ML HW3

R13922146 葉富銘

Problem 5

To disprove this, we can simply construct two hypothesis sets H_1 and H_2, whose VC dimension are both 0, but $d_{vc}(H_1 \cup H_2)$ is 1 :

$H_1 = \{h_1\}$, $H_2 = \{h_2\}$, $h_1(\mathbf{x}) = 0$, so it always predicts 0 ; $h_2(\mathbf{x}) = 1$, so it always predicts 1.

It is trivial that both H_1 and H_2 can't shatter even one point, so their VC dimension is 0. but $H_1 \cup H_2$ can shatter one point, and can't shatter two points, so its VC dimension is 1.

Therefore, $d_{vc}(H_1 \cup H_2) = 1 > d_{vc}(H_1) + d_{vc}(H_2) = 0$, hence is disproved.

Problem 6

To get an ideal mini-target, we need to minimize the expected cost. Let the cost of a false negative be $C_{\mathrm{FP}} = 1$, and the cost of a false positive be $C_{\mathrm{FN}} = 10$.

1. The expected cost of misclassifying as +1 : Cost(+1) = C_{FP}(P(y = -1|$\mathbf{x}$)) = 1 − P(y = +1|$\mathbf{x}$)
2. The expected cost of misclassifying as -1 : Cost(-1) = C_{FN}(P(y = +1|$\mathbf{x}$)) = 10(P(y = +1|$\mathbf{x}$))

Upon decision, we should choose the classification whose cost is lower. so we classify as +1 if Cost(+1) ≤ Cost(−1), which lead to 1 − P(y = +1|$\mathbf{x}$) ≤ 10(P(y = +1|$\mathbf{x}$)), hence P(y = +1|$\mathbf{x}$) ≥ $\frac{1}{11}$. And we classify as -1 if P(y = +1|$\mathbf{x}$) < $\frac{1}{11}$

Given above, we can construct $f_{\mathrm{MKT}}(\mathbf{x})$ = sign(P(y = +1|$\mathbf{x}$) - $\frac{1}{11}$), hence $\alpha = \frac{1}{11}$.

Problem 7

$E^{(2)}_{out}(h) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) = y]\!] + \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) \neq y]\!]$

1. $\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) = y]\!] \leq \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\![h(\mathbf{x}) \neq f(\mathbf{x})]\!] = E^{(1)}_{out}(h)$, since if $h(\mathbf{x}) \neq y$ and $f(\mathbf{x}) = y$, then $h(\mathbf{x}) \neq f(\mathbf{x})$ must be true.
2. $\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) \neq y]\!] \leq \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![f(\mathbf{x}) \neq y]\!] = E^{(2)}_{out}(f)$, which is obvious.

Combine the result above, we can derive the inequality :

$E^{(2)}_{out}(h) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) = y]\!] + \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})}[\![h(\mathbf{x}) \neq y, f(\mathbf{x}) \neq y]\!] \leq E^{(1)}_{out}(h) + E^{(2)}_{out}(f)$, hence proved.

Problem 8

According to the assumption that X^TX is invertible, $\mathbf{w}_{\text{LIN}} = (X^TX)^{-1}X^Ty$.

We can construct $X_{\text{LUCKY}} = XD$, D is a diagonal matrix with $D_{0,0} = 1126$ and other diagonal elements are 1. so X_{LUCKY} comes from X with every x_0 being changed to 1126 instead of 1.

$\mathbf{w}_{\text{LUCKY}} = (X_{\text{LUCKY}}^TX_{\text{LUCKY}})^{-1}X_{\text{LUCKY}}^Ty = ((XD)^TXD)^{-1}(XD)^Ty = (D^TX^TXD)^{-1}D^TX^Ty$.

Since D is a diagonal matrix, it is invertible. Hence $\mathbf{w}_{\text{LUCKY}}$ can be further rewrited to $D^{-1}(X^TX)^{-1}(D^T)^{-1}D^TX^Ty$ $= D^{-1}(X^TX)^{-1}X^Ty = D^{-1}\mathbf{w}_{\text{LIN}}$, and hence $\mathbf{w}_{\text{LIN}} = D\mathbf{w}_{\text{LUCKY}}$ is proved.

