Home Work 2

Machine Learning Techniques

R04323050

經濟碩三　陳伯駒

1.

Let $s = -y_n \left[A(\mathbf{w}_{svm}^T \cdot \phi(\mathbf{x}_n + b_{svm}) + B)\right]$, then

$$\frac{\partial F}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^s} \cdot e^s \cdot \frac{\partial s}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} \frac{e^s}{1+e^s}(-y_n)z_n = \frac{-1}{N}\sum_{n=1}^{N} y_n z_n p_n$$

$$\frac{\partial F}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^s} \cdot e^s \cdot \frac{\partial s}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} \frac{e^s}{1+e^s}(-y_n) = \frac{-1}{N}\sum_{n-=1}^{N} y_n p_n$$

$$\therefore\ \nabla F(A,B) = \begin{bmatrix} \dfrac{-1}{N}\displaystyle\sum_{n=1}^{N} y_n z_n p_n \\ \dfrac{-1}{N}\displaystyle\sum_{n-=1}^{N} y_n p_n \end{bmatrix}$$

2.

$$\frac{\partial^2 F}{\partial A^2} = \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot z_n \cdot \frac{\partial p_n}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} -y_n z_n p_n(1-p_n)(-y_n z_n) = \frac{1}{N}\sum_{n=1}^{N} z_n^2 \cdot p_n \cdot (1-p_n)$$

$$\frac{\partial^2 F}{\partial A \partial B} = \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot z_n \cdot \frac{\partial p_n}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} -y_n z_n p_n(1-p_n)(-y_n) = \frac{1}{N}\sum_{n=1}^{N} z_n \cdot p_n \cdot (1-p_n)$$

$$\frac{\partial^2 F}{\partial B \partial A} = \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot \frac{\partial p_n}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} -y_n p_n(1-p_n)(-y_n z_n) = \frac{1}{N}\sum_{n=1}^{N} z_n \cdot p_n \cdot (1-p_n)$$

$$\frac{\partial^2 F}{\partial B^2} = \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot \frac{\partial p_n}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} -y_n p_n(1-p_n)(-y_n) = \frac{1}{N}\sum_{n=1}^{N} p_n \cdot (1-p_n)$$

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

3.

Now $K(\mathbf{x},\mathbf{x}') = \exp(-\gamma \cdot \|\mathbf{x}-\mathbf{x}'\|^2)$. By the exercise in slides in class 3, if $\mathbf{x} \neq \mathbf{x}'$, $K(\mathbf{x},\mathbf{x}') = 0$ when $\gamma \longrightarrow \infty$.

Hence, $\boldsymbol{\beta} = (\lambda \mathrm{I} + K)^{-1}\mathbf{y} = \frac{1}{\lambda} \cdot \mathbf{y}$

4.

$$\begin{aligned} e_t = E_{test}(g_t) &= \frac{1}{M}\sum_{m=1}^{M}\left[(g_t(\tilde{x}_m))^2 + \tilde{y}_m^2 - 2 \cdot g_t(\tilde{x}_m) \cdot \tilde{y}_m\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}(g_t(\tilde{x}_m))^2 + \frac{1}{M}\sum_{m=1}^{M}(g_0 - \tilde{y}_m)^2 - \frac{2}{M}\sum_{m=1}^{M} g_t(\tilde{x}_m) \cdot \tilde{y}_m \\ &= s_t + e_0 - \frac{2}{M}\sum_{m=1}^{M} g_t(\tilde{x}_m) \cdot \tilde{y}_m \end{aligned}$$

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

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

5.

step 1. 從 $\mathcal{D}_{train}$ 得到 $g_1^-, g_2^-, ..., g_T^-$。

step 2. 從 $\mathcal{D}_{val}$ 的 $\{(\mathbf{x}_n, y_n)\}$，轉換至 $\mathbf{z}_n = \Phi^-(\mathbf{x}_n) = (g_1^-(\mathbf{x}), g_2^-(\mathbf{x}), ..., g_T^-(\mathbf{x}))$

step 3. Linear Blending：

i. 使用 $LinearModel(\mathbf{z}_n, y_n)$，得到 $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_T)$。

ii. 重新使用所有資料 $\mathcal{D}$，得到 $\Phi(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), ..., g_T(\mathbf{x}))$。

iii. $G_{LINB}(\mathbf{x}) = \underset{\mathbf{w}=\mathbf{x}}{LinearModel}(\Phi(\mathbf{x}))$

6.

