Machine Learning Foundations HW 3

R09946006 | 何青儒 | HO, Ching-Ru | Nov 18th, 2020

---

1. $[\mathbf{b}]$ $N = 30$
   - Substitute $\sigma = 0.1$ and $d = 11$ into $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$.
   - We can get $0.06 \geq (0.1)^2 \times \left(1 - \frac{11+1}{N}\right)$, making the smallest $N = 30$.
2. $[\mathbf{a}]$ There exists at least one solution for the normal equation.
   - If $\mathrm{X}^T\mathrm{X}$ is invertible, which means that $(\mathrm{X}^T\mathrm{X})^{-1}$ exists, we can get an unique solution easily.
   - If $\mathrm{X}^T\mathrm{X}$ is singular, which means that $(\mathrm{X}^T\mathrm{X})^{-1}$ doesn't exists, we can only solve this problem by using psuedo-inverse term. There can exist many solution (non-unique), but we can usually find "one" solution by some mathematic tools or programming.
3. $[\mathbf{c}]$ multiplying each of the n-th row of X by $\frac{1}{n}$ (which is equivalent to scaling the n-th example by $\frac{1}{n}$)
   - Consider 3 conditions:
     - (1) multiply a matrix A on the left side of X (means AX) .
     - (2) multiply a matrix B on the right side of X (means XB) .
     - (3) multiply a scalar c on the left side of X (means cX).
   - (1) if new X become AX, we can get new $\mathrm{H}^* = \mathrm{AX}((\mathrm{AX})^T\mathrm{AX})^{-1}(\mathrm{AX})^T$, changed.
     $$\begin{aligned} &= \mathrm{AX}(\mathrm{X}^T\mathrm{A}^T\mathrm{AX})^{-1}\mathrm{X}^T\mathrm{A}^T \\ &= \mathrm{AXX}^{-1}\mathrm{A}^{-1}(\mathrm{A}^T)^{-1}(\mathrm{X}^T)^{-1}\mathrm{X}^T\mathrm{A}^T \\ &\neq \mathrm{H} \end{aligned}$$
   - (2) if new X become XB, we can get new $\mathrm{H}^* = \mathrm{XB}((\mathrm{XB})^T\mathrm{XB})^{-1}(\mathrm{XB})^T$,
     $$\begin{aligned} &= \mathrm{XB}(\mathrm{B}^T\mathrm{X}^T\mathrm{XB})^{-1}\mathrm{B}^T\mathrm{X}^T \\ &= \mathrm{X}(\mathrm{BB}^{-1})\mathrm{X}^{-1}(\mathrm{X}^T)^{-1}\left((\mathrm{B}^T)^{-1}\mathrm{B}^T\right)\mathrm{X}^T \\ &= \mathrm{XX}^{-1}(\mathrm{X}^T)^{-1}\mathrm{X}^T \\ &= \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \\ &= \mathrm{H} \end{aligned}$$
     unchanged.
   - (3) if new X become cX, we can get new $\mathrm{H}^* = \mathrm{cX}((\mathrm{cX})^T\mathrm{cX})^{-1}(\mathrm{cX})^T$, unchanged.
     $$\begin{aligned} &= \mathrm{cX}(\mathrm{cX}^T\mathrm{cX})^{-1}\mathrm{cX}^T \\ &= \mathrm{c}^2\mathrm{X}(\mathrm{c}^2\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \\ &= \mathrm{c}^2\frac{1}{\mathrm{c}^2}\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \\ &= \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \\ &= \mathrm{H} \end{aligned}$$
   - $[\mathbf{a}]$ equivalent to the third condition cX where $\mathrm{c} = 2$, making $\mathrm{H}^* = \mathrm{H}$, so H is unchanged.
   - $[\mathbf{b}]$ equivalent to the second condition XB where $\mathrm{B} = \begin{bmatrix} 1 & & \cdots & 0 \\ & 2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & i \end{bmatrix}$, making $\mathrm{H}^* = \mathrm{H}$, so H is unchanged.
   - $[\mathbf{c}]$ equivalent to the second condition AX where $\mathrm{B} = \begin{bmatrix} 1 & & \cdots & 0 \\ & 2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & i \end{bmatrix}$, making $\mathrm{H}^* \neq \mathrm{H}$, so H is changed.

