Machine Learning Techniques HW2
B05902109 資工二 柯上優

1.

$$\mathrm{F(A,B)} = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(Az_n+B)))$$

$$\frac{\partial \mathrm{F(A,B)}}{\partial \mathrm{A}} = \frac{1}{N}\sum_{n=1}^{N} \frac{-y_n \mathrm{z_n} \exp\left(-y_n(Az_n+B)\right)}{\left(1+\exp\left(-y_n(Az_n+B)\right)\right)} = \frac{1}{N}\sum_{n=1}^{N} -y_n \mathrm{z_n p_n}$$

$$\frac{\partial \mathrm{F(A,B)}}{\partial \mathrm{B}} = \frac{1}{N}\sum_{n=1}^{N} \frac{-y_n \exp\left(-y_n(Az_n+B)\right)}{\left(1+\exp\left(-y_n(Az_n+B)\right)\right)} = \frac{1}{N}\sum_{n=1}^{N} -y_n \mathrm{p_n}$$

$$\nabla \mathrm{F(A,B)} = \frac{1}{N}\sum_{n=1}^{N} [-y_n \mathrm{z_n p_n}, -y_n \mathrm{p_n}]^{\mathrm{T}}$$

2.

By MLF hw3, we know that

$$\mathrm{H(F)} = \nabla^2 \mathrm{F(A,B)}$$

So we can derive that

$$\mathrm{H(F)} = \begin{bmatrix} \frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}^2} & \frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}\,\partial \mathrm{B}} \\ \frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}\,\partial \mathrm{B}} & \frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{B}^2} \end{bmatrix}$$

$$\frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}^2} = \frac{1}{N}\sum_{n=1}^{N} -y_n \mathrm{z_n} \frac{-y_n \mathrm{z_n} \exp\left(-y_n(Az_n+B)\right)}{\left(1+\exp\left(-y_n(Az_n+B)\right)\right)^2} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 \mathrm{z_n^2 p_n}(1-\mathrm{p_n})$$

$$\frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}\,\partial \mathrm{B}} = \frac{1}{N}\sum_{n=1}^{N} -y_n \mathrm{z_n} \frac{-y_n \exp\left(-y_n(Az_n+B)\right)}{\left(1+\exp\left(-y_n(Az_n+B)\right)\right)^2} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n \mathrm{p_n}(1-\mathrm{p_n})$$

$$\frac{\partial^2 \mathrm{F(A,B)}}{\partial \mathrm{A}^2} = \frac{1}{N}\sum_{n=1}^{N} -y_n \frac{-y_n \exp\left(-y_n(Az_n+B)\right)}{\left(1+\exp\left(-y_n(Az_n+B)\right)\right)^2} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 \mathrm{p_n}(1-\mathrm{p_n})$$

$$\mathrm{H(F)} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 \begin{bmatrix} \mathrm{z_n^2 p_n(1-p_n)} & z_n \mathrm{p_n(1-p_n)} \\ z_n \mathrm{p_n(1-p_n)} & \mathrm{p_n(1-p_n)} \end{bmatrix}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \begin{bmatrix} \mathrm{z_n^2 p_n(1-p_n)} & z_n \mathrm{p_n(1-p_n)} \\ z_n \mathrm{p_n(1-p_n)} & \mathrm{p_n(1-p_n)} \end{bmatrix}$$

其中 y_n 平方必為 1。

3.

When gamma goes to infinite, the kernel matrix looks like an identity matrix, which for same x (in $\mathrm{K(i,i)}$, where $0 < \mathrm{i} \leq \mathrm{n}$) minus itself is zero and exponetail goes to 1, and different x (as the assignment promises) minus other will make exponential goes to 0.

By the slide,

want $\nabla E_{\text{aug}}(\beta) = \mathbf{0}$: one analytic solution

$$\beta = (\lambda \mathrm{I} + \mathrm{K})^{-1}\mathbf{y}$$

We can know that optimal beta is

$$\beta = (\lambda \mathrm{I} + \mathrm{I})^{-1} y = \frac{\mathrm{y}}{\lambda + 1}$$

4.

