ML HW1

R09922A21 郭蕙綺

1、(D) I think I can use online learning to solve this problem

Training data can be collected as: 1 month history attributes, such as views on website... and the ground truth is the predicted chance of making a purchase in the next 7 days.

We can collect data every month so that we can improve hypothesis through receiving data instances.

2. (E) $y_{n(t)} W_{t+1}^T X_{n(t)} = y_{n(t)} [W_t^T + y_{n(t)} X_{n(t)} \cdot \eta_t] X_{n(t)} > 0$

$\Rightarrow y_{n(t)} \cdot W_t^T \cdot X_{n(t)} + y_{n(t)} \cdot y_{n(t)} \cdot X_{n(t)} \cdot X_{n(t)} \cdot \eta_t > 0$

$\Rightarrow \overbrace{\cancel{y_{n(t)}} \cdot \cancel{y_{n(t)}}}^{=1} \cdot X_{n(t)} \cdot X_{n(t)} \cdot \eta_t > -y_{n(t)} \cdot W_t^T \cdot X_{n(t)}$ ($\because y_{n(t)} = \pm 1, \therefore [y_{n(t)}]^2 = 1$)

$\Rightarrow \eta_t > \dfrac{-y_{n(t)} \cdot W_t^T \cdot X_{n(t)}}{\|X_{n(t)}\|^2}$ $\therefore$ Choose (E).

3. $W_{t+1} \leftarrow W_t + y_{n(t)} X_{n(t)} \cdot \eta_t$

$W_f^T W_{t+1} = W_f^T (W_t + y_{n(t)} X_{n(t)} \cdot \eta_t)$

$\geq W_f^T W_t + \min_n y_n W_f^T X_n \eta_t$

$\|W_{t+1}\|^2 = \| W_t + y_{n(t)} X_{n(t)} \cdot \eta_t \|^2$

$= \|W_t\|^2 + 2 y_{n(t)} W_t^T X_{n(t)} \cdot \eta_t + \eta_t^2 \| y_{n(t)} X_{n(t)} \|^2$

$\leq \|W_t\|^2 + 0 + \max_n \|X_n\|^2 \cdot \eta_t^2$

Let $\min_n y_n W_f^T X_n = \rho$, $\max_n \|X_n\|^2 = R^2$

$W_f^T W_1 \geq W_f^T W_0 + \eta_0 \rho$ $\quad \|W_1\|^2 \leq \|W_0\|^2 + R^2 \cdot \eta_0^2$

$W_f^T W_2 \geq \cancel{W_f^T W_1} + \eta_1 \rho$ $\quad \|W_2\|^2 \leq \cancel{\|W_1\|^2} + R^2 \cdot \eta_1^2$

$\vdots$ $\quad \vdots$

+) $W_f^T W_T \geq \cancel{W_f^T W_{T-1}} + \eta_{T-1} \rho$ $\quad$ +) $\|W_T\|^2 \leq \cancel{\|W_{T-1}\|^2} + R^2 \eta_{T-1}^2$

$W_f^T W_T \geq \rho \cdot \sum_{t=0}^{T-1} \eta_t$ $\quad \|W_T\|^2 \leq R^2 \cdot \sum_{t=0}^{T-1} \eta_t^2$

$\dfrac{W_f^T W_T}{\|W_f\| \|W_T\|} \geq \dfrac{\rho \cdot \sum_{t=0}^{T-1} \eta_t}{R \cdot \sqrt{\sum_{t=0}^{T-1} \eta_t^2}}$

if $\dfrac{\sum_{t=0}^{T-1} \eta_t}{\sqrt{\sum_{t=0}^{T-1} \eta_t^2}}$ is strongly increasing, that W_T can be W_f

$\longrightarrow$ For (A), $\dfrac{\sum_{t=0}^{T-1} 2^{-t}}{\sqrt{\sum_{t=0}^{T-1} 2^{-2t}}} = \dfrac{1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{T-1}}}{\sqrt{1 + \frac{1}{4} + \frac{1}{16} + \cdots + \frac{1}{4^{T-1}}}}$

$= \dfrac{2 \cdot 1 \cdot (1 - \frac{1}{2^T})}{\sqrt{\frac{1 \cdot (1 - \frac{1}{4^T})}{\frac{3}{4}}}} = \sqrt{3} \cdot \dfrac{(1 - \frac{1}{2^T})}{\sqrt{(1 - \frac{1}{4^T})}}$

since $(1 - \frac{1}{4^T}) > (1 - \frac{1}{2^T})$, both of them < 1,

so, $(1 - \frac{1}{2^T}) < \sqrt{(1 - \frac{1}{4^T})}$ $\Rightarrow$ not increasing.