- [**d**] equivalent to the second condition XB where $\mathrm{B} = \begin{bmatrix} 1 & & \cdots & 0 \\ & 2 & & \\ 1 & & \ddots & \vdots \\ 1 & & \cdots & i \end{bmatrix}$, the first column n-th row becomes 1 when random chosen n-th one, making $\mathrm{H}^* = \mathrm{H}$, so H is unchanged.

4. [**e**] 4
   - $f(y) = \begin{cases} \theta, & y_n = 1 \\ 1-\theta, & y_n = 0 \end{cases}$, we can see $\nu = \frac{1}{N}\sum_{n=1}^{N} y_n$ as E_{out} due to ν is the expected value of sample space, and see θ as E_{in} due to $E_{in}(\theta) = \mathbb{P}(\text{Head}) \times 1 + \mathbb{P}(\text{Tail}) \times 0 = \theta$ theoretically.
   - The first statement is true due to Hoeffding's inequality we've seen before.
   - The second statement is true due to the definition of likelihood method.
   - The third statement is true due to the definition of E_{in} we've seen before.
   - The forth statement is true because we can get the gradient $(-\nabla E_{in} = \frac{2}{N}\sum_{n=1}^{N} y_n = 2\nu)$ by the third statement.
5. [**a**] $\left(\frac{1}{\hat{\theta}}\right)^N$
   - $Y \sim \text{Uni}(0, \theta)$, the probability dense function (pdf) is $f(y|\theta) = \begin{cases} \frac{1}{\theta}, & y \in [0, \theta] \\ 0, & \text{otherwise} \end{cases}$
   - $\mathcal{L}(\theta) = \prod_y f(y|\theta) = (\frac{1}{\theta})(\frac{1}{\theta})\cdots(\frac{1}{\theta}) = \left(\frac{1}{\theta}\right)^N$, since $\hat{\theta} \geq \max(y_1, y_2, \cdots, y_N)$, ensure $\hat{\theta}$ is the biggest θ. Thus, the likelihood of an uniform distribution is $\left(\frac{1}{\hat{\theta}}\right)^N$.
6. [**b**] $\text{err}(\mathbf{w}, \mathbf{x}, y) = \max(0, -y\mathbf{w}^T\mathbf{x})$
   - Consider when $y_n = 1$:
     - When $\text{sign}(\mathbf{w}_t^T\mathbf{x}_n) = 1$ (classification), then $\left(y_n\mathbf{w}_t^T\mathbf{x}_n\right) > 1$, $\text{err}(\mathbf{w}, \mathbf{x}, y) = 0$
     - When $\text{sign}(\mathbf{w}_t^T\mathbf{x}_n) = -1$ (misclassification), then $\left(y_n\mathbf{w}_t^T\mathbf{x}_n\right) < 1$, $\text{err}(\mathbf{w}, \mathbf{x}, y) = -y\mathbf{w}^T\mathbf{x}$
   - Consider when $y_n = -1$:
     - When $\text{sign}(\mathbf{w}_t^T\mathbf{x}_n) = 1$ (misclassification), then $\left(y_n\mathbf{w}_t^T\mathbf{x}_n\right) < 1$, $\text{err}(\mathbf{w}, \mathbf{x}, y) = -y\mathbf{w}^T\mathbf{x}$
