Machine Learning Foundations HW 4

R09946006 | 何青儒 | HO Ching-Ru | Dec 4th, 2020

1. [**c**] $\left|e^x - \left(\frac{3+3e^2}{8}\right)x\right|$

   - The magnitude of deterministic noise is the **area** between linear hypotheses $h(x) = w \cdot x$ we found and target function $f(x) = e^x$. However, we don't know which w can make the area be minimum, we need to find the size of area function $h(x; w)$ first.
   - $$\begin{aligned} h(x;w) &= \int_0^2 (wx - e^x)^2 dx \\ &= \int_0^2 (w^2x^2 - 2wxe^x + e^{2x}) dx \\ &= \frac{1}{3}w^2x^3 - 2w(xe^x - e^x) + \frac{1}{2}e^{2x}\Big|_{x=0}^{x=2} \\ &= \left(\frac{3}{8}w^2 - 2we^2 + \frac{1}{2}e^4\right) - \left(2w + \frac{1}{2}\right) \\ &= \frac{3}{8}w^2 - 2we^2 - 2w + \frac{1}{2}e^4 - \frac{1}{2} \end{aligned}.$$
   - Then try to find w which can make the area smallest.
   - $\frac{\partial}{\partial w}\left(\frac{3}{8}w^2 - 2we^2 - 2w + \frac{1}{2}e^4 - \frac{1}{2}\right) = \frac{16}{3}w - 2e^2 - 2 = 0,\ w = \frac{3}{8}\left(1 + e^2\right)$.
   - Then minus e^x (target function) by the result above, the magnitude of deterministic is $\left|e^x - \left(\frac{3+3e^2}{8}\right)x\right|$.

2. [**b**] 1

   - By $\mathcal{A}(\mathcal{D})$ will return a best $h \in \mathcal{H}$ which can make $E_{in}(\cdot)$ smallest. However, there exist anther $h^* \in \mathcal{H}$ can make $E_{out}(\cdot)$ smallest (noticed that we can't get h^* unless "cheating") , both hypothesis h and h^* might not be equal.
   - If $h = h^*$ (few condition), both $E_{in}(\cdot)$ and $E_{out}(\cdot)$ will be the smallest, making $\mathbb{E}[E_{in}(\mathcal{A}(\mathcal{D})] = \mathbb{E}[E_{out}(\mathcal{A}(\mathcal{D})]$.
   - If $h \neq h^*$ (frequently condition), $E_{out}(h)$ will not be minimized (because only h^* can return a minimum result in $E_{out}(\cdot)$), making $\mathbb{E}[E_{in}(\mathcal{A}(\mathcal{D})] < \mathbb{E}[E_{in}(\mathcal{A}(\mathcal{D})]$.
   - As above, the third statement, $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathcal{A}(\mathcal{D}))] > \mathbb{E}_{\mathcal{D}}[E_{\text{out}}(\mathcal{A}(\mathcal{D}))]$, is always false. The condition will never happen because $\mathcal{A}(\mathcal{D})$ will only return h rather than h^*.