For (B), $\dfrac{\sum_{t=0}^{T-1} 0.6211}{\sqrt{\sum_{t=0}^{T-1} 0.6211^2}} = \dfrac{T \cdot 0.6211}{0.6211 \cdot \sqrt{T}}$

$\Rightarrow T > \sqrt{T}$ $\Rightarrow$ strongly increasing.

For (D), $\dfrac{\sum_{t=0}^{T-1} (\frac{1}{1+t})}{\sqrt{\sum_{t=0}^{T-1} (\frac{1}{1+t})^2}} = \boxed{\dfrac{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{T}}{\sqrt{1 + (\frac{1}{2})^2 + (\frac{1}{3})^2 + \cdots + (\frac{1}{T})^2}}}$

Let $\checkmark = k$, $k^2 = \dfrac{1^2 + (\frac{1}{2})^2 + \cdots + (\frac{1}{T})^2 + 2(\frac{1}{2} \cdot \frac{1}{3} + \cdots + \frac{1}{T \cdot (T-1)}}{1 + (\frac{1}{2})^2 + (\frac{1}{3})^2 + \cdots + (\frac{1}{T})^2}$

$\Rightarrow k^2 > 1 \Rightarrow k > 1$ $\Rightarrow$ strongly increasing.

For (C), (E), from problem 2, we can easily know that (C) cannot halt.

(E) can halt.

4. According to "Classol Handout" p.52.

start from $W_0 = 0$, after T mistake corrections,

$$1 \ge \frac{W_f^T W_T}{\|W_f\| \|W_T\|} \ge \frac{T\rho}{\|W_f\| \cdot \sqrt{T} R} \quad (\rho = \min_n y_n W_f^T x_n,\ R^2 = \max_n \|x_n\|^2)$$

$$\Rightarrow \|W_f\| \cdot \sqrt{T} \cdot R \ge T\rho \Rightarrow \|W_f\|^2 \cdot T \cdot R^2 \ge T^2 \rho^2 \Rightarrow T \le \frac{\|W_f\|^2 \cdot R^2}{\rho^2}$$

In this problem, $f(x) = W_f^T \cdot X$, W_f can be $[\overset{w_0}{-0.5}\ w_1\ w_2 \ldots\ w_d]$

$w_i = 1$ if w_i is spam-like

$= -1$ if w_i is not spam-like

$X_n = [\overset{x_0}{1}, x_1, x_2, x_3 \ldots x_d]$

$x_i = 1$ if word i is in email X_n

$= 0$ if word i is not in email X_n

$\rho = \min_n y_n W_f^T X_n = 0.5$ since $W_f^T X_n = -0.5 + \sum_{i=1}^{d} w_i x_i$, $y_n = \mathrm{sign}(W_f^T X_n)$

$R^2 = \max_n \|X_n\|^2 = \sqrt{\max_n (1 + x_1^2 + x_2^2 + \cdots + x_d^2)}^{\,2} = (1+m)$ since there are at most m distinct words in each email.