Problem 9

9.

$$\tilde{h}(x) = \frac{1}{2}\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}+1\right)$$

$$\Rightarrow \max \text{ likelihood}(\underline{w}) \propto \prod_{n=1}^{N} h(y_n\underline{x}_n) = \prod_{n=1}^{N} \frac{y_n}{2}\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}+1\right)$$

$$\Rightarrow \max \ln \prod_{n=1}^{N} \frac{y_n}{2}\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}+1\right)$$

$$\Rightarrow \min \frac{1}{N}\sum_{n=1}^{N} -\ln\left(\underbrace{\frac{y_n}{2}\left(\boxed{\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}}^{\,\square}+1\right)}_{\bigcirc}\right) = \tilde{E}_{in}(\underline{w})$$

To minimize $\tilde{E}_{in}(w)$, Find $\nabla\tilde{E}_{in}(\underline{w}) = 0$

$$\Rightarrow \frac{\partial \tilde{E}_{in}(\underline{w})}{\partial w_i} = \left(\frac{\partial -\ln(\bigcirc)}{\partial \bigcirc}\right)\left(\frac{\partial\left(\frac{y_n}{2}(\square+1)\right)}{\partial \square}\right)\left(\frac{\partial\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}\right)}{\partial x_i}\right)$$

$$= \left(\frac{-1}{\bigcirc}\right)\left(\frac{y_n}{2}\right)\left(\frac{x_i}{(1+\underline{w}^T\underline{x})^{\frac{3}{2}}}\right)$$

$$= \frac{-1}{\frac{y_n}{2}\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}+1\right)} \times \frac{y_n}{2} \times \frac{x_i}{(1+\underline{w}^T\underline{x})^{\frac{3}{2}}}$$

$$\Rightarrow \nabla \tilde{E}_{in}(w) = \frac{-1}{\frac{y_n}{2}\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}+1\right)} \times \frac{y_n}{2} \times \frac{\underline{x}}{(1+\underline{w}^T\underline{x})^{\frac{3}{2}}}$$

#

(Derivation of $\left(\frac{\partial\left(\frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}\right)}{\partial x_i}\right)$;)

Let $f(w) = \frac{\underline{w}^T\underline{x}}{\sqrt{1+(\underline{w}^T\underline{x})^2}}$, let $\underline{w}^T\underline{x} = z$,

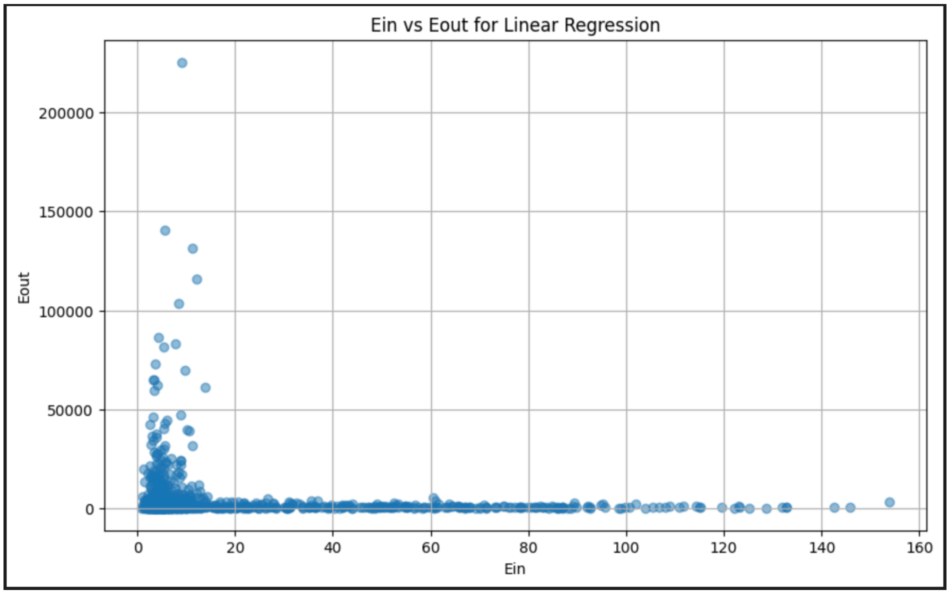
$$g(z) = \frac{z}{\sqrt{1+z^2}} \Rightarrow \frac{\partial f(w)}{\partial w_i} = \frac{\partial g(z)}{\partial w_i}$$

$$\Rightarrow \frac{\partial g(z)}{\partial w_i} = \frac{\partial g(z)}{\partial z}\frac{\partial z}{\partial w_i} = \frac{\sqrt{1+z^2} - z\cdot\frac{1}{2}(1+z^2)^{-\frac{1}{2}}\cdot 2z}{1+z^2}\cdot x_i$$

$$= \frac{\sqrt{1+z^2} - \frac{z^2}{\sqrt{1+z^2}}}{1+z^2}\cdot x_i = \frac{x_i}{(1+z^2)^{\frac{3}{2}}} = \frac{x_i}{(1+\underline{w}^T\underline{x})^{\frac{3}{2}}}$$