使用$\mathrm{g_0}(x)$計算$\mathrm{E_{test}}(g_0)$可以得到

$$\mathrm{E_{test}}(g_0) = \frac{1}{M}\sum_{m=1}^{M}(0 - \tilde{y}_m)^2 = \frac{1}{M}\sum_{m=1}^{M}\tilde{y}_m^2$$

所求為

$$\sum_{m=1}^{M} g_t(\tilde{x}_m)\tilde{y}_m = \frac{M}{2}(s_t - e_t + \mathrm{E_{test}}(g_0)) = \frac{M}{2}(s_t - e_t + \mathrm{e_0})$$

5.

In section 4 ‘Test-set Blending’ of the paper teacher Lin gave us, we know that if we set

$$\mathrm{r} = \mathrm{y} = [\mathrm{y_0}, \mathrm{y_2}, \dots\dots, \mathrm{y_T}]$$
$$\mathrm{Z} = \mathrm{g} = [\mathrm{g_0}, \mathrm{g_1}, \dots\dots, \mathrm{g_T}]^{\mathrm{T}}$$
$$\mathrm{w} = \alpha = [\alpha_0, \alpha_1, \dots\dots \alpha_{\mathrm{T}}]$$

by ridge regression, to calculate optimal weights, we will get

$$\mathrm{w} = (Z^T Z + \lambda I)^{-1} Z^T r = (Z^T Z + \lambda I)^{-1} Q$$

still, we don' t know r, but in the paper it use RMSE

$$\mathrm{e_m} = \sqrt{\frac{\|r - z_m\|^2}{N}}$$

$$\mathrm{z_m^T} r = \frac{\|r\|^2 + \|z_m\|^2 - N e_m^2}{2} = \mathrm{Q}$$

And by the paper and Problem 4, we can estimate that

$$\mathrm{e_m} \approx \tilde{e}_m \ \ and \ \ \|r\|^2 \approx N \tilde{e}_0^2$$

Where we can get $\tilde{e}_m$ by submit $\mathrm{g_m}$ to the judge system and check leaderboard' s RMSE, $\tilde{e}_0$ can be get by submit a all-answer-zero $\mathrm{g_0}$ and check the RMSE.

In the end, all we have to do is submit all the $\mathrm{g_t}$ , $0 \le \mathrm{t} \le \mathrm{T}$, to the system and record their RMSE, and calculate the optimal $\alpha$ (or call $\mathrm{w}$ )

6.

Set training set =$\{(x_1, 2x_1 - x_1^2), (x_2, 2x_2 - x_2^2)\}$,

With the form $\mathrm{h(x)} = \mathrm{w_1} x + w_0$, we ca get the mean square error

$$\mathrm{error} = (\mathrm{w_1} x_1 + w_0 - (2x_1 - x_1^2))^2 + (\mathrm{w_1} x_2 + w_0 - (2x_2 - x_2^2))^2$$

with partial differential,

$$\frac{\partial \mathrm{error}}{\partial \mathrm{w_1}} = 2\mathrm{x_1}\left(\mathrm{w_1} x_1 + w_0 - (2x_1 - x_1^2)\right) + 2x_2(\mathrm{w_1} x_2 + w_0 - (2x_2 - x_2^2)) = 0$$

$$\frac{\partial \mathrm{error}}{\partial \mathrm{w_0}} = 2\left(\mathrm{w_1} x_1 + w_0 - (2x_1 - x_1^2)\right) + 2(\mathrm{w_1} x_2 + w_0 - (2x_2 - x_2^2)) = 0$$

We can get the min w

$$\mathrm{w_1} = -x_1 - x_2 + 2, w_0 = x_1 x_2$$

And we know

$$\mathrm{h(x)} = (-x_1 - x_2 + 2)\mathrm{x} + (x_1 x_2)$$

for this question,

$$\bar{g}(x) = \lim_{T \to inf} \frac{1}{T} \sum_{t=1}^{T} g_t = E(-x_1 - x_2 + 2)x + E(x_1 x_2) = \mathrm{x} + \frac{1}{4}$$

7.

By the content of the slide, to make g_2 diverse to g_1, we product each with the other percentage, that is,

$$\frac{u_{+}^{(2)}}{u_{-}^{(2)}} = \frac{1-87\%}{87\%} = \frac{13\%}{87\%} = \frac{13}{87}$$

8.