$\|W_f\|^2 = ((-0.5)^2 + d) = 0.25 + d$

$\therefore T \le \frac{(0.25+d)\cdot(1+m)}{(0.5)^2} = (4d+1)(m+1)$ #

5. Let W_{PLA}'s update equation: $W_{t+1} \leftarrow W_t + y_{n(t)} X_{n(t)}$

multiclass PLA update equation: $W_y^{(t+1)} \leftarrow W_y^{(t)} + X_{n(t)}$

$W_{y'}^{(t+1)} \leftarrow W_{y'}^{(t)} + X_{n(t)}$

Initial weight of W_{PLA} & multiclass PLA are 0

When we find a mistake $X_{n(t)}$,

if $y_n(t) = \begin{cases} +1 & \text{in binary PLA. update eq: } W_{t+1} \leftarrow W_t + X_n(t) \\ 2 & \text{in multiclass PLA: } W_{y=2}^{(t+1)} \leftarrow W_{y=2}^{(t)} + X_n(t) \\ & W_{y=1}^{(t+1)} \leftarrow W_{y=1}^{(t)} - X_n(t) \end{cases}$

if $y_n(t) = \begin{cases} -1 & \text{in binary PLA: } W_{t+1} \leftarrow W_t - X_n(t) \\ 1 & \text{in multiclass: } W_{y=2}^{(t+1)} \leftarrow W_{y=2}^{(t)} - X_n(t) \\ & W_{y=1}^{(t+1)} \leftarrow W_{y=1}^{(t)} + X_n(t) \end{cases}$

conclusion: Since initial weight are all 0, W_2 do the same calculation as W_{PLA}, W_1 do the opposite calculation from W_{PLA}

Thus, I choose (B) #

6 self-supervised learning is to learn 'physical knowledge' before actual tasks.

The images at similar time stamps should contain similar objects. (knowledge)

As a result, we can pair the images to train our model. (label of an image pair is the differance of time stamps.) → self-defined goal

After the model learn that knowledge, images that are taken at similar time stamps can be mapped to similar vectors.

Also, this model can be a pretrained model of other tasks.

7. Each article can belong to several different categories. → multilabel classification.

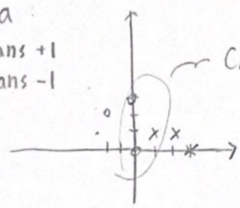
Having few (1126) labeled data → semi-supervised learning + batch learning

Each tag can be labeled as 0 (not belong to an article) or 1 (belong to an article) → raw features.

8. Data:

o means +1

x means −1

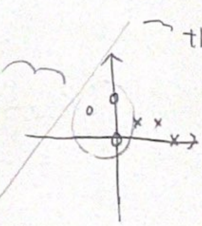


Choose these 3 examples as D. We can easily imagine that after PLA learning, a line that can seperate these 3 examples can classify the other 3 examples correctly.

So, the smallest $E_{ots}(g) = 0$.

The largest $E_{ots}(g)$ would occurred when choosing 3 examples are all lebeled as "+1" or "−1".

For example, choose 3 examples that labeled as +1:



this is initial w.

then, we don't need to update w and we would misclassify all the other examples.

So, the largest $E_{ost}(g)$ would be 1. #

9. (A) $\mathbb{E}[\hat{\theta}] = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}[|h(x_n)+y_n|]\right) = \mathbb{E}_{x\sim p}[|h(x) \neq f(x)|]$.

(B) $\mathbb{E}[\hat{\theta}] = \mathbb{E}\left(\frac{1}{N}\sum_{n=1} x_n\right)$, P is a Bernouli distribution, $\mathbb{E}(X) = \theta$

(D) $\mathbb{E}[\hat{\theta}] = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} x_n^2\right) = \mathbb{E}(X^2)$. ∵ Gaussian distribution. ∴ $\mathbb{E}(X^2) = Var(X) + [\mathbb{E}(X)]^2$

∵ P is zero-mean, ∴ $\mathbb{E}(X^2) = Var(X) + 0 = Var(X) = \theta$

(C) Let $Y = \max\{x_1, x_2, \dots, x_N\}$

$$E(Y) = \sum_{y=1}^{M} y\cdot P(Y=y) = \sum_{y=1}^{M} y\cdot P(X_1 \le y, X_2 \le y, \dots, X_n \le y) = \sum_{y=1}^{M} y\cdot\left(\frac{y}{M}\right)^n = \frac{1}{M^n}\sum_{y=1}^{M} y^{n+1}$$

