Machine Learning Techniques HW2

B04902004 王佑安

1. $\frac{\delta F}{\delta A} = \frac{1}{N}\sum_{n=1}^{N} \frac{-y_n z_n}{1+exp(y_n(Az_n+B))}$

   $= \frac{1}{N}\sum_{n=1}^{N} (-y_n z_n)\frac{exp(-y_n(Az_n+B))}{1+exp(-y_n(Az_n+B))}$

   $= \frac{1}{N}\sum_{n=1}^{N} -y_n z_n p_n$

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

   $= \frac{1}{N}\sum_{n=1}^{N} (-y_n)\frac{exp(-y_n(Az_n+B))}{1+exp(-y_n(Az_n+B))}$

   $= \frac{1}{N}\sum_{n=1}^{N} -y_n p_n$

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

2. $H(F) = \nabla^2 F(A,B) = \begin{bmatrix} \frac{\delta^2 f}{\delta A^2} & \frac{\delta^2 f}{\delta A \delta B} \\ \frac{\delta^2 f}{\delta B \delta A} & \frac{\delta^2 f}{\delta B^2} \end{bmatrix}$

   $= \sum_{n=1}^{N} \begin{bmatrix} \frac{\delta -y_n z_n p_n}{\delta A} & \frac{\delta -y_n z_n p_n}{\delta B} \\ \frac{\delta -y_n p_n}{\delta A} & \frac{\delta -y_n p_n}{\delta B} \end{bmatrix}$

   $= \sum_{n=1}^{N} \begin{bmatrix} -y