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

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

Homework 2

~~**Essay and Programming, Due 21:00, Thursday, October 6, 2022**~~

Essay and Programming, Due 21:00, Thursday, October 13, 2022

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 150%

If have any questions in this homework, contact TA 陳冠亦 (r11521609), and 李國碩 (d10521013).

1. **(0%)** Please fill in this survey. Link: https://forms.gle/5STcaEPg8rgkNQKq7

2. **(20%)** Name your essay `boolean_hypothesis.pdf`.

   (a) 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.

   (b) Now consider a three-bit representation with the following ground truth. List all the possible target functions.

x	y
000	○
001	●
010	○
011	○
100	●

   In this problem, you need to report your answer in a pdf file.

3. **(30%)** Name your essay `biasVar100.pdf`.

   (a) Repeat the bias and variance example from the lecture with a training data $\mathcal{D}$ consisting of 100 points. Plot error, bias and variance versus x. Calculate and report the expected out-of-sample error and its bias and variance components.

   (b) Compare your results with the training data $\mathcal{D}$ consisting of only 2 points from the lecture. Write a short essay to rationalize and comment your findings.

In this problem, you need to report your answer in a pdf file.

4. (60%)

(a) Consider the bias and variance example covered in the class. Suppose we now have a hypothesis set consisting of all horizontal lines $h(x) = b$. The input variable x is uniformly distributed in the interval $[-1, +1]$. The training data $\mathcal{D}$ consists of only <u>two points</u> $\{x_1, x_2\}$. The target function $f(x) = \sin(\pi x)$. The data set is $\mathcal{D} = \{(x_1, \sin(\pi x_1)), (x_2, \sin(\pi x_2))\}$. The learning algorithm returns the line at the midpoint $b = \frac{\sin(\pi x_1)+\sin(\pi x_2)}{2}$ as $g^{(\mathcal{D})}$ ($\mathcal{H}$ consists of functions of the form $h(x) = b$). Modify the code distributed in <u>Hw2-4.ipynb template from Google Colab</u> and record the bias and variance.

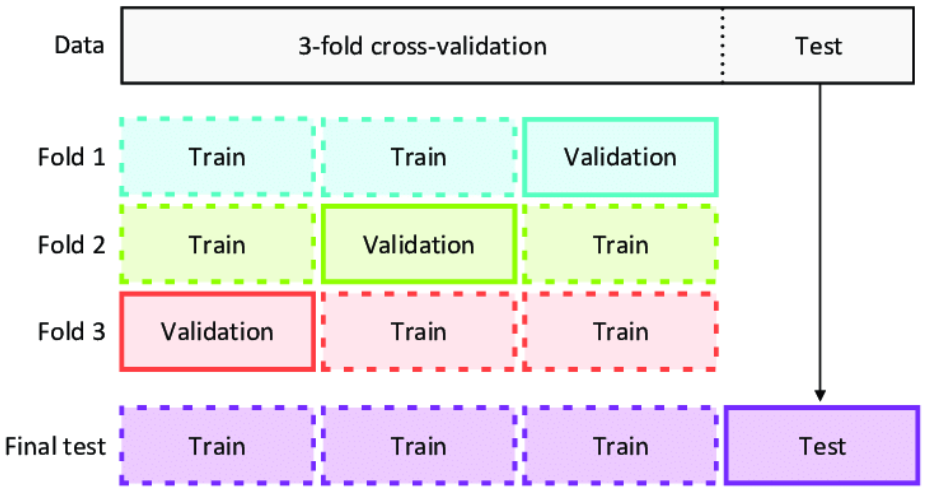
Write a function called "output" to return (bias, variance), turn .ipynb into .py and submit .py file named "`Hw2-4.py`"

(b) Name your essay `h0_100.pdf`. Now increase your training data $\mathcal{D}$ to <u>100 points</u>. Calculate and report the expected out-of-sample error and its bias and variance components. Compare your results with the training data $\mathcal{D}$ consisting of only <u>2 points</u> from **(a)**. Write a short essay to rationalize and comment your findings.

5. (40%) Cross validation can be used to help us find the best hyperparameters in an algorithm. In this exercise, we will use an in-house *C&RT tree with random forest bagging algorithm* as a black box to predict whether the class of test data is 1 or -1. In order to find the best hyperparameter combination of this classification algorithm, you will do two parts following the instructions in Hw2-5.py template. Submit `Hw2-5.py`

(a) Record your best accuracy in validation, best tree number, best tree depth and test accuracy from the holdout validation. All the codes have been implemented. You will receive 40% of the total score.

(b) Please implement a 3-fold cross validation, take the average of accuracy from 3 folds and observe which hyperparameter combination gives the best results. Record your best accuracy in validation, best tree number, best tree depth and test accuracy from the 3-fold cross validation. You will receive 60% of the total score.



- **Submission Format:** Please compress all the five files (boolean_hypothesis.pdf, biasVar100.pdf, h0_100.pdf, HW2-4.py, HW2-5.py) into yourStudentId_hw2.zip, then upload it to NTU COOL.