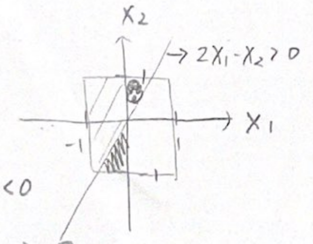
$$= \frac{1}{M^n}(1+2^{n+1}+3^{n+1}+\dots+M^{n+1}) = \frac{1+2^{n+1}+\dots+(M-1)^{n+1}}{M^n} + M \neq M \ \#$$

Choose C #

10. For $E_{out}(h_2)$, error occar when
$$\begin{cases} x_1 > 0,\ x_2 < 0 \rightarrow \frac{1}{2}\times\frac{1}{2} = \frac{1}{4} \\ x_1 < 0,\ x_2 > 0 \rightarrow \frac{1}{2}\times\frac{1}{2} = \frac{1}{4} \end{cases}$$

∴ $E_{out}(h_2) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

For $E_{out}(h_1)$, error occur when
$$\begin{cases} 2x_1 - x_2 > 0,\ x_1 < 0 \quad — ① \\ 2x_1 - x_2 < 0,\ x_1 > 0 \quad — ② \end{cases}$$



$x_1 < 0$

solution①

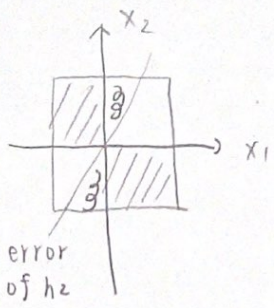
solution②

Prob: $\frac{\frac{1}{2}\times 1\times\frac{1}{2}}{4} = \frac{1}{16}$ — ①

Prob: $\frac{\frac{1}{2}\times 1\times\frac{1}{2}}{4} = \frac{1}{16}$ — ②

∴ $E_{out}(h_1) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$ #

11.



error of h_2

3 error of h_1

$E_{in}(h_2) = E_{in}(h_1)$ means the number of error is the same

When 0 error occur: $(1 - \frac{1}{2} - \frac{1}{8})^4 = \frac{81}{4096}$

When 1 error occur: $\frac{1}{8}\times\frac{1}{2}\times(1-\frac{1}{2}-\frac{1}{8})^2 \times \frac{4!}{2!} = \frac{9\times 4}{4096}\times 12 = \frac{432}{4096}$

When 2 error occur: $\frac{1}{8}\times\frac{1}{8}\times\frac{1}{2}\times\frac{1}{2}\times\frac{4!}{2!\,2!} = \frac{16\times 6}{4096} = \frac{96}{4096}$

Since there's no overlap, 3 and 4 errors are impossible

Prob: $\frac{81}{4096} + \frac{432}{4096} + \frac{96}{4096} = \frac{609}{4096}$ #

12.

number	dices (green)
1	B (5B)
2	A, B, D
3	D (5D)
4	A (5A)
5	D (5D)
6	A B C

5A, 5B, 5C, 5D can get some number are purely green. $\rightarrow (\frac{1}{4})^5 \times 4 = \frac{4}{1024}$

For number "2" is purely green:

4A
- 1B : $(\frac{1}{4})^5 \cdot \frac{5!}{4!} = \frac{5}{1024}$
- 1D : $(\frac{1}{4})^5 \cdot \frac{5!}{4!} = \frac{5}{1024}$

3A $\left(\frac{1}{4}\right)^5 \cdot$
- 2B-0D : $\frac{5!}{3!2!} = \frac{10}{1024}$
- 1B-1D : $\frac{5!}{3!} = \frac{20}{1024}$
- 0B-2D : $\frac{5!}{3!2!} = \frac{10}{1024}$

2A $\left(\frac{1}{4}\right)^5$
- 3B-0D : $\cdot \frac{5!}{2!3!} = \frac{10}{1024}$
- 2B-1D : $\frac{5!}{2!2!} = \frac{30}{1024}$
- 1B-2D : $\frac{5!}{2!2!} = \frac{30}{1024}$
- 0B-3D : $\frac{5!}{2!3!} = \frac{10}{1024}$

