National Taiwan University **Introduction to Machine Learning and Deep Learning**

Department of Civil Engineering Instructor: C.-S. CHEN

Homework 1

Essay and Programming, Due 21:00, Wednesday, October 13, 2021

Late submission within 24 hours: score*0.9;

Late submission before post of solution: score*0.8 (the solution will usually be posted within a week); no late submission after the post of solution)

Total 80%

1. **(20%)** Name your essay `boolean_hypothesis.pdf`. The learning example considered in the lecture indicates the entire Boolean hypothesis set $\mathcal{H}$ over a n-bit representation of input space is 2^{2^n}. If we denote binary output by ●/○ for visual clarity, we can list all the possible hypotheses h_i for a one-bit representation of input space below:

x	h_1	h_2	h_3	h_4
0	○	○	●	●
1	○	●	○	●

   Tabulate all the possible hypotheses h_i for a two-bit representation of input space.
   In this problem you need to report you answer in pdf file.

2. **(30%)** The curse of dimensionality commonly occurred in machine learning refers to phenomena that arise when learning from data in high-dimensional spaces that do not occur in low-dimensional settings. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality.

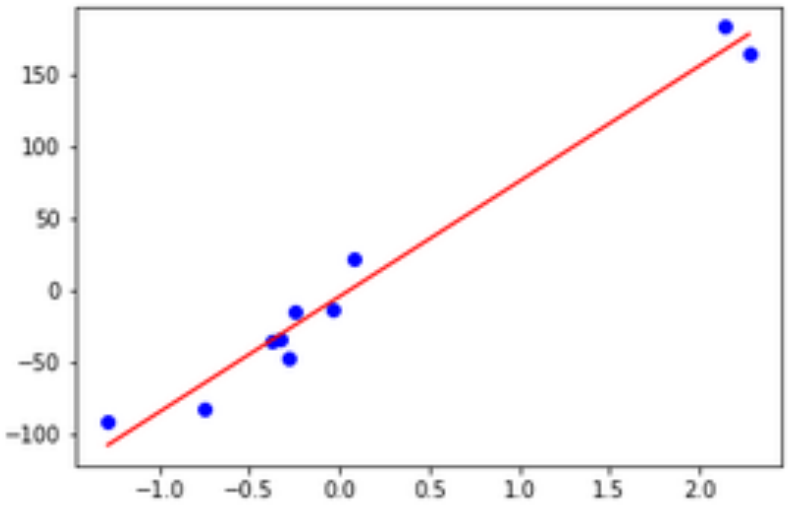
   Please follow the instructions in Hw1-2.ipynb template from Google Colab, and complete your codes under the `# TODO` annotations.
   Colab link:
   https://colab.research.google.com/drive/1Shrz1JWvxud4IM3LUnHsfutSoh_x_vUF?usp=sharing
   In this problem you need to save your answer into csv file.

3. **(30%)** For a set of n samples $(x_0, y_0), (x_1, y_1), \cdots, (x_{n-1}, y_{n-1})$, you can easily fit a line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ for these samples. This fitting is called simple linear regression. Using the least square fitting criterion, we can show:

$$\hat{\beta}_1 = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n-1}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean. Please follow the instructions in Hw1-3.ipynb template from Google Colab and complete your codes under the `# TODO` annotations to compute $\hat{\beta}_0$ and $\hat{\beta}_1$ from a set of samples and plot thse samples and fitting line.
And a sample plot (`line width = 3, marker size = 20`):



Colab link:
https://colab.research.google.com/drive/1fUw-FeXv-74th1xeSGzzKsdTu6xICUrP?usp=sharing
In this problem you need to submit the figure and save $\hat{\beta}_0$ $\hat{\beta}_1$ value into csv file.

Homework Submission Format (Important!!!):
Pack two csv files from problem 2 and problem 3 into one folder, and name this folder <csv>.
Pack the pdf file from problem 1 and figure from problem 3 into one folder, and name this folder <pdf&figure>.
Compress these two folders into < your_student_id_ hw1>.zip, and upload it to NTU COOL.

- **Example:**