# COMP341 Introduction to Artificial Intelligence
# HW6: Machine Learning

- This homework is about machine learning. You are given one dataset for binary classification and three datasets for regression.
- The binary classificiation includes images. You are going to extract features from these images as part of your homework.
- You will implement k-Nearest Neighbors and linear regression. You are also going to compare these to logistic regression and ridge regression, respectively.
- By submitting this homework, you **agree** to fully comply with Koç University Student Code of Conduct, and accept any punishment in case of failure to comply.
- Make sure you read the submission instructions (the last section of this document) carefully.

## Introduction

In this homework, you will implement a feature extraction method and two machine learning methods. The datasets are provided to you. Cross-validation to handle overfitting, performance visualization, data loading, regression data pre-processing are handled for you. Even though we strongly suggest that you interpret the results, a report will not be needed.

As a result of these, detailed explanations and help with installing extra modules will not be provided. The provided code has detailed comments and hints, it is up to you to go over them. Furthermore, learning to install needed modules for python will be a valuable experience.

## Programming

### Preliminaries

You are going to need the following python modules:

- Numpy (possible Numpy+Mkl)
- Matplotlib
- OpenCV
- Scikit-Learn

If you are using a Linux distribution, I would assume that you know what you are doing and that you probably do not need any help with installing these. (Hint: pip or apt-get)

If you are using windows, you can use pip. I suggest you use the libraries provided here with pip instead of their defaults: `https://www.lfd.uci.edu/~gohlke/pythonlibs/`.

If you are using Mac OS, you can also use pip with the default modules.

Hint for using pip: `python -m pip install <whl file OR package name>`

## Datasets and the ML Problems

### Classification

In the classification part of the homework, you are going to classify whether an image involves **wood** or **metal**. The images are provided under two folders. You are going to use the k-Nearest Neighbor and the Logistic Regression algorithms for this. You are going to only implement the kNN approach, logistic regression is provided to you.
**WARNING:** Using existing kNN implementations, for example the scikit-learn version, is prohibited and will not get any credit. You can however use these to verify your code.

In addition to implementing the kNN method, you are going to implement a feature extraction approach, involving histograms. Look at the code for further details.

### Regression

In the regression part of the homework, you are given three regression datasets. The inputs and targets are defined and extracted for you. You are going to use linear regression and ridge regression (AKA linear regression with Tikhonov regularization). You are going to only implement linear regression. Ridge regression is provided to you.

**WARNING:** Using existing linear regression implementations, for example the scikit-learn version, is prohibited and will not get any credit. You can however use these to verify your code. You are also allowed to use matrix operations provided by Numpy.

## Implementation

There are 5 python files. You are going to complete the code in *data.py* and *learners.py* for this project. YOu would also want to play around with *main.py* for debugging. The entry point of the code is the *main.py*. You can directly call it as:

```
python main.py
```

If you are getting `ImportError: No module named ...` type errors, go back to Preliminaries. If not, you would get a `*** Method not implemented: ...` message.
At this point, what you need to do should be obvious. First read all the comments in *main.py*, then move on to the needed parts of the code. Specifically, you need to complete

- `extract` method of the `SaturationHistogramExtractor` class in file *data.py*
- `fit` and `predict` methods of the `knnClassifier` class in file *learners.py*
- `fit` and `predict` methods of the `LinearRegression` class in file *learners.py*

There are enough comments for you to go on. Do not forget that you have access to the slides as well. If there is something that you do not know, for example what a histogram is or how an image is represented, internet is your friend. Self-learning is part of this homework.

## Outputs

The code that is provided to you prints to the standard output and saves performance figures as PNG files. We have provided example outputs for you under the *example-outputs* folder. If your machine is 64-bit, you need to get the exact results (the plot colors and size might change but the numebrs should be the same) since we are fixing the random seed. If you are not getting the exact results, either your implementation is wrong or your version of the random number generator (RNG) is different than ours. Seek the instructor or the TAs if you think this is the problem.

## Submission

The deadline is January $5^{th}$ at 11.59PM. You are going to submit a compressed archive through the blackboard site. This should extract to a folder with your student ID which includes *learners.py* and *data.py*.

- The submitted file can have *zip*, *tar* or *tar.gz* format.
- This compressed file should extract to a folder with your *student identification number* with the two leading zeros removed which should have 5 digits.
- The previous point is very important, I do not want to see multiple folders (apart from operating system ones such as MACOSX or DS Store). I do not want to play inception with your code.
- Inside the folder, you should only have *learners.py* and *data.py*. Anything else will be deleted. If they interfere with the grading process, I will simply ignore your submission.
- One advice is after creating the compressed file, move it to your desktop and extract it. Then check if all the above criteria is met.
- **DO NOT SUBMIT CODE THAT DOES NOT TERMINATE OR THAT BLOWS UP THE MEMORY**. I will take these as malicious acts and will proceed accordingly.

The late policy is in the syllabus.