By the g' s rule,

$$g_{s,i,\theta}(\mathbf{x}) = s \cdot \text{sign}\Big(x_i - \theta\Big),$$

where $i \in \{1, 2, \cdots, d\}, d$ is the finite dimensionality of the input space, $s \in \{-1, +1\}, \theta \in \mathbb{R}$, and $\text{sign}(0) = +1$

We know the formula

$$2\mathrm{dM} + 2$$

By M choises in [0, M], positive and negative have double choises, d dimentionality, and the definition

Two decision stumps $g$ and $\hat{g}$ are defined as the *same* if $g(\mathbf{x}) = \hat{g}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$.

make all positive and all negative be the same.

So the answer is 22.

9.

In question 8, we know that the number of classifier is $2\mathrm{dM} + 2$, but for K(x, x'), there is a range that will make some sets: $s * \text{sign}(x - \theta) * s * \text{sign}(x' - \theta) = -1$, which will decrease the number of K(x, x').

The range number is $\|x' - x\|$, and the possible classifier number is $2\|x' - x\|$ because of both positive and negative.

Finally, we can know that $\mathrm{K}(\mathrm{x}, \mathrm{x}') = 2\mathrm{dM} + 2 - 4\|x' - x\|$

10.

$$\mathrm{q_t} = \begin{cases} 1 \ , for \ \ \theta \leq x_i \ \ and \ \ s = 1 \\ \quad 0 \ \ , otherwise \end{cases}$$

$$, where \ \ \mathrm{q_t} \ \ match \ \ to \ \ the \ \ g_{s,i,\theta}(x) = s * sign(x_i - \theta)$$

Proof:

For one dimension, we know that $\varphi_{\mathrm{hi}}(\mathrm{x})$ has $\mathrm{x}$ terms with $\mathrm{q_t} = 1$, $\varphi_{\mathrm{hi}}(\mathrm{x}')$ has $\mathrm{x}'$ terms with $\mathrm{q_t} = 1$, and those terms also have $\mathrm{h_t}(x) = 1$. When we calculate $\varphi_{\mathrm{hi}}(\mathrm{x})\varphi_{\mathrm{hi}}(\mathrm{x}')$, only the terms with both $\mathrm{q_t} = 1$ and $\mathrm{h_t}(x) = 1$ will remain. So the one in $\max(\mathrm{x}, \mathrm{x}')$ will lose some terms, which is $\|x - x'\|$, and finally remaining terms are equal to $\min(\mathrm{x}, \mathrm{x}')$, leading to $\varphi_{\mathrm{hi}}(\mathrm{x})\varphi_{\mathrm{hi}}(\mathrm{x}') = \min(\mathrm{x}, \mathrm{x}')$.

For more than one dimension, it is exactly the same condition in the 'i' dimention, and the other dimensions are all $\mathrm{q_t} = 0$, that is , the result is actually same to the $\mathrm{K_{hi}}(x, x')$.

11.

gamma = 32 , lamda = 0.001 , E_in = 0
gamma = 32 , lamda = 1 , E_in = 0
gamma = 32 , lamda = 1000 , E_in = 0
gamma = 2 , lamda = 0.001 , E_in = 0
gamma = 2 , lamda = 1 , E_in = 0
gamma = 2 , lamda = 1000 , E_in = 0
gamma = 0.125 , lamda = 0.001 , E_in = 0
gamma = 0.125 , lamda = 1 , E_in = 0.03
gamma = 0.125 , lamda = 1000 , E_in = 0.2425

Minimum E_in = 0 happens in 7 combinations:
All combination with gamma = 32 and 2, and (gamma, lamda) = (0.125, 0.001)

12.

gamma = 32 , lamda = 0.001 , E_out = 0.45
gamma = 32 , lamda = 1 , E_out = 0.45
gamma = 32 , lamda = 1000 , E_out = 0.45
gamma = 2 , lamda = 0.001 , E_out = 0.44
gamma = 2 , lamda = 1 , E_out = 0.44
gamma = 2 , lamda = 1000 , E_out = 0.44
gamma = 0.125 , lamda = 0.001 , E_out = 0.46
gamma = 0.125 , lamda = 1 , E_out = 0.45
gamma = 0.125 , lamda = 1000 , E_out = 0.39

Minimum E_out = 0.39 happens in (gamma, lamda) = (0.125, 1000).