     - When $\text{sign}(\mathbf{w}_t^T\mathbf{x}_n) = -1$ (classification), then $\left(y_n\mathbf{w}_t^T\mathbf{x}_n\right) > 1$, $\text{err}(\mathbf{w}, \mathbf{x}, y) = 0$
   - Thus, $\text{err}(\mathbf{w}, \mathbf{x}, y) = \max(0, -y\mathbf{w}^T\mathbf{x})$
7. [**a**] $+y_n\mathbf{x}_n \exp(-y_n\mathbf{w}^T\mathbf{x}_n)$
   - From $\nabla\text{err}_{exp}(\mathbf{w}, \mathbf{x}, y) = \left(\frac{\partial \text{err}_{exp}}{\partial \mathbf{w}}\right) = \left(\frac{\partial \exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{\partial \mathbf{w}}\right) = -y_n\mathbf{x}_n \exp(-y_n\mathbf{w}^T\mathbf{x}_n)$
   - We can get $-\nabla\text{err}_{exp}(\mathbf{w}, \mathbf{x}, y) = y_n\mathbf{x}_n \exp(-y_n\mathbf{w}^T\mathbf{x}_n)$
8. [**b**] $-(\mathrm{A}_E(\mathbf{u}))^{-1}\mathbf{b}_E(\mathbf{u})$
   - $$\begin{aligned} E(\mathbf{w}) &\simeq E(\mathbf{u}) + \mathbf{b}_E(\mathbf{u})^T(\mathbf{w}-\mathbf{u}) + \frac{1}{2}(\mathbf{w}-\mathbf{u})^T\mathrm{A}_E(\mathbf{u})(\mathbf{w}-\mathbf{u}) \\ &= E(\mathbf{u}) + \mathbf{b}_E(\mathbf{u})^T(\mathbf{w}-\mathbf{u}) + \frac{1}{2}\left(\mathbf{w}^T\mathrm{A}_E(\mathbf{u})\mathbf{w} - 2\mathbf{w}^T\mathrm{A}_E(\mathbf{u})\mathbf{u} + \mathbf{u}^T\mathrm{A}_E(\mathbf{u})\mathbf{u}\right) \end{aligned}$$
   - Try to minimized: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = 0 + \mathbf{b}_E(\mathbf{u}) + \frac{2}{2}\mathrm{A}_E(\mathbf{u})\mathbf{w} - \frac{2}{2}\mathrm{A}_E(\mathbf{u})\mathbf{u} = 0$
   - We can get $\mathbf{b}_E(\mathbf{u}) + \mathrm{A}_E(\mathbf{u})(\mathbf{w}-\mathbf{u}) = 0,\ \ \mathbf{w} = \mathbf{u} - (\mathrm{A}_E(\mathbf{u}))^{-1}\mathbf{b}_E(\mathbf{u})$, and finally get $\mathbf{v} = -(\mathrm{A}_E(\mathbf{u}))^{-1}\mathbf{b}_E(\mathbf{u})$.
9. [**b**] $\frac{2}{N}\mathrm{X}^T\mathrm{X}$
   - From linear regression's $E_{in}(\mathbf{w}) = \frac{1}{N}||\mathrm{X}\mathbf{w} - \mathrm{y}||^2 = \frac{1}{N}(\mathbf{w}^T\mathrm{X}^T\mathrm{X}\mathbf{w} - 2\mathbf{w}^T\mathrm{X}^T\mathrm{y} + \mathrm{y}^T\mathrm{y})$
   - Since $\mathbf{b}_E(\mathbf{w}) = \nabla E_{in}(\mathbf{w}) = \frac{2}{N}(\mathrm{X}^T\mathrm{X}\mathbf{w} - \mathrm{X}^T\mathrm{y})$
   - Thus $\mathrm{A}_E(\mathbf{w}) = \nabla^2 E_{in}(\mathbf{w}) = \frac{\partial \mathbf{b}_E(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{N}(\mathrm{X}^T\mathrm{X})$
   - For any given $\mathbf{w} = \mathbf{w}_t$, the result holds.
10. [**b**] $(h_k(\mathbf{x}) - [y = k])x_i$