3. [**d**] $2\mathrm{X}^T\mathrm{X} + N\sigma^2\mathrm{I}_{d+1}$

   - $$\begin{aligned} \mathbb{E}(\mathrm{X}_h^T\mathrm{X}_h) &= \mathbb{E}\left(\begin{bmatrix} | & & | & | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N & \tilde{\mathbf{x}}_1 & \cdots & \tilde{\mathbf{x}}_N \\ | & & | & | & & | \end{bmatrix} \cdot \begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_N & - \\ - & \tilde{\mathbf{x}}_1 & - \\ & \vdots & \\ - & \tilde{\mathbf{x}}_N & - \end{bmatrix}\right) \\ &= \mathbb{E}\left(\begin{bmatrix} | & & | & | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N & \mathbf{x}_1 + \epsilon & \cdots & \mathbf{x}_N + \epsilon \\ | & & | & | & & | \end{bmatrix} \cdot \begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_N & - \\ - & \mathbf{x}_1 + \epsilon & - \\ & \vdots & \\ - & \mathbf{x}_N + \epsilon & - \end{bmatrix}\right) \\ &= \mathbb{E}\left(\begin{bmatrix} | & & | & | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N & \mathbf{x}_1 + \epsilon & \cdots & \mathbf{x}_N + \epsilon \\ | & & | & | & & | \end{bmatrix} \cdot \begin{bmatrix} x_{11} \\ \vdots \\ x_{N1} \\ x_{11} + \epsilon \\ \vdots \\ x_{N1} + \epsilon \end{bmatrix} + \cdots + \begin{bmatrix} | & & | & | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N & \mathbf{x}_1 + \epsilon & \cdots & \mathbf{x}_N + \epsilon \\ | & & | & | & & | \end{bmatrix} \cdot \begin{bmatrix} x_{1d} \\ \vdots \\ x_{Nd} \\ x_{1d} + \epsilon \\ \vdots \\ x_{Nd} + \epsilon \end{bmatrix}\right) \\ &= \mathbb{E}\left(\mathrm{X}^T\mathrm{X} + \mathrm{X}^T\mathrm{X}\right) + N \cdot \mathrm{Var}\left(\begin{bmatrix} \epsilon & \cdots & \epsilon \\ \vdots & \ddots & \vdots \\ \epsilon & \cdots & \epsilon \end{bmatrix}\right) \\ &= \mathrm{X}^T\mathrm{X} + \mathrm{X}^T\mathrm{X} + N\sigma^2\mathrm{I}_{d+1} = 2\mathrm{X}^T\mathrm{X} + N\sigma^2\mathrm{I}_{d+1} \end{aligned}.$$

4. [**e**] $2\mathrm{X}^T\mathbf{y}$

- $$\begin{aligned}
\mathbb{E}\left(\mathrm{X}_h^T\mathbf{y}_h\right) &= \mathbb{E}\left(\begin{bmatrix} x_{11} \\ \vdots \\ x_{1d}\end{bmatrix} y_1 + \cdots + \begin{bmatrix} x_{N1} \\ \vdots \\ x_{Nd}\end{bmatrix} y_N + \begin{bmatrix} \tilde{x_{11}} \\ \vdots \\ \tilde{x_{1d}}\end{bmatrix} y_1 + \cdots + \begin{bmatrix} \tilde{x_{N1}} \\ \vdots \\ \tilde{x_{Nd}}\end{bmatrix} y_1\right) \\
&= \mathbb{E}\left(\begin{bmatrix} x_{11} \\ \vdots \\ x_{1d}\end{bmatrix} y_1 + \cdots + \begin{bmatrix} x_{N1} \\ \vdots \\ x_{Nd}\end{bmatrix} y_N + \begin{bmatrix} x_{11}+\epsilon \\ \vdots \\ x_{1d}+\epsilon\end{bmatrix} y_1 + \cdots + \begin{bmatrix} x_{N1}+\epsilon \\ \vdots \\ x_{Nd}+\epsilon\end{bmatrix} y_N\right) \\
&= \mathbb{E}\left(\begin{bmatrix} x_{11} \\ \vdots \\ x_{1d}\end{bmatrix} y_1 + \cdots + \begin{bmatrix} x_{N1} \\ \vdots \\ x_{Nd}\end{bmatrix} y_N + \begin{bmatrix} x_{11} \\ \vdots \\ x_{1d}\end{bmatrix} y_1 + \cdots + \begin{bmatrix} x_{N1} \\ \vdots \\ x_{Nd}\end{bmatrix} y_N + \begin{bmatrix} \epsilon \\ \vdots \\ \epsilon\end{bmatrix} y_1 + \cdots + \begin{bmatrix} \epsilon \\ \vdots \\ \epsilon\end{bmatrix} y_N\right) \\
&= \mathbb{E}\left(\begin{bmatrix} -\mathbf{x}_1- \\ \vdots \\ -\mathbf{x}_N-\end{bmatrix}\mathbf{y} + \begin{bmatrix} -\mathbf{x}_1- \\ \vdots \\ -\mathbf{x}_N-\end{bmatrix}\mathbf{y} + \begin{bmatrix} \epsilon & \cdots & \epsilon \\ \vdots & \ddots & \vdots \\ \epsilon & \cdots & \epsilon\end{bmatrix}\mathbf{y}\right) \\
&= \mathrm{X}^T\mathbf{y} + \mathrm{X}^T\mathbf{y} + \mathbf{0} \\
&= 2\mathrm{X}^T\mathbf{y}
\end{aligned}.$$
- Noticed that $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ because they are generated i.i.d. from a multivariate normal distribution.

