Machine Learning Foundations HW 1

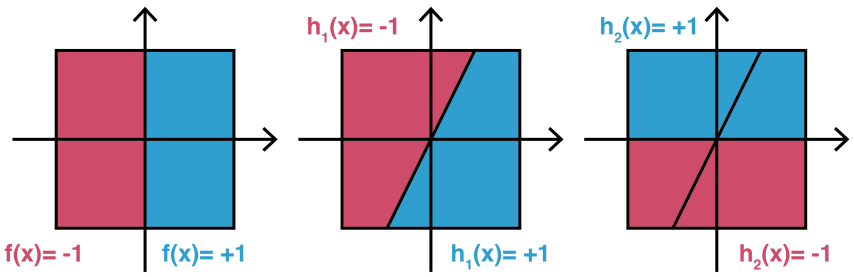
R09946006 | 何青儒 | HO, Ching-Ru | Oct 14th, 2020

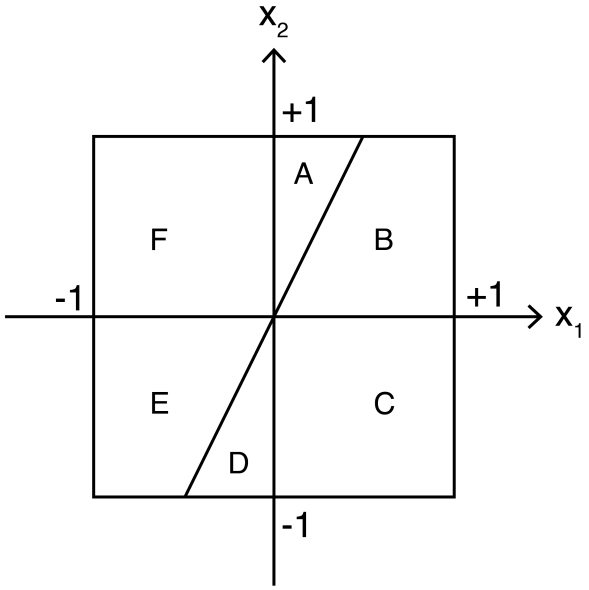
---

1. [**d**] ranking mango images by the quality of the mangoes
    - Assume that we have enough associated mango images with different rankings, we can use supervised learning if we have already classified the label and find the different features of different rankings, e.g., a premium mango may have beautiful appearance or color. This makes us to figure out some method about ranking algorithm.
2. [**e**] get a data set that contains spams and non-spams, for all words in the data set, let the machine decide its “spam score”; sum the score up for each email; let the machine optimize a threshold that achieves the best precision of spam detection; classify the email as a spam iff the email is of score more than the threshold
    - [**a**]: Only depends on coins, which is meaningless to spam email classifier.
    - [**b**]: Different people might have different idea about spam emails. It's not objective and common.
    - [**c**]: The original data set only contain 3 people, which is not objective and common, too.
    - [**d**]: "10 words" in this condition seems strange. It's not robust.
    - [**e**]: Follow idea in [**d**], but if we set a threshold and let the machine decide it, it's more objective than "10 words" which set in choice [**d**].
3. [**d**] unchanged
    - From the concept: Dr. Short only scales down the value in $\mathbf{x}_n$ rather than scales down the dimension in $\mathbf{x}_n$. It have no effect to change the worst-case speed.
    - From math proof:
        - From $\mathbf{w}_{t+1} := \mathbf{w}_t + \eta y_{n(t)}\mathbf{x}_{n(t)}$, where $\eta = \frac{1}{4}$, we can define $\tilde{\mathbf{x}}_{n(t)} = \eta \mathbf{x}_{n(t)} = \frac{1}{4}\mathbf{x}_{n(t)}$.
        - $\tilde{R}^2 = \max ||\tilde{\mathbf{x}}_n^2|| = \max ||\frac{1}{4}\mathbf{x}_n^2|| = \frac{1}{16}\max ||\mathbf{x}_n^2|| = \frac{1}{16}R^2$.
        - $\tilde{\rho} = \min y_n \frac{T_f^t}{||w_f||}\tilde{\mathbf{x}_n} = \min y_n \frac{T_f^t}{||w_f||}\frac{1}{4}\mathbf{x}_n = \frac{1}{4}\min y_n \frac{T_f^t}{||w_f||}\mathbf{x}_n = \frac{1}{4}\rho$.