假設兩組訓練樣本為：$(x_1, 2x_1 - x_1^2), (x_2, 2x_2 - x_2^2)$。

The loss function would be: $\mathcal{L} = [w_1x_1 + w_0 - (2x_1 - x_1^2)]^2 + [w_1x_2 + w_0 - (2x_2 - x_2^2)]^2$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_1} = 2(w_1x_1 + w_0 - 2x_1 + x_1^2) \cdot x_1 + 2(w_1x_2 + w_0 - 2x2 + x_2^2) \cdot x_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial w_2} = 2(w_1x_1 + w_0 - 2x_1 + x_1^2) + 2(w_1x_2 + w_0 - 2x2 + x_2^2) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} w_1^* = 2 - x_1 - x_2 \\ w_2^* = x_1x_2 \end{cases}$$

$\therefore\ \bar{g}(x) = E\left[(2 - x_1 - x_2) \cdot x + x_1x_2\right] = x + \frac{1}{4}$。[1]

7.

$$\epsilon_t = \frac{\sum_{n=1}^{N} \mu_n^{(t)} [\![y_n \neq g_t(x_n)]\!]}{\sum_{n=1}^{N} \mu_n^{(t)}} = \frac{\sum_{n=1}^{N} \frac{1}{N} [\![y_n \neq g_t(x_n)]\!]}{\sum_{n=1}^{N} \frac{1}{N}} = \frac{\sum_{n=1}^{N} [\![y_n \neq g_t(x_n)]\!]}{N} = 0.13$$

$$\therefore\ \frac{\mu_+^2}{\mu_-^2} = \frac{\blacklozenge_t^{-1}}{\blacklozenge_t} = \frac{\epsilon_t}{1 - \epsilon_t} = \frac{0.13}{0.87} = \frac{13}{87} \text{ (where } \blacklozenge_t = \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}}\text{)}$$

8.

通過不同維度、不同的 θ 取在兩點之間及不同的方向，有決策樹共 $2d \times (M - 0) = 2dM$ 種。但仍須考量有全正、全負兩種情形，$\therefore$ 總共有 $2dM + 2 = 2 \times 2 \times 5 + 2 = 22$ 種。

---

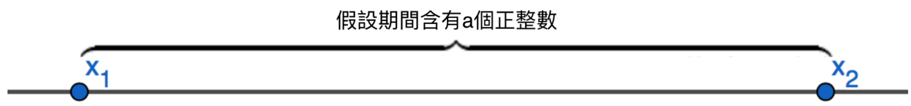
[1] $x_1, x_2 \overset{i.i.d.}{\sim} U(0, 1)$

$E(x_1) = \frac{1}{2}, E(x_2) = \frac{1}{2}, E(x_1x_2) = E(x_1) \cdot E(x_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

9.

$$
\begin{aligned}
K_{ds} &= \sum_{j=1}^{|\mathcal{G}|} sign(x_{ij} - \theta_j) \cdot sign(x'_{ij} - \theta_j) \\
&= sign(x_{i1} - \theta_1) \cdot sign(x'_{i1} - \theta_1) + sign(x_{i2} - \theta_2) \cdot sign(x'_{i2} - \theta_2) + \cdots \\
&+ sign(x_{i|\mathcal{G}|} - \theta_{|\mathcal{G}|}) \cdot sign(x'_{i|\mathcal{G}|} - \theta_{|\mathcal{G}|})
\end{aligned}
$$

上式的項數 $|\mathcal{G}| = 2dM + 2$，包含了所有可能的情況 (s, i, θ)，現在我們進一步來檢查有多少為 $+1$ 項、有多少為 -1 項：



如上圖，假設 x_1, x_2 之間存在 a 個整數，則總共有 $2 \times (a + 1)$ 種 hypothesis 使得分類結果乘積為 -1。對於題目給定的 $\mathbf{x}, \mathbf{x}'$ 為 integers，因此總共包含的分類結果乘積為 -1 的項數為：$2 \cdot \|\mathbf{x} - \mathbf{x}'\|$

$\therefore$ 分類結果乘積為 $+1$ 的項數為：$2dM + 2 - 2 \times \|\mathbf{x} - \mathbf{x}'\|$

$$
\begin{aligned}
K_{ds} &= (+1) \times [2dM + 2 - 2 \times \|\mathbf{x} - \mathbf{x}'\|] + (-1) \times [2 \times \|\mathbf{x} - \mathbf{x}'\|] \\
&= 2dM + 2 - 4 \times \|\mathbf{x} - \mathbf{x}'\|
\end{aligned}
$$

10.