5. [**d**] $\frac{\gamma_i}{\gamma_i+\lambda}$
   - Find the minimum $\mathbf{w} \in \mathbb{R}^{d+1}$.
   - Refer to the course slide, the optimal solution of regularized linear regression $\mathbf{w} = \left(\mathrm{Z}^T\mathrm{Z} + \lambda\mathrm{I}\right)^{-1}\mathrm{Z}^T\mathbf{y} = (\Gamma + \lambda\mathrm{I})^{-1}\mathrm{Z}^T\mathbf{y}$.
   - Noticed that $\mathrm{Z}^T\mathrm{Z} = (\mathrm{XQ})^T\mathrm{XQ} = \mathrm{Q}^T\mathrm{X}^T\mathrm{XQ} = \mathrm{Q}^T(\mathrm{Q}\Gamma\mathrm{Q}^T)\mathrm{Q} = \Gamma$.
   - When $\lambda > 0$, the solution $\mathbf{u}$ is $(\Gamma + \lambda\mathrm{I})^{-1}\mathrm{Z}^T\mathbf{y}$. And when $\lambda = 0$, the solution $\mathbf{v}$ is $(\Gamma)^{-1}\mathrm{Z}^T\mathbf{y}$. The difference is $(\lambda\mathrm{I})^{-1} = \mathrm{I}\frac{1}{\lambda} = \begin{bmatrix} 1/\gamma_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\gamma_d\end{bmatrix}$. So $u_i/v_i = \frac{1/(\gamma_i+\lambda)}{1/\gamma_i} = \frac{\gamma_i}{\gamma_i+\lambda}$.

6. [**a**] $C = \left(\frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2 + \lambda}\right)^2$
   - Find the minimum $w \in \mathbb{R}$ by $\frac{\partial}{\partial w}\left(\frac{1}{N}\sum_{n=1}^N (wx_n - y_n)^2 + \frac{\lambda}{N}w^2\right) = 0$.
   - $$\begin{aligned}
&\frac{\partial}{\partial w}\left(\frac{1}{N}\sum_{n=1}^N (wx_n - y_n)^2 + \frac{\lambda}{N}w^2\right) \\
=&\frac{2}{N}\left(\sum_{n=1}^N (wx_n^2 - x_n y_n)\right) + \frac{2w\lambda}{N} \\
=&\frac{2}{N}w\sum_{n=1}^N x_n^2 - \frac{2}{N}\sum_{n=1}^N x_n y_n + \frac{2w\lambda}{N} \\
=&w\left(\frac{2}{N}\sum_{n=1}^N x_n^2 + \frac{2\lambda}{N}\right) - \frac{2}{N}\sum_{n=1}^N x_n y_n \\
=&w\left(\sum_{n=1}^N x_n^2 + \lambda\right) - \sum_{n=1}^N x_n y_n \\
=&0
\end{aligned}.$$
   - The optimal solution $w* = \dfrac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2 + \lambda}$, thus $C = (w^*)^2 = \left(\dfrac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2 + \lambda}\right)^2$.

