ML HW4

B06303126 Lo Yun Chien

December 2020

Question 1

Correct answer:(c)
Since deterministic error define as the difference between target function and hypothesis, by using the distribution is uniform, the optimal w is to minimize the below value

$$\int_0^2 (e^x - wx)^2 dx$$

Then using differentiation

$$\begin{aligned}
&\int_0^2 2(e^x - wx)(-x)dx = 0 \\
&\Longrightarrow w\int_0^2 x^2 dx = \int_0^2 e^x x dx \\
&\Longrightarrow w\frac{8}{3} = 2e^2 - e^2 + 1 \\
&\Longrightarrow w = \frac{3(e^2+1)}{8}
\end{aligned}$$

So deterministic error is

$$|e^x - \frac{3(e^2+1)}{8}x|$$

Question 2

Correct answer:(b)
Consider for any D, define

$$h^* = \arg\min_{h\in H} E_{out}(h)$$

Since $E_{out}(h^*) \leq E_{out}(A(D))$ for all D, one has $E[E_{out}(h^*)] \leq E[E_{out}(A(D))]$.
And $E_{in}(A(D)) \leq E_{in}(h^*)$ for all D, one has $E[E_{in}(A(D))] \leq E[E_{in}(h^*)]$.
Recall that E_{out} is defined on the distribution, take expectation average in-sample D, which means $E[E_{out}(h^*)] = E[E_{in}(h^*)]$, in the end we have

$$E[E_{in}(A(D))] \leq E[E_{out}(A(D))]$$

Question 3

Correct answer:(d)
Define $\tilde{X}$ is a matrix which column vectors are $\tilde{x}_n$'s. A is matrix which column vectors are ϵ_n generating for x_n.

$$\begin{aligned} E[X_h^T X] &= E[X^T X + \tilde{X}^T \tilde{X}] \\ &= X^T X + E[X^T X + X^T A + A^T X + A^T A] \\ &= 2X^T X + N\sigma^2 I \end{aligned}$$

Question 4

Correct answer: (e)
Using the same notation as Q3, we have

$$\begin{aligned} E[X_h^T y] &= E[X^T y + \tilde{X}^T y] \\ &= X^T y + E[X^T y + A^T y] \\ &= 2X^T y \end{aligned}$$

Question 5

Correct answer:(c)
From class lecture, we derive the optimal v and u are

$$v = (Z^T Z)^{-1} Z^T y$$

$$u = (Z^T Z + \lambda I)^{-1} Z^T y$$

Since $Z^T Z = Q^T X^T X Q = Q^T (Q\Gamma Q^T)^T (Q\Gamma Q^T) Q = \Gamma^2$, to compare$u_i, v_i$, neglect $Z^T y$, we'll get $\frac{v_i}{u_i} = \frac{\gamma_i^2}{\gamma_i^2 + \lambda}$

Question 6

Correct answer:(a)
Since $w = (X^T X + \lambda I)^{-1} X^T y$, in one dimensional condition, we have

$$w = \frac{\sum_{i=1}^{n} x_n y_n}{\sum_{i=1}^{n} x^2 + \lambda}$$

Then C is the square of it.

Question 7

Correct answer:(d)
Consider adding 2K example forms $y_{n+1}, y_{n+2}..y_{n+2K}$, using $N + 2K$ examples to minimize square

error, one has best y is $\frac{\sum_{i=1}^{N} y_n + K}{N+2k}$. Then we can rewrite the form of E_{in}

$$\frac{1}{N+2K}\sum_{i=1}^{N+2K}(y-y_n)^2 = \frac{1}{N+2K}(\sum_{i=1}^{N}(y-y_n)^2 + \sum_{i=N+1}^{N+2K}(y-y_n)^2)$$

Observe the last term

$$\sum_{i=N+1}^{N+2K}(y-y_n)^2 = K(y^2+(y-1)^2) = 2K(y-0.5)^2 - \frac{1}{2}K$$

Without concern of the constant, $\Omega(y)$ would be $(y-0.5)^2$

Question 8

Correct answer: (b)
Since they are equivalent, one has $\tilde{w}^T\Gamma^{-1}x = w^Tx$, which is $\tilde{w}^T = w^T\Gamma$, Then for the regularizer, we have $\tilde{w}^T\tilde{w} = w^T\Gamma^2 w$.

Question 9

Correct answer:(b)
In order to let two problem yield the same w, we want

$$\lambda\sum_{i=0}^{d}\beta_i w_i^2 = \sum_{i=1}^{K}(w^T\tilde{x}_k - \tilde{y}_k)^2$$

Where $K = d+1$, expand the RHS

$$\sum_{i=1}^{K}(w^T\tilde{x}_k - \tilde{y}_k)^2 = \sum_{i=0}^{d}(\tilde{x}_k^T ww^T\tilde{x}_k - 2w^T\tilde{x}_k\tilde{y}_k + \tilde{y}_k^2)$$

Since LHS only left w_i^2, then the last two term should be 0, then $\tilde{y} = 0$, the coefficient term is $\tilde{x}_k^2$, so $\tilde{X} = \sqrt{\lambda}\sqrt{B}$.

Question 10

Correct answer:(e)
If exclude one positive example, the left majority will predict negative for all and get $E_{out} = 1$, on the other hand, it's the same. So the error is always 1.

Question 11

Correct answer:(c)
While doing leave out out, if there at least two positive and negative examples, there are only two instance can violate the rule which are the biggest negative one and the smallest positive one.

Question 12

Correct answer: (e)
While predict with a constant, take (x_1, y_1) out, we predict $y = 1$, error is 1, take (x_2, y_2) out, we predict $y = 0$, error is 4. take (x_3, y_3) out, we predict $y = 1$, error is 1. So $E_{loocv} = 2$.
While using linear hypothesis, take (x_1, y_1) out, we have $h_1 = \frac{2}{\rho+3}x + \frac{6}{\rho+3}$, error is $(\frac{12}{\rho+3})^2$. take (x_2, y_2) out, predict constant, error is 4, take (x_3, y_3) out, we have $h_3 = \frac{2}{\rho-3}x - \frac{6}{\rho-3}$, error is $(\frac{-12}{\rho-3})^2$ so we get

$$(\frac{12}{\rho+3})^2 + (\frac{-12}{\rho-3})^2 = 2$$

Using wolfram, we get the answer is (e)

Question 13

Correct answer:(d)

$$\begin{aligned} Var[E_{val}(h)] &= Var[\frac{\sum_{i=1}^{K} err(h(x_i), y_i)}{K}] \\ &= \frac{\sum_{i=1}^{K} Var[err(h(x), y)]}{K^2} \\ &= \frac{Var[err(h(x), y)]}{K} \end{aligned}$$

Since they are i.i.d, we don't need compute covariance term. Thus we yield such result.

Question 14

Correct answer: (c)
2D perceptron can only create 14 dichotomy, then at least 2 cases E_{in} wouldn't be 0. for these two case, we can only correctly specify 3 points, then $E_{in} = \frac{1}{4}$, since each dichotomy has the same probability appear. Then the expectation is $\frac{1}{4} * \frac{2}{16} = \frac{2}{64}$.

Question 15

Correct answer: (a)
Under more general case, $E_{out}(g)$ is

$$E_{out}(g) = p\epsilon_+ + (1-p)\epsilon_-$$

If we want to hypothesis equally good, which means

$$p\epsilon_+ + (1-p)\epsilon_- = (1-p)$$

Solving this equation, we get the answer is (a)