        - After scaling, the worst-case speed of PLA becomes $\tilde{T} \leq \frac{\tilde{R}^2}{\tilde{\rho}^2} = \frac{(1/16)R^2}{(1/16)\rho^2} = \frac{R^2}{\rho^2}$, where can proof $\tilde{T} = T \leq \frac{R^2}{\rho^2}$.
4. [**c**] 2
    - We know $\mathbf{w}_{t+1} := \mathbf{w}_t + \eta y_{n(t)}\mathbf{x}_{n(t)} = \mathbf{w}_t + y_{n(t)}\frac{\mathbf{x}_{n(t)}}{||\mathbf{x}_{n(t)}||}$.
    - From $||\mathbf{w}_{t+1}||^2 = ||\mathbf{w}_t + y_{n(t)}\frac{\mathbf{x}_{n(t)}}{||\mathbf{x}_{n(t)}||}||^2 \leq ||\mathbf{w}_t||^2 + ||y_{n(t)}\frac{\mathbf{x}_{n(t)}}{||\mathbf{x}_{n(t)}||}||^2 \leq ||\mathbf{w}_t||^2 + \max ||y_n \frac{\mathbf{x}_n}{||\mathbf{x}_n||}||^2$, we get $||\mathbf{w}_T||^2 \leq T(\max ||y_n \frac{\mathbf{x}_n}{||\mathbf{x}_n||}||^2)$.
    - From another definition, $\mathbf{w}_f^T\mathbf{w}_{t+1} = \mathbf{w}_f^T(\mathbf{w}_t + y_{n(t)}\frac{\mathbf{x}_{n(t)}}{||\mathbf{x}_n(t)||}) = \mathbf{w}_f^T\mathbf{w}_t + y_{n(t)}\mathbf{w}_f^T\frac{\mathbf{x}_{n(t)}}{||\mathbf{x}_n(t)||}) \geq \mathbf{w}_f^T\mathbf{w}_t + \min y_n \mathbf{w}_f^T(\frac{\mathbf{x}_n}{||\mathbf{x}_n||})$, we get $\mathbf{w}_f^T\mathbf{w}_T \geq Ty_n||\mathbf{w}_f||\min \frac{\mathbf{w}^T + f\mathbf{x}_n}{||\mathbf{x}_n||\cdot||\mathbf{w}_f||}$.
    - $1 \geq \cos\theta_T = \frac{\mathbf{w}_f^T\mathbf{w}_T}{||\mathbf{w}_f^T||\cdot||\mathbf{w}_T||} \geq \frac{Ty_n||\mathbf{w}_f||\min \frac{\mathbf{w}^T + f\mathbf{x}_n}{||\mathbf{x}_n||\cdot||\mathbf{w}_f||}}{||\mathbf{w}_f^T||\sqrt{T}\sqrt{\max ||y_n \frac{\mathbf{x}_n}{||\mathbf{x}_n||}||^2}}$, and we know
    $1 \geq \frac{Ty_n||\mathbf{w}_f||\min \frac{\mathbf{w}^T + f\mathbf{x}_n}{||\mathbf{x}_n||\cdot||\mathbf{w}_f||}}{||\mathbf{w}_f^T||y_n\sqrt{T}} = \frac{T\hat{\rho}}{\sqrt{T}} \geq \frac{Ty_n||\mathbf{w}_f||\min \frac{\mathbf{w}^T + f\mathbf{x}_n}{||\mathbf{x}_n||\cdot||\mathbf{w}_f||}}{||\mathbf{w}_f^T||\sqrt{T}\sqrt{\max ||y_n \frac{\mathbf{x}_n}{||\mathbf{x}_n||}||^2}}$, then move ρ to LHS, $\frac{1}{\hat{\rho}} \geq \sqrt{T}$, thus $\hat{\rho}^{-2} \geq T$.
    - $p = -2$, and $\hat{\rho}^{-2}$ becomes the upper bound of T.
5. [**d**] $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)} \cdot \lfloor \frac{-y_{(n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)}}{||\mathbf{x}_{n(t)}||^2} + 1 \rfloor$