7. [**d**] $(y - 0.5)^2$
   - Fine the minimum $y \in \mathbb{R}$ by $\frac{\partial}{\partial y}\left(\frac{1}{N}\sum_{n=1}^N (y - y_n)^2 + \frac{2K}{N}\Omega(y)\right) = 0$.
   - $$\begin{aligned}
&\frac{\partial}{\partial y}\left(\frac{1}{N}\sum_{n=1}^N (y - y_n)^2 + \frac{2K}{N}\Omega(y)\right) \\
=&\frac{1}{N}\sum_{n=1}^N 2(y - y_n) + \frac{2K}{n}\Omega'(y) \\
=&\frac{2}{N}\left(\sum_{n=1}^N y - \sum_{n=1}^N y_n\right) + \frac{2K}{N}\Omega'(y) \\
=&\frac{2}{N}\left(Ny - \sum_{n=1}^N y_n\right) + \frac{2K}{N}\Omega'(y) \\
=&2y - \frac{2}{N}\sum_{n=1}^N y_n + \frac{2K}{N}\Omega'(y) \\
=&0
\end{aligned}.$$
   - [**a**] Assume $\Omega(y) = f_a = (y+1)^2$, $f_a' = 2(y+1)$, the result doesn't equal to the first function.
   - [**b**] Assume $\Omega(y) = f_b = (y+0.5)^2$, $f_b' = 2(y+0.5)$, the result doesn't equal to the first function.
   - [**c**] Assume $\Omega(y) = f_c = y^2$, $f_c' = 2y$, the result doesn't equal to the first function.
   - [**d**] Assume $\Omega(y) = f_d = (y-0.5)^2$, $f_d' = 2(y-0.5)$, the result equals to the first function.
   - [**e**] Assume $\Omega(y) = f_e = (y-1)^2$, $f_e' = 2(y-1)$, the result doesn't equal to the first function.
   - In fact, $\Omega(y)$ can have many possible candidates, but in here, we just fit five choices into $\Omega(y)$ then try.

8. [**b**] $\mathbf{w}^T\Gamma^2\mathbf{w}$
   - Both function should be equivalent, so:
     - $\tilde{\mathbf{w}}^T\Phi(\mathbf{x}_n) = \tilde{\mathbf{w}}^T\Gamma^{-1}\mathbf{x}$ must equal to $\mathbf{w}^T\mathbf{x}_n$.
     - $(\tilde{\mathbf{w}}^T\tilde{\mathbf{w}})$ must equal to $\Omega(\mathbf{w})$.

- Thus, $\mathbf{w}^T = \tilde{\mathbf{w}}\Gamma^{-1}$, $\tilde{\mathbf{w}}^T = \mathbf{w}^T\Gamma$, and $\tilde{\mathbf{w}} = (\mathbf{w}^T\Gamma)^T = \Gamma^T\mathbf{w}$. Noticed that the inverse of diagonal matrix equals to itself, $\Gamma^T = \Gamma$. Therefore, $(\tilde{\mathbf{w}}^T\tilde{\mathbf{w}}) = \mathbf{w}^T\Gamma\Gamma\mathbf{w} = \mathbf{w}^T\Gamma^2\mathbf{w}$.