13.

lamda = 0.01 , E_in = 0.3175
lamda = 0.1 , E_in = 0.3175
lamda = 1 , E_in = 0.3175
lamda = 10 , E_in = 0.32
lamda = 100 , E_in = 0.3125

Minimum E_in = 0.3125 happens in lamda = 100.

14.

lamda = 0.01 , E_out = 0.36

lamda = 0.1 , E_out = 0.36

lamda = 1 , E_out = 0.36

lamda = 10 , E_out = 0.37

lamda = 100 , E_out = 0.39

Minimun E_out = 0.36 happens in lamda = 0.01, 0.1, and 1.

15.

lamda = 0.01 , E_in = 0.3225

lamda = 0.1 , E_in = 0.3225

lamda = 1 , E_in = 0.3225

lamda = 10 , E_in = 0.3225

lamda = 100 , E_in = 0.3175

After numbers of program running, I consider that 'lamda = 100' has a smaller E_in = 0.175, which is also the one that it is smaller in Question 13. I think it is for same data, despite we randomly pick some of them to generate g, they still can get close conclusion by voting. Those classified easily in Question 13 are also classified in this problem, and those ambiguous are misjudged here, too.

16.

lamda = 0.01 , E_out = 0.37

lamda = 0.1 , E_out = 0.37

lamda = 1 , E_out = 0.37

lamda = 10 , E_out = 0.38

lamda = 100 , E_out = 0.39

After numbers of program running, I consider that 'lamda = 0.01, 0.1, and 1' have a same smaller E_in = 0.37, which are also the one that it is smaller in Question 14. There is a interesting finding that in Question 15 and Question 16, their E_in and E_out is obviously a little higher than those in Question 13 and Question 14. It may due to voting system has some g come from bias data set, and their judgement can interfere the final result.

17.

Reference: B05902028 王元益

In the course slide 204 Page. 10, we have

$$\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m K_2(x,x') - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to}\ \ \sum_{n=1}^{N} y_n\alpha_n = 0;\ \ 0 \le \alpha_\mathrm{n} \le C, for\ \ n = 1,2,\dots,N$$

now we change it into

$$\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m (K_1(x,x') + \kappa) - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to}\ \ \sum_{n=1}^{N} y_n\alpha_n = 0;\ \ 0 \le \alpha_\mathrm{n} \le C, for\ \ n = 1,2,\dots,N$$

we can derive that

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \kappa = \frac{1}{2}\kappa\sum_{n=1}^{N}\alpha_n y_n\sum_{m=1}^{N}\alpha_m y_m = 0, \text{for}\ \ \sum_{n=1}^{N} y_n\alpha_n = 0$$

Now we know that $\kappa$ doesn't effect the optimal solution $\alpha$, and it is exactly the same $g_{svm}$ .

18.

Same as above. We change it into

$$\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m (K_1(x,x') + \gamma(x) + \gamma(x')) - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to}\ \ \sum_{n=1}^{N} y_n\alpha_n = 0;\ \ 0 \le \alpha_\mathrm{n} \le C, for\ \ n = 1,2,\dots,N$$

derive the $\gamma(x)$ and $\gamma(x')$

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \gamma(x_n) = \frac{1}{2}\sum_{n=1}^{N}\alpha_n y_n\gamma(x_n)\sum_{m=1}^{N}\alpha_m y_m = 0, \text{for}\ \ \sum_{m=1}^{N}\alpha_m y_m = 0$$

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \gamma(x'_m) = \frac{1}{2}\sum_{m=1}^{N}\alpha_m y_m\gamma(x'_m)\sum_{n=1}^{N}\alpha_n y_n = 0, \text{for}\ \ \sum_{n=1}^{N}\alpha_n y_n = 0$$

So we can know that despite the fact that K1 is not a valid kernel, $\gamma(x)$ won't change the optimal solution $\alpha$, and it is the same $g_{svm}$ .