- From $\mathbf{w}_{t+1} := \mathbf{w}_t + \eta_{n(t)} y_{n(t)} \mathbf{x}_{n(t)}$, we can multiply $y_{n(t)}$ to LHS/RHS's left side. And multiply $\mathbf{x}_{n(t)}$ to both right side.
- Then we can get $y_{n(t)} \mathbf{w}_{t+1} \mathbf{x}_{n(t)} := y_{n(t)} \mathbf{w}_t \mathbf{x}_{n(t)} + \eta_{n(t)} y_{n(t)} y_{n(t)} \mathbf{x}_{n(t)} \mathbf{x}_{n(t)} = y_{n(t)} \mathbf{w}_t \mathbf{x}_{n(t)} + \eta_{n(t)} ||\mathbf{x}_{n(t)}||^2 > 0$, which given in problem.
- Now, because $y_{n(t)} \mathbf{w}_t \mathbf{x}_{n(t)} + \eta_{n(t)} ||\mathbf{x}_{n(t)}||^2 > 0$, we can keep $\eta_{n(t)}$ in LHS, and move others to the RHS.
- Thus, $\eta_{n(t)} > \frac{y_{n(t)} \mathbf{w}_t \mathbf{x}_{n(t)}}{||\mathbf{x}_{n(t)}||^2}$, chose [**d**] which is larger (at least plus 1 and find the floor) than lower bound of $\eta_{n(t)}$.

6. [**d**] 4
   - From $\mathbf{w}_{t+1} := \mathbf{w}_t + \eta_t \cdot y_{n(t)} \mathbf{x}_{n(t)}$, we can define different η_t in problem 5's choices.
   - [**a**] and [**b**] have the constant η_t, they only rescale the vector $\mathbf{w}$ and $\mathbf{x}$ but never change the sign of the prediction, so both choices will halt due to the proof in *"notes about convergence of PLA"*.
   - [**c**] and [**d**] have the positive η_t. Because we only change "perfect line" when $\text{sign}(\mathbf{w}_t^T \mathbf{x}_{n(t)}) \neq y_{n(t)}$ and make $y_{n(t)} \mathbf{w}_t^T \mathbf{x}_{n(t)} \leq 0$, it doesn't change the sign even we multiply a positive η_t. So both choices will still halt.
   - However, [**e**] has a negative η_t, this will remain the sign though we find a mistake. So this choice might have some problem to find the "perfect line".
   - Thus, only choice [**a**], [**b**], [**c**], [**d**] will halt in someday.
7. [**e**] reinforcement learning
   - "They get the feedback from the judge environment" means they learn from exploration and exploitation, which is similar to reinforcement learning.
8. [**b**] structured learning, semi-supervised learning, batch learning, raw features
   - We have to classify steering, braking, and signaling-before-turning, so it might be structured learning.
   - We have only 100 hours of labeled video of 200 hours of fully video, so it might be semi-supervised learning.
   - We have get all videos we needed to use before training beginning, so it might be batch learning rather than online learning.
   - We only have the raw videos and have to detect the pixels in them, so it might be raw features.
9. [**e**] $(0, 1)$
   - No matter use PLA or human learning algorithm to find the $h(x)$, we might get a full-correct on test data set ($\mathcal{U} \diagdown \mathcal{D}$), which means all $h(x) = y$, making $E_{out}(h) = \frac{1}{3} \cdot 0 = 0$. On the other hand, if we get a full-error prediction, which means all $h(x) \neq y$, making $E_{out}(h) = \frac{1}{3} \cdot 3 = 1$.
10. [**b**] $\frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$
    - From $\mathbb{P}(|\nu - \mu|) = \delta \leq 2 \exp(-2\epsilon^2 N)$, we can get $\frac{1}{2\epsilon^2} \ln \frac{2}{\delta} \leq N$.
11. [**c**] $\frac{1}{32}$
    - The combination of example are $[-1, -1]^T, [-1, 1]^T, [1, -1]^T, [1, 1]^T$, and only we sample $[-1, -1]$ and $[1, 1]^T$ can make $[h(\mathbf{x}_n) \neq y_n]$. The probability is $\frac{1}{2}$ for each time, and five time will be $(\frac{1}{2})^5 = \frac{1}{32}$.
12. [**d**] $\frac{3843}{32768}$
    -