1A
- 4B-0D : $\frac{5}{1024}$
- 3B-1D : $\frac{20}{1024}$
- 2B-2D : $\frac{30}{1024}$
- 1B-3D : $\frac{20}{1024}$
- 0B-4D : $\frac{5}{1024}$

0A
- 4B-1D : $\frac{5}{1024}$
- 3B-2D : $\frac{10}{1024}$
- 2B-3D : $\frac{10}{1024}$
- 1B-4D : $\frac{5}{1024}$

Total: $\frac{240}{1024}$

For num "6", also 3 kinds of dices (A, B, C). ∴ similar to this situation, only minus duplicate scenario. (only have "A", "B")

Num "6" : $\frac{240-5-10-10-5}{1024} = \frac{210}{1024}$ (scenario that contains "0D")

Total: $\frac{4+240+210}{1024} = \frac{454}{1024}$ #

```python
import random
import numpy as np
from tqdm import tqdm

train_data = [np.array([1.0]+[float(x) for x in i.strip().split()]) for i in
open("hw1_train.dat").readlines()]

#Q13

W_pla = []
for i in tqdm(range(1000)):
    r_seed = i*11
    random.seed(r_seed)
    w = [0.0] *11
    w = np.array(w)
    accuracy = 0
    while True:
        data_id = random.randint(0,99)
        sign = 1.0 if np.sum(w * train_data[data_id][:11]) > 0 else -1.0
        if sign != train_data[data_id][-1]:
            w += train_data[data_id][-1]*train_data[data_id][:11]
            accuracy = 0
        else:
            accuracy += 1
        if accuracy == 500:
            W_pla.append(w)
            break
length = 0
for w_ in W_pla:
    length += np.sum(np.power(w_,2))
print(length/1000)

#Q14
for idx in range(100):
    train_data[idx][:11] = train_data[idx][:11]*2
W_pla = []
for i in tqdm(range(1000)):
    r_seed = i*11
    random.seed(r_seed)
    w = [0.0] *11
    w = np.array(w)
    accuracy = 0
    while True:
        data_id = random.randint(0,99)
        sign = 1.0 if np.sum(w * train_data[data_id][:11]) > 0 else -1.0
        if sign != train_data[data_id][-1]:
            w += train_data[data_id][-1]*train_data[data_id][:11]
            accuracy = 0
        else:
            accuracy += 1
        if accuracy == 500:
            W_pla.append(w)
            break
length = 0
for w_ in W_pla:
    length += np.sum(np.power(w_,2))
print(length/1000)

#Q15
for idx in range(100):
    l = np.sqrt(np.sum(np.power(train_data[idx][:11],2)))
    train_data[idx][:11] = train_data[idx][:11]/l
W_pla = []
for i in tqdm(range(1000)):
    r_seed = i*11
    random.seed(r_seed)
    w = [0.0] *11
    w = np.array(w)
    accuracy = 0
    while True:
        data_id = random.randint(0,99)
        sign = 1.0 if np.sum(w * train_data[data_id][:11]) > 0 else -1.0
        if sign != train_data[data_id][-1]:
            w += train_data[data_id][-1]*train_data[data_id][:11]
            accuracy = 0
        else:
            accuracy += 1
        if accuracy == 500:
            W_pla.append(w)
            break
length = 0
for w_ in W_pla:
    length += np.sum(np.power(w_,2))
print(length/1000)

#Q16
train_data = [np.array([0.0]+[float(x) for x in i.strip().split()]) for i in
open("hw1_train.dat").readlines()]

W_pla = []
for i in tqdm(range(1000)):
    r_seed = i*11
    random.seed(r_seed)
    w = [0.0] *11
    w = np.array(w)
    accuracy = 0
    while True:
        data_id = random.randint(0,99)
        sign = 1.0 if np.sum(w * train_data[data_id][:11]) > 0 else -1.0
        if sign != train_data[data_id][-1]:
            w += train_data[data_id][-1]*train_data[data_id][:11]
            accuracy = 0
        else:
            accuracy += 1
        if accuracy == 500:
            W_pla.append(w)
            break

length = 0
for w_ in W_pla:
    length += np.sum(np.power(w_,2))
print(length/1000)
```