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, Wednesday, October 2, 2024

Late submission within 24 hours: score*0.9;
Late submission before the post of solution: score*0.8 (the solution will usually be posted within a week); no late submission after the post of solution)

Total 170%

1. (10%)

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_i
0	○
0	●
1	○
1	●

Consider a subset of five-bit representation 0_1_1. Please list all the possible hypotheses h_i in `HW2_report_template.docx`. (hint: below are two of hypotheses from this subset)

x	h_i
00111	○
00111	●

2. (30%) As we mentioned in the lecture, learning is only feasible in a *probabilistic* way (PAC: probably approximately correct). We can predict something useful outside the training set $\mathcal{D}$ using only $\mathcal{D}$ if a stable probability structure for both the in-sample and out-of-sample data exists. We can play around `Boolean_Learning.ipynb` distributed in class to reinforce our understanding. To make your results reproducible, always use random seed 12 and 30 when sampling the training and testing examples.

(a) Copy `Boolean_Learning.ipynb` and change the file name to `HW2_2.ipynb`.

(b) Repeat what we did in the class using 5 and 12 training examples. You should obtain an error rate of 0.28 and 0.127.

(c) Keep the probability distribution of training examples and alter the probability distribution of testing examples so that only the first four elements from the input space $\mathcal{X}$ can be chosen. Report your error rate using 5 training examples and 12 training examples.

3. (30%)

(a) Repeat the bias and variance example from the lecture with a training data $\mathcal{D}$ consisting of <u>**20** points</u>. Plot error, bias and variance versus x. Calculate and report the expected out-of-sample error and its bias and variance components.
Please fill answers in `HW2_report_template.docx.`

(b) Compare your results with the training data $\mathcal{D}$ consisting of only <u>2 points</u> from the lecture. Write a short essay to rationalize and comment on your findings in `HW2_report_template.docx.`

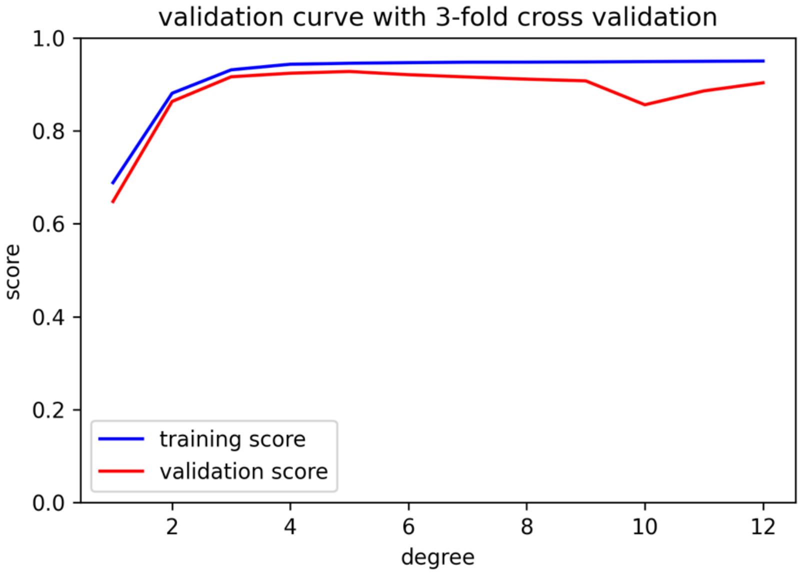
4. (60%)

(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>2 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 $h(x) = b$). Write a program to compute the bias and variance. Please fill answers in `HW2_report_template.docx.`

(b) Now increase your training data $\mathcal{D}$ to <u>20 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 on your findings in `HW2_report_template.docx.`

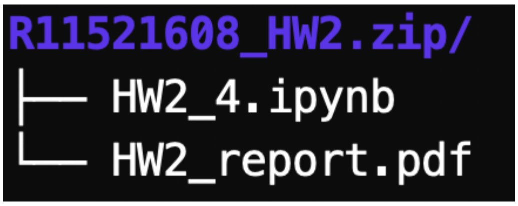
5. (40%)

The intuitive bias and variance example considered in the lecture indicates a low-order polynomial tends to have a high bias (underfit) and a high-order polynomial tends to have high variance (overfit). It is sometimes helpful to plot the influence of a single hyperparameter (in this case, the polynomial degree) on the training score and the validation score to find out whether the estimator is overfitting or underfitting for some hyperparameter values. The function in scikit-learn `validation_curve` can help in this case. Start from `HW2_4.ipynb` where 100 data points were already generated. Please plot the mean score of training and validation curves using a polynomial regression model with **3-fold** cross validation in `HW2_report_template.docx`. Below is a sample plot.



Submission Format

Convert HW2_report_template.docx to HW2_report.pdf, then place HW2_report.pdf and HW2_4.ipynb into a folder named {yourStudentID}_HW2 and compress it into a ZIP file for upload to NTU COOL. Below are the file formats for upload.



```
R11521608_HW2.zip/
├── HW2_4.ipynb
└── HW2_report.pdf
```