Problem 10

When the E_{in} is small , there are many E_{out} which are large, but when E_{in} get larger, E_{out} decrease rapidly. I infer that it's because in some experiments, we may select the data which is not general enough, or even the edge case. And $\mathbf{w}_{lin}$ match the selected data well, which let E_{in} be small, but it performs bad on others, which let E_{out} very large.



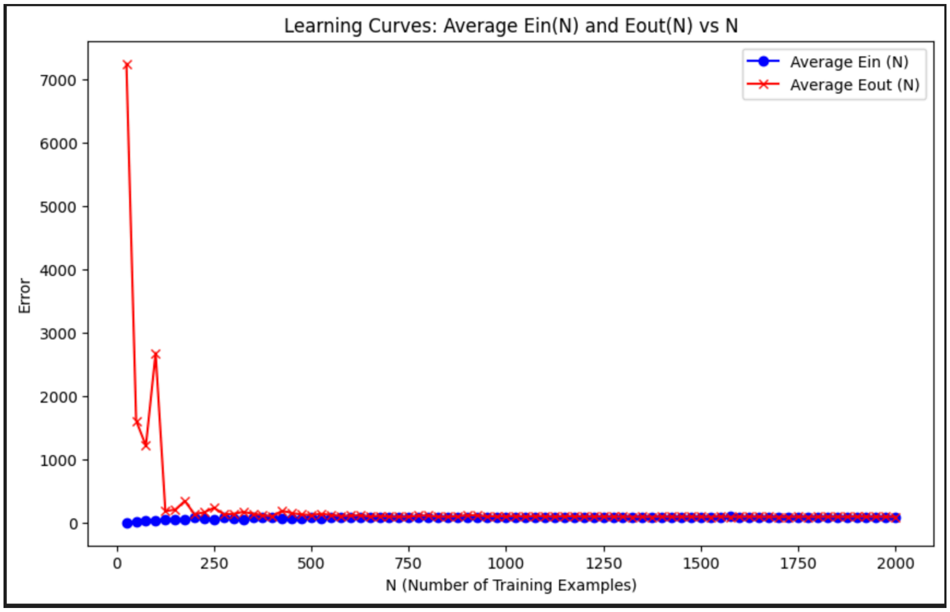

```python
def _exepriments(X, y, N, times):

    Ein_list = []
    Eout_list = []

    length = len(X)

    for _ in range(times):

        # randomly selected examples
        train_indices = np.random.choice(np.arange(length), size=N, replace=False)
        X_train, y_train = X[train_indices], y[train_indices]

        # construct the remaining examples
        test_indices = list(set(range(len(X))) - set(train_indices))
        X_test, y_test = X[test_indices], y[test_indices]

        w_lin = linear_regression(X_train, y_train)

        Ein = mean_squared_error(X_train, y_train, w_lin)
        Eout = mean_squared_error(X_test, y_test, w_lin)

        Ein_list.append(Ein)
        Eout_list.append(Eout)

    return Ein_list, Eout_list
```

```python
def linear_regression(X, y):

    X_pseudo_inv = np.linalg.pinv(X)
    w_lin = X_pseudo_inv @ y

    return w_lin


def mean_squared_error(X, y, w):

    predictions = X @ w
    errors = predictions - y

    return np.mean(errors ** 2)
```

```python
def plot_scatter(Ein_list, Eout_list):

    plt.figure(figsize=(10, 6))
    plt.scatter(Ein_list, Eout_list, alpha=0.5)
    plt.xlabel('Ein')
    plt.ylabel('Eout')
    plt.title('Ein vs Eout for Linear Regression')
    plt.grid(True)
    plt.show()
```

Problem 11

$\overline{E}_{out}$ is high and fluctuates in the beginning, I infer that it's because N is not big enough so the difference between $\overline{E}_{in}$ and $\overline{E}_{out}$ is large and not stable. But when N gets bigger, both of them converge at some level, which match what is teached in the class.