- $\frac{\partial \mathrm{err}(\mathrm{W},\mathbf{x},y)}{\partial \mathrm{W}_{ik}} = \frac{\partial(-\ln h_y(\mathbf{x}))}{\partial \mathrm{W}_{ik}} = \frac{\partial\left(-\sum_{k=1}^{K}[y=k]\ln h_k(\mathbf{x})\right)}{\partial \mathrm{W}_{ik}} = \frac{-[y=k]\cdot\left(\sum_{k=1}^{K}\partial \ln h_k(\mathbf{x})\right)}{\partial \mathrm{W}_{ik}}$
  $= \frac{-[y=k]\cdot\left(\sum_{k=1}^{K}\partial(\ln \exp(\mathbf{w}_y^T\mathbf{x}) - \ln\sum_{i=1}^{N}\exp(\mathbf{w}_i^T\mathbf{x})\right)}{\partial \mathrm{W}_{ik}} = \frac{-[y=k]\cdot\left(\sum_{k=1}^{K}\partial(\ln \exp(\mathbf{w}_y^T\mathbf{x}) - \sum_{i=1}^{N}\ln\exp(\mathbf{w}_i^T\mathbf{x})\right)}{\partial \mathrm{W}_{ik}}$
  $= -\left([y=k] - \left(\frac{\exp\,(\mathbf{w}_y^T\mathbf{x})}{\sum_{i=1}^{K}\exp(\mathbf{w}_i^T\mathbf{x})}\right)\right)x_i = -([y=k] - h_k(\mathbf{x}))\,x_i = (h_k(\mathbf{x}) - [y=k])x_i$

11. [**e**] $\mathbf{w}_2^* - \mathbf{w}_1^*$
    - Since $K=2$, there are only two classes, turning the problem become binary classification, $y \in \{0,1\}$. Noticed that $\theta(s) = \frac{1}{1+\exp(-s)}$ is sigmoid function in below.
    - While $y_n = 1$, $y_n' = 2y_n - 3 = -1$, making
      $h_1(\mathbf{x}) = \mathbb{P}(\text{classification} = -1|\,\mathbf{x},\mathbf{w}) = \frac{\exp(\mathbf{w}_y^T\mathbf{x})}{\sum_{i=1}^{K}\exp(\mathbf{w}_i^T\mathbf{x})}$.
      $\overset{K=2}{=\!=\!=} \frac{\exp(\mathbf{w}_1^{*T}\mathbf{x})}{\exp(\mathbf{w}_1^{*T}\mathbf{x})+\exp(\mathbf{w}_2^{*T}\mathbf{x})} = \frac{1}{1+\exp\left((\mathbf{w}_2^{*T}-\mathbf{w}_1^{*T})\cdot\mathbf{x}\right)} = \theta(-\mathbf{w}_y^T\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}_y^T\mathbf{x})}$
    - While $y_n = 2$, $y_n' = 2y_n - 3 = 1$, making
      $h_2(\mathbf{x}) = \mathbb{P}(\text{classification} = 1|\,\mathbf{x},\mathbf{w}) = \frac{\exp(\mathbf{w}_y^T\mathbf{x})}{\sum_{i=1}^{K}\exp(\mathbf{w}_i^T\mathbf{x})}$.
      $\overset{K=2}{=\!=\!=} \frac{\exp(\mathbf{w}_2^{*T}\mathbf{x})}{\exp(\mathbf{w}_1^{*T}\mathbf{x})+\exp(\mathbf{w}_2^{*T}\mathbf{x})} = \frac{1}{1+\exp\left((\mathbf{w}_1^{*T}-\mathbf{w}_2^{*T})\cdot\mathbf{x}\right)} = \theta(\mathbf{w}_y^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}_y^T\mathbf{x})}$
    - Finally, check the summation of probability, where
      $\mathbb{P}(\text{classification} = -1) + \mathbb{P}(\text{classification} = 1) = h_1(\mathbf{x}) + h_2(\mathbf{x}) = \theta(-\mathbf{w}_y^T\mathbf{x}) + \theta(\mathbf{w}_y^T\mathbf{x}) = 1.$
    - Due to the exponential term in the denominator, $\mathbf{w}_y^T = \mathbf{w}_2^* - \mathbf{w}_1^*$.