Area	$f(x)$	$h_1(x)$	$h_2(x)$	$\mathbb{P}$
A	$+1$	$-1, \neq f(x)$	$+1, = f(x)$	$1/16$
B	$+1$	$+1, = f(x)$	$+1, = f(x)$	$3/16$
C	$+1$	$+1, = f(x)$	$-1, \neq f(x)$	$1/4$
D	-1	$+1, \neq f(x)$	$-1, = f(x)$	$1/16$
E	-1	$-1, = f(x)$	$-1, = f(x)$	$3/16$
F	-1	$-1, = f(x)$	$+1, \neq f(x)$	$1/4$

Condition	Area	$\mathbb{P}$
$f(x) = h_1(x)$, and $f(x) = h_2(x)$	**B, E**	$\frac{3}{16} + \frac{3}{16} = \frac{3}{8}$
$f(x) = h_1(x)$, but $f(x) \neq h_2(x)$	**C, F**	$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
$f(x) \neq h_1(x)$, but $f(x) = h_2(x)$	**A, D**	$\frac{1}{16} + \frac{1}{16} = \frac{2}{16}$

- To ensure $E_{in}(h_1) = E_{in}(h_2)$, $h_1(x)$ and $h_2(x)$ should have same "trouble (or, error)". They can all correct (which mean $f(x) = h_1(x) = h_2(x)$), the probability is $(\frac{3}{8})^5$, or both have one mistake with probability $(\frac{3}{8})^3(\frac{1}{2})(\frac{1}{8})(\frac{5!}{3!1!1!})$, or both have two mistakes with probability $(\frac{3}{8})(\frac{1}{2})^2(\frac{1}{8})^2(\frac{5!}{1!2!2!})$ (noticed that they can't make more than 2 mistakes).
- Sum of the probability = $(\frac{3}{8})^5 + (\frac{3}{8})^3(\frac{1}{2})(\frac{1}{8})(\frac{5!}{3!1!1!}) + (\frac{3}{8})(\frac{1}{2})^2(\frac{1}{8})^2(\frac{5!}{1!2!2!}) = \frac{3843}{32768}$

13. [**b**] $C = d$

- $h_i(\mathbf{x}) = \begin{cases} \text{sign}(x_i), i \in \{1, d\} \\ -\text{sign}(x_i - d), i \in \{d+1, 2d\} \end{cases}$.
- $h_1(\mathbf{x}) = \text{sign}(x_1)$ have same BAD DATA with $h_{d+1}(\mathbf{x}) = -\text{sign}(x_1)$.
- $h_2(\mathbf{x}) = \text{sign}(x_2)$ have same BAD DATA with $h_{d+2}(\mathbf{x}) = -\text{sign}(x_2)$.
- $\vdots$ (omitted)
- $h_d(\mathbf{x}) = \text{sign}(x_d)$ have same BAD DATA with $h_{2d}(\mathbf{x}) = -\text{sign}(x_d)$.