Suppose N be the numbers of inputs, and we let $q_t = \begin{cases} 1 & ,\text{if} \quad t \le N \\ 0 & ,\text{if} \quad t > N \end{cases}$, then:

$$
\Phi_{hi}(\mathbf{x}) = (\underbrace{1, 1, \cdots, 1}_{x_1\text{個}}, \underbrace{0, 0, \cdots, 0}_{N-x_1\text{個}}, \underbrace{1, 1, \cdots, 1}_{x_2\text{個}}, \underbrace{0, 0, \cdots, 0}_{N-x_2\text{個}}, \cdots, \underbrace{1, 1, \cdots, 1}_{x_d\text{個}}, \underbrace{0, 0, \cdots, 0}_{N-x_d\text{個}})
$$

Each bin value x_i is encoded with N bits such that x_i first bits are equal to 1 and $N - x_i$ bits are equal to 0. If we consider the dot product on this binary representation between $\Phi_{hi}(\mathbf{x})$ and $\Phi_{hi}(\mathbf{x}')$, we obtain for each bin than $\min\{x_i, x'_i\}$ components products equal to 1 and $N - \min\{x_i, x'_i\}$ equal to 0.

$\therefore$ we cna write $K_{hi}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \min\{x_i, x'_i\} = \Phi_{hi}(\mathbf{x}) \cdot \Phi_{hi}(\mathbf{x}')$

11.

$\gamma = 32, \lambda = 0.001 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.45

$\gamma = 32, \lambda = 1 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.45

$\gamma = 32, \lambda = 1000 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.45

$\gamma = 2, \lambda = 0.001 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.44

$\gamma = 2, \lambda = 1 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.44

$\gamma = 2, \lambda = 1000 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.44

$\gamma = 0.125, \lambda = 0.001 \Rightarrow$ Ein$=$ 0.0, Eout$=$ 0.46

$\gamma = 0.125, \lambda = 1 \Rightarrow$ Ein$=$ 0.03, Eout$=$ 0.45

$\gamma = 0.125, \lambda = 1000 \Rightarrow$ Ein$=$ 0.2425, Eout$=$ 0.39

$\therefore$ the minimum Ein is 0.0, the combination expect ($\gamma = 0.125, \lambda = 1$) and ($\gamma = 0.125, \lambda = 1000$) can achieve the value.

12.

According to Q11, the combination ($\gamma = 0.125, \lambda = 1000$) can minimize Ein$=$ 0.39.

13.

$\lambda = 0.01 \Rightarrow$ Ein$=$ 0.3175, Eout$=$ 0.36

$\lambda = 0.1 \Rightarrow$ Ein$=$ 0.3175, Eout$=$ 0.36

$\lambda = 1 \Rightarrow$ Ein$=$ 0.3175, Eout$=$ 0.36

$\lambda = 10 \Rightarrow$ Ein$=$ 0.32, Eout$=$ 0.37

$\lambda = 100 \Rightarrow$ Ein$=$ 0.3125, Eout$=$ 0.39

$\therefore \lambda = 100$ 時可以讓 $E_{in} = 0.3125$ 達到最小

14.

According to Q13, $\lambda = 0.01, 0.1, 1$ 可以同時地讓 $E_{out} = 0$ 達到最小。

15.

$\lambda = 0.01 \Rightarrow$ Ein$= 0.3175$, Eout$= 0.36$

$\lambda = 0.1 \Rightarrow$ Ein$= 0.32$, Eout$= 0.36$

$\lambda = 1 \Rightarrow$ Ein$= 0.32$, Eout$= 0.36$

$\lambda = 10 \Rightarrow$ Ein$= 0.322$, Eout$= 0.36$

$\lambda = 100 \Rightarrow$ Ein$= 0.3225$, Eout$= 0.37$

$\therefore \lambda = 0.01$ 時可以讓 $E_{in} = 0.3175$ 達到最小，但使用 bagging 的方法整體而言的 E_{in} 會比 Q13 and Q14 的大一些。

16.