12. [**e**] $[-7, 0, 0, 2, -2, 3]$
    - Find the answer by the code below:

```
import numpy as np

# xi = [x1, x2, yi]
x1 = [0, 1, -1]
x2 = [1, -0.5, -1]
x3 = [-1, 0, -1]
x4 = [-1, 2, +1]
x5 = [2, 0, +1]
x6 = [1, -1.5, 1]
x7 = [0, -2, 1]
X = [x1, x2, x3, x4, x5, x6, x7]

a = [-9, -1, 0, 2, -2, 3]
b = [-5, -1, 2, 3, -7, 2]
c = [9, -1, 4, 2, -2, 3]
d = [2, 1, -4, -2, 7, -4]
e = [-7, 0, 0, 2, -2, 3]
choice = [a, b, c, d, e]
choiceName = ["a", "b", "c", "d", "e"]

def phi(x):
return [1, x[0], x[1], x[0]*x[0], x[0]*x[1], x[1]*x[1]]

print("Notaion: predict | reality | result")
print("===================")
for j in range(len(choice)):
    print("For choice", choiceName[j], ":", choice[j])
    ct = 0
    for i in range(len(X)):
        if np.sign(np.dot(chosen[j], phi(X[i]))) == X[i][2]:
            print(int(np.sign(np.dot(chosen[j], phi(X[i])))),"|", X[i][2], "| ○")
```

```
            ct = ct + 1
        else:
            print( int(np.sign(np.dot(chosen[j], phi(X[i])))),"|", X[i][2],"| ×")
    print("=================== find ", ct , "/ 7 correct in choice", choiceName[j])
```

- Output:

```
Notaion: predict | reality | result
===================
For choice a : [-9, -1, 0, 2, -2, 3]
-1 | -1 | ◯
-1 | -1 | ◯
-1 | -1 | ◯
1 | 1 | ◯
-1 | 1 | ×
1 | 1 | ◯
1 | 1 | ◯
=================== find  6 / 7 correct in choice a
For choice b : [-5, -1, 2, 3, -7, 2]
-1 | -1 | ◯
0 | -1 | ×
-1 | -1 | ◯
1 | 1 | ◯
1 | 1 | ◯
1 | 1 | ◯
-1 | 1 | ×
=================== find  5 / 7 correct in choice b
For choice c : [9, -1, 4, 2, -2, 3]
1 | -1 | ×
1 | -1 | ×
1 | -1 | ×
1 | 1 | ◯
1 | 1 | ◯
1 | 1 | ◯
1 | 1 | ◯
=================== find  4 / 7 correct in choice c
For choice d : [2, 1, -4, -2, 7, -4]
-1 | -1 | ◯
-1 | -1 | ◯
-1 | -1 | ◯
-1 | 1 | ×
-1 | 1 | ×
-1 | 1 | ×
-1 | 1 | ×
=================== find  3 / 7 correct in choice d
For choice e : [-7, 0, 0, 2, -2, 3]
-1 | -1 | ◯
-1 | -1 | ◯
-1 | -1 | ◯
1 | 1 | ◯
1 | 1 | ◯
1 | 1 | ◯
1 | 1 | ◯
=================== find  7 / 7 correct in choice e
```

13. [**b**] $2(\log_2 d + 1)$

   - For any $\Phi_{(k)}(\mathbf{x}) = (1, x_k)$ is equivalent to the decision stump where $m_{\mathcal{H}}(N) = 2N$.
   - If there have d transform function (here, meaning $\Phi_{(k)}(\mathbf{x}) = (1, x_k)$), $m_{\mathcal{H}}(N) = 2Nd$.
   - From the definition of d_{vc},

     $2^N \leq 2Nd$
     $\Rightarrow \frac{2^N}{2} \leq Nd$
     $\Rightarrow 2^{N-1} \leq Nd$
     $\Rightarrow N - 1 \leq \log_2 N + \log_2 d < \frac{N}{2} + \log_2 d \text{ (by hints, } \log_2 a < \frac{a}{2} \text{ for } N)$
     $\Rightarrow 2N - 2 < N + 2\log_2 d$
     $\Rightarrow N < 2\log_2 d + 2 = 2(\log_2 d + 1)$

14. [**d**] 0.60

15. [**c**] 1800

16. [**c**] 0.56

17. [**b**] 0.50

18. [**a**] 0.32

19. [**b**] 0.36

20. [**d**] 0.44