- Thus, $\mathbb{P}(\text{BAD } \mathcal{D} \text{ for } \mathcal{H})$.
  $\leq \mathbb{P}(\text{BAD } \mathcal{D} \text{ for } h_1) + \cdots + \mathbb{P}(\text{BAD } \mathcal{D} \text{ for } h_{2d})$
  $= \mathbb{P}(\text{BAD } \mathcal{D} \text{ for } h_1) + \cdots + \mathbb{P}(\text{BAD } \mathcal{D} \text{ for } h_d)$
  $\leq 2\exp(-2\epsilon N) + \cdots + 2\exp(-2\epsilon N)$
  $= d \cdot 2\exp(-2\epsilon N)$
- Thus $C = d$.

14. [**d**] five green 4's
    - The "green 3" can be come from dice B and D.
    - [**a**] The "green 1" is impossible found in any dice.
    - [**b**] The "orange 2" can be come from dice C.
    - [**c**] The "green 2" can be come from dice A, B, and D.
    - [**d**] The "green 4" can be come from dice A, B.
    - [**e**] The "green 5" can be come form dice D.
    - Both probability of "five green 4" and "five green 4" are $(\frac{1}{4} + \frac{1}{4})^5$
15. [**c**] $\frac{274}{1024}$
    - Notation: $(number, dice)$= if we want find a *number* is all green in the *dice*.
      - $\mathbb{P}(1, \emptyset) = 0$
      - $\mathbb{P}(2, A \vee B \vee D)$
        $= \mathbb{P}(\text{全 } A) + \mathbb{P}(\text{全 } B) + \mathbb{P}(\text{全 } D)$
        $+ \mathbb{P}(\text{僅 } AB) + \mathbb{P}(\text{僅 } BD) + \mathbb{P}(\text{僅 } AD) + \mathbb{P}(\text{僅 } ABD)$
      - $\mathbb{P}(3, B \vee D) = \mathbb{P}(\text{全 } B) + \mathbb{P}(\text{全 } D) + \mathbb{P}(\text{僅 } BD)$
      - $\mathbb{P}(4, A \vee B) = \mathbb{P}(\text{全 } A) + \mathbb{P}(\text{全 } B) + \mathbb{P}(\text{僅 } AB)$
      - $\mathbb{P}(5, D) = \mathbb{P}(\text{全 } D)$
      - $\mathbb{P}(6, A \vee C) = \mathbb{P}(\text{全 } A) + \mathbb{P}(\text{全 } C) + \mathbb{P}(\text{僅 } AC)$
    - Compute the union set, $\mathbb{P}(\text{some number that is purely green})$
      $= \mathbb{P}(\text{全 } A) + \mathbb{P}(\text{全 } B) + \mathbb{P}(\text{全 } C) + \mathbb{P}(\text{全 } D)$
      $+ \mathbb{P}(\text{僅 } AC) + \mathbb{P}(\text{僅 } AB) + \mathbb{P}(\text{僅 } BD) + \mathbb{P}(\text{僅 } AD)$
      $+ \mathbb{P}(\text{僅 } ABD)$
      $= (1 + 1 + 1 + 1 + 30 + 30 + 30 + 30 + 150)/1024 = 274/1024$
16. [**b**] 11
    - I get $\text{med}(update\_times) = 11$ times in my program.
17. [**b**] -5
    - I get $\text{med}(w_o) = -7$ in my program.
18. [**d**] 17
    - I get $\text{med}(update\_times) = 16$ times in my program.
19. [**d**] 17
    - I get $\text{med}(update\_times) = 17$ times in my program.
20. [**d**] 17
    - I get $\text{med}(update\_times) = 16$ times in my program.