According to Q13, $\lambda = 0.01, 0.1, 1, 10$ 同時時可以讓 $E_{out} = 0.36$ 達到最小，經過 bagging 的過程可以讓 E_{out} 預測率平均表現較穩定、值也較低，其結果也符合我們所預期，bagging 的方式可以有效地緩和 overfitting 的問題。

17.

By the slides in class 3, we know:

$$g_{svm}(\mathbf{x}) = sign(\sum_{\text{SV indices n}} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b)$$
$$= c$$

s denotes as the support vector.

Now we are using $K_2 = K_1(\mathbf{x}, \mathbf{x}') + \kappa$. To make the result same, we then let $\alpha'_n = \alpha_n$ and $C' = C$, we can solve the dual problem based on K_2, α'_n and C':

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha'_n \alpha'_m y_n y_m K_2 - \sum_{n=1}^{N} \alpha'_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \alpha'_n y_n = 0$$
$$0 \leq \alpha'_n \leq C', \text{ for } n = 1, 2, \ldots N$$

the we know the optimal separating hyperplane will be:

$$
\begin{aligned}
g_{svm}(\mathbf{x}) &= sign(\sum_{\text{SV}} \alpha'_n y_n K_2(\mathbf{x}_n, \mathbf{x}) + b) \\
&= sign(\sum_{\text{SV}} \alpha'_n y_n K_2(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha'_n y_n K_2(\mathbf{x}_n, \mathbf{x}_s)) \\
&= sign(\sum_{\text{SV}} \alpha_n y_n (K_1(\mathbf{x}_n, \mathbf{x}) + \kappa) + y_s - \sum_{\text{SV}} \alpha_n y_n (K_1(\mathbf{x}_n, \mathbf{x}_s) + \kappa)) \\
&= sign(\sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}_s))
\end{aligned}
$$

, which is equivalent to the solution of original problem under optimal $\boldsymbol{\alpha}$

18.

Similar to the previous problem, now we are using $K_3(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + r(\mathbf{x}) + r(\mathbf{x}')$. We still let $\boldsymbol{\alpha}'' = \boldsymbol{\alpha}$ and $C'' = C$, we can solve the dual problem based on $K_3, \boldsymbol{\alpha}'' = \boldsymbol{\alpha}$ and $C'' = C$:

$$
\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha''_n \alpha''_m y_n y_m K_3 - \sum_{n=1}^{N} \alpha''_n
$$

$$
\begin{aligned}
\text{subject to} \quad & \sum_{n=1}^{N} \alpha''_n y_n = 0 \\
& 0 \le \alpha''_n \le C'', \text{ for } n = 1, 2, \ldots N
\end{aligned}
$$

the we know the optimal separating hyperplane will be:

$$
\begin{aligned}
g_{svm}(\mathbf{x}) &= sign(\sum_{\text{SV}} \alpha''_n y_n K_3(\mathbf{x}_n, \mathbf{x}) + b) \\
&= sign[\sum_{\text{SV}} \alpha''_n y_n K_3(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha''_n y_n K_3(\mathbf{x}_n, \mathbf{x}_s)) \\
&= sign(\sum_{\text{SV}} \alpha_n y_n (K_1(\mathbf{x}_n, \mathbf{x}) + r(\mathbf{x}) + r(\mathbf{x}')) + y_s \\
&- \sum_{\text{SV}} \alpha_n y_n (K_1(\mathbf{x}_n, \mathbf{x}_s) + r(\mathbf{x}) + r(\mathbf{x}'))] \\
&= sign[\sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}_s) \\
&+ \underbrace{\sum_{\text{SV}} \alpha_n y_n (r(\mathbf{x}) - r(\mathbf{x}_s))}_{=[r(\mathbf{x})-r(\mathbf{x}_s)]\cdot \sum_{\text{SV}} \alpha_n y_n = [r(\mathbf{x}) - r(\mathbf{x}_s)] \cdot 0 = 0} \quad] \\
&= sign(\sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha_n y_n K_1(\mathbf{x}_n, \mathbf{x}_s))
\end{aligned}
$$

, which is equivalent to the solution of original problem under optimal $\boldsymbol{\alpha}$