9. [**b**] $\tilde{\mathrm{X}} = \sqrt{\lambda}\cdot\sqrt{\mathrm{B}}$, $\tilde{\mathbf{y}} = \mathbf{0}$
   - The term $\sum_{i=0}^{d}\beta_i w_i^2$ can be tensor into the matrix $\mathbf{w}^2\mathrm{B}$.
   - Find the minimum $\mathbf{w}_1 \in \mathbb{R}^{d+1}$ in the first function by $\frac{\partial}{\partial \mathbf{w}_1}\left(\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}_1^T\mathbf{x}_n - y_n\right)^2 + \frac{\lambda}{N}\sum_{i=0}^{d}\beta_i w_i^2\right) = 0$.
   - $$\begin{aligned}&\frac{\partial}{\partial \mathbf{w}_1}\left(\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}_1^T\mathbf{x}_n - y_n\right)^2 + \frac{\lambda}{N}\sum_{i=0}^{d}\beta_i w_i^2\right)\\ =&\frac{\partial}{\partial \mathbf{w}_1}\left(\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}_1^T\mathbf{x}_n - y_n\right)^2 + \frac{\lambda}{N}\mathbf{w}_1^2\mathrm{B}\right)\\ =&\frac{2}{N}\sum_{n=1}^{N}\left(\mathbf{w}_1^T\mathbf{x}_n^2 - \mathbf{x}_n y_n\right) + \frac{2}{N}\mathbf{w}_1\mathrm{B}\\ =&\mathbf{w}_1\left(\sum_{n=1}^{N}\mathbf{x}_n + \lambda\mathrm{B}\right) - \sum_{n=1}^{N}\mathbf{x}_n y_n\\ =&0\end{aligned}.$$
   - Next, find the minimum $\mathbf{w}_2 \in \mathbb{R}^{d+1}$ in the second function by $\frac{\partial}{\partial \mathbf{w}_2}\left(\frac{1}{N+K}\left(\sum_{n=1}^{N}(\mathbf{w}_2^T\mathbf{x}_n - y_n)^2 + \sum_{k=1}^{K}(\mathbf{w}_2^T\tilde{\mathbf{x}}_n - \tilde{y}_n)^2\right)\right) = 0$.
   - $$\begin{aligned}&\frac{\partial}{\partial \mathbf{w}_2}\left(\frac{1}{N+K}\left(\sum_{n=1}^{N}(\mathbf{w}_2^T\mathbf{x}_n - y_n)^2 + \sum_{k=1}^{K}(\mathbf{w}_2^T\tilde{\mathbf{x}}_n - \tilde{y}_n)^2\right)\right)\\ =&\frac{1}{N+K}\left(\sum_{n=1}^{N}2\mathbf{x}_n(\mathbf{w}_2^T\mathbf{x}_n - y_n) + \sum_{k=1}^{K}2\tilde{\mathbf{x}}_n(\mathbf{w}_2^T\tilde{\mathbf{x}}_n - \tilde{y}_n)\right)\\ =&\frac{2}{N+K}\left(\sum_{n=1}^{N}\mathbf{w}_2^T\mathbf{x}_n^2 - \sum_{n=1}^{N}\mathbf{x}_n y_n + \sum_{k=1}^{K}\mathbf{w}_2^T\tilde{\mathbf{x}}_n^2 - \sum_{k=1}^{K}\tilde{\mathbf{x}}_n\tilde{y}_n\right)\\ =&0\end{aligned}.$$
   - The optimal solution is $\mathbf{w}_1 = \dfrac{\sum_{n=1}^{N}\mathbf{x}_n y_n}{\sum_{n=1}^{N}\mathbf{x}_n^2 + \lambda\mathrm{B}}$ in first function, and $\mathbf{w}_2 = \dfrac{\sum_{n=1}^{N}\mathbf{x}_n y_n - \sum_{n=1}^{N}\tilde{\mathbf{x}_n}\tilde{y_n}}{\sum_{n=1}^{N}\mathbf{x}_n^2 + \sum_{n=1}^{N}\tilde{\mathbf{x}}^2}$ in second function.
   - Compare with both coefficient, $\mathbf{w}_1$ should equal to $\mathbf{w}_2$. We can get $\tilde{\mathrm{X}} = \sqrt{\lambda}\cdot\sqrt{\mathrm{B}}$ and $\tilde{\mathbf{y}} = \mathbf{0}$.
10. [**e**] 1
    - There have N positive samples and N negative samples. Assume we **leave one positive sample** into validation set. The classification algorithm $\mathcal{A}_{\text{majority}}$ will return **negative** in training dataset, because the number of positive samples is $N-1$ and the number of negative samples is N. However, the validation set used by $\mathcal{A}_{\text{majority}}$ will return **positive** due to there has one positive example.
