight vectors and also report the accuracy. Also send us the excel files on email. Now argue why this loss function is "easy" to optimize.

**Hint**: Can you observe from the plot that the function $L_1(z)$ can be approximated by set of tangent lines and this function is given by the upper envelope.

# Problem 4: Training a ML model using SVMs

This problem is entire described in the following colab file
`https://colab.research.google.com/drive/1vFs6ECjiSBGdY3PCG6G9yO_RQVzfyup_?usp=sharing`.

Solve it on colab and submit it to us by pasting the url on the uploaded pdf to gradescope.

# References

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.