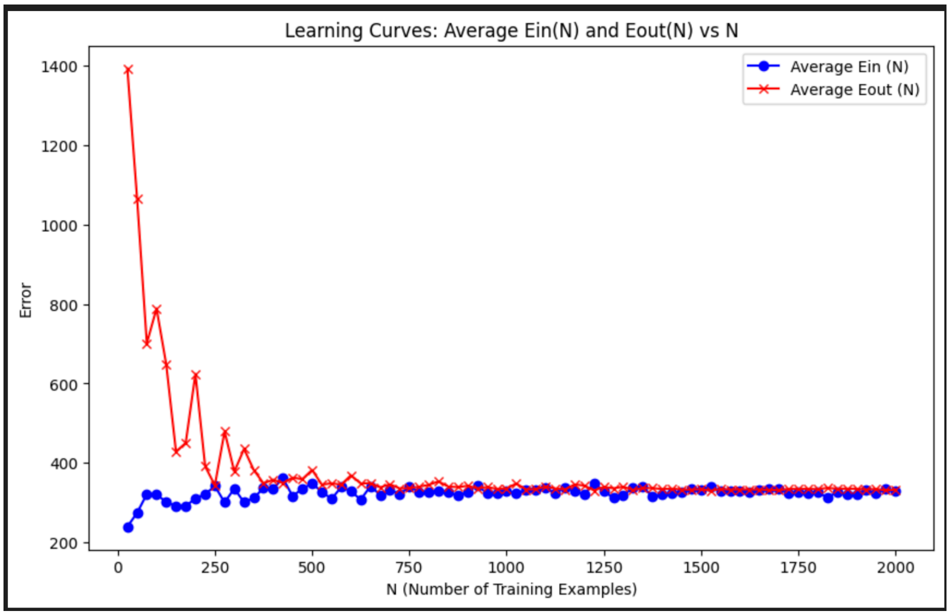



```python
def exepriments(X, y, times):

    Ein_avg_list, Eout_avg_list = [], []

    for i in range(25, 2001, 25):
        Ein_list, Eout_list = _exepriments(X, y, i, times)

        Ein_avg = sum(Ein_list) / len(Ein_list)
        Eout_avg = sum(Eout_list) / len(Eout_list)

        Ein_avg_list.append(Ein_avg)
        Eout_avg_list.append(Eout_avg)

    return Ein_avg_list, Eout_avg_list
```

```python
def plot_learning_curve(Ein_avg_list, Eout_avg_list):

    N_values = list(range(25, 2001, 25))

    plt.figure(figsize=(10, 6))
    plt.plot(N_values, Ein_avg_list, label="Average Ein (N)", color='blue', marker='o')
    plt.plot(N_values, Eout_avg_list, label="Average Eout (N)", color='red', marker='x')
    plt.xlabel("N (Number of Training Examples)")
    plt.ylabel("Error")
    plt.title("Learning Curves: Average Ein(N) and Eout(N) vs N")
    plt.legend()
    plt.show()
```

Problem 12

Compared to the previous problem, the fluctuation is larger and last longer until N becomes much more larger. I infer that it's because we only use the first 2 features, hence doesn't have enough power to deal with target complexity. But again as N gets bigger, both of them converge at some level.




```python
# problem 12
# slice the matrix
X_reduced = X[:, :3]

Ein_list, Eout_list = exepriments(X_reduced, labels, times)
plot_learning_curve(Ein_list, Eout_list)
```

Problem 13

$B(N,k)$: maximum posssible $m_H(N)$ when break point at k, it means that it can't shatter any length-k subvectors inside the length-N vector.

To shatter N points, the hypothesis set must implement all posibble 2^N dichotomies.

Now consider each term of $\sum_{i=0}^{k-1}\binom{N}{i}$ as the number of dichotomies that contain exactly i points labeled 1, that is :

- $\binom{N}{0}$: dichotomy that has no point labeled 1
- $\binom{N}{1}$: dichotomies that have exactly one point labeled 1
- $\binom{N}{2}$: dichotomies that have exactly two points labeled 1
- ...
- $\binom{N}{k-1}$: dichotomies that have exactly (k-1) points labeled 1

It is trivial that although there is a hyphotheiss set contains all dichotomies above, it can't shatter k points since it doesn't contain any dichotomy which has at least k points labeled 1, so it can't shatter any length-k subvectors inside the length-N vector.

Hence $B(N,k) \geq \sum_{i=0}^{k-1}\binom{N}{i}$ is proved, and given the proof $B(N,k) \leq \sum_{i=0}^{k-1}\binom{N}{i}$ in lecture, we can finally prove that $B(N,k) = \sum_{i=0}^{k-1}\binom{N}{i}$