    - On the other hand, if we leave **one negative sample** into validation set, the training set will return **positive** and the validation set will return **negative**.
    - Generally, $\mathcal{A}_{\text{majority}}$ will always return different classification between training set and validation set. So $E_{\text{loocv}}(\mathcal{A}_{\text{majority}}) = 1$.
11. [**c**] $2/N$
    - Refer to Problem 16 of Homework 2, the decision stump model's $\theta \in \{-1\} \cap \{\frac{x_i' + x_{i+1}'}{2}\}$. Means the model will try to train θ being the middle point ($\frac{1}{2}$) of **the most largest negative sample** and **the most smallest positive sample**.
    - If we leave **the most largest negative sample** out, the model will use **the second largest negative sample** and **the most smallest positive sample** to train θ. Sometimes, both points' middle point will not equal to **the most largest negative sample**, if $\theta <$(the most largest negative sample), it will let the classifier classify the most largest negative sample as positive, making an error in validation set.
    - On the other hand, If we leave **the most smallest positive sample** out, the model will use **the most largest negative sample** and **the second smallest positive sample** to train θ. Sometimes, both points' middle point will not equal to **the most smallest positive sample**, if $\theta <$(the most smallest positive sample), it will let the classifier classify the most smallest positive sample as negative, making an error in validation set.
    - There will be no error when we leave out others sample. Only when we leave out the these 2 samples: **the most largest negative sample** and **the most smallest positive sample**, might cause θ an error in validation set. So the tightest upper bound on the leave-one-out error on the decision stump model is $2/N$.
12. [**e**] $\sqrt{81 + 36\sqrt{6}}$
    - Denote point $\alpha = (x_1, y_1) = (3, 0)$, point $\beta = (x_2, y_2) = (\rho, 2)$, and point $\gamma = (x_3, y_3) = (-3, 0)$
    - In constant model $h(x) = w_0$:
      - Leave α out, use β and γ for model, $h(x) = 1$, the squared error is $(0-1)^2$.
      - Leave β out, use α and γ for model, $h(x) = 0$, the squared error is $(2-0)^2$.
      - Leave γ out, use α and β for model, $h(x) = 1$, the squared error is $(0-1)^2$.
      - $\text{err}(h(x), y) = \frac{1}{3}\left((0-1)^2 + (2-0)^2 + (0-1)^2\right) = 2$.
    - In linear model $h(x) = w_0 + w_1 x$:
      - Leave α out, use β and γ for model, $h(x) = \frac{2}{\rho-3}x - \frac{6}{\rho-3}$, the squared error is $(-\frac{6}{\rho-3} - \frac{6}{\rho-3})^2$.
      - Leave β out, use α and γ for model, $h(x) = 0$, the squared error is $(2-0)^2$.
      - Leave γ out, use α and β for model, $h(x) = \frac{2}{\rho+3}x + \frac{6}{\rho+3}$, the squared error is $(\frac{6}{\rho+3} + \frac{6}{\rho+3})^2$.
      - $\text{err}(h(x), y) = \frac{1}{3}\left((-\frac{6}{\rho-3} - \frac{6}{\rho-3})^2 + (2-0)^2 + (\frac{6}{\rho+3} + \frac{6}{\rho+3})^2\right) = 4 + 144\left(\frac{1}{(\rho-3)^2} + \frac{1}{(\rho+3)^2}\right)$.

- Both err should be equal (be "tied"). Solve $2 = 4 + 144\left(\frac{1}{(\rho-3)^2} + \frac{1}{(\rho+3)^2}\right)$, $\rho = \sqrt{81 + 36\sqrt{6}}$.

13. [**d**] $\frac{1}{K}$
    - By definition, $$\begin{aligned}&\mathrm{Var}_{\mathcal{D}_{\mathrm{val}}\sim\mathcal{P}^K}[E_{\mathrm{val}}(h)]\\ =&\mathrm{Var}_{(\mathbf{x},y)\sim\mathcal{P}}[\frac{1}{K}\sum_{n=1}^{K}\mathrm{err}(h(\mathbf{x}_n), y_n)]\\ =&\frac{1}{K^2}\mathrm{Var}_{(\mathbf{x},y)\sim\mathcal{P}}[\sum_{n=1}^{K}\mathrm{err}(h(\mathbf{x}_n), y_n)]\\ =&\frac{1}{K^2}\cdot\left(\mathrm{Var}_{(\mathbf{x},y)\sim\mathcal{P}}[\mathrm{err}(h(\mathbf{x}_1), y_1) + \cdots + \mathrm{err}(h(\mathbf{x}_K), y_K)]\right)\\ =&\frac{1}{K^2}\cdot\left(K\cdot\mathrm{Var}_{(\mathbf{x},y)\sim\mathcal{P}}[\mathrm{err}(h(\mathbf{x}), y)]\right)\\ =&\frac{1}{K}\cdot\left(\mathrm{Var}_{(\mathbf{x},y)\sim\mathcal{P}}[\mathrm{err}(h(\mathbf{x}), y)]\right)\end{aligned}.$$
    - Noticed that Var(A+B) = Var(A) + Var(B) + 2Cov(A,B), however the examples generate from a i.i.d. distribution, so the covariance term $\mathrm{Cov}[\mathrm{err}(h(\mathbf{x}_i), y_i), \mathrm{err}(h(\mathbf{x}_j), y_j)] = 0$, $\forall i, j$.
14. [**c**] $2/64$
    - There has $2^4 = 16$ binary combination for 4 points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$.
    - $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ form a rectangle. Under most of condition, the 2D perceptron can be shattered, $E_{in}(\mathbf{w}) = 0$ Only in two condition: if the set of diagonal points has different classification to another set of diagonal points, such as $\begin{bmatrix}\bigcirc & \times\\ \times & \bigcirc\end{bmatrix}$ and $\begin{bmatrix}\times & \bigcirc\\ \bigcirc & \times\end{bmatrix}$, the 2D perceptron cannot be shattered.
    - The probability is $\frac{1}{4} + \frac{1}{4}$(Symmetry for $\bigcirc$ and $\times$) $= \frac{2}{4}$. And the expectation $\mathbb{E}_{y_1,y_2,y_3,y_4} = \frac{1}{16}(E_{in}(\mathbf{w})) = \frac{1}{16}(\frac{2}{4}) = \frac{2}{64}$.
15. [**a**] $p = \frac{1-\epsilon_-}{\epsilon_+ - \epsilon_- + 1}$
    - By definition, $$\begin{aligned}E_{out} &= \frac{1}{N}\sum[g(x) \neq y] = \frac{1}{N}\sum[g(x) = -1, y = +1] + \frac{1}{N}\sum[g(x) = +1, y + -1]\\ &= \frac{1}{N}\sum\mathbb{P}(g(x) = -1|y = +1)\mathbb{P}(y = +1) + \frac{1}{N}\sum\mathbb{P}(g(x) = +1|y = -1)\mathbb{P}(y = -1)\\ &= \frac{1}{N}N\cdot(\frac{1}{2}\epsilon_+) + \frac{1}{N}\cdot(\frac{1}{2}\epsilon_-)\\ &= \frac{1}{2}\epsilon_+ + \frac{1}{2}\epsilon_-\end{aligned}.$$
    - Noticed that by the rule of conditional probability, from the first line to the second line due to $\mathbb{P}(A\cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
    - In another distribution, $\mathbb{P}(y = +1) = p$, $\mathbb{P}(y = -1) = 1 - p$.
    - By the same definition, $$\begin{aligned}E_{out}(g_c) &= \mathbb{P}(g(x) = -1|y = +1)\cdot\mathbb{P}(g(x) = +1|y = -1)\\ &= \epsilon_+ p + \epsilon_-(1-p)\\ &= 1 - p\end{aligned}.$$
    - Solve p, $p = \frac{1-\epsilon_-}{\epsilon_+ - \epsilon_- + 1}$.
16. [**b**] -2
17. [**a**] -4
18. [**e**] 0.14
19. [**d**] 0.13
20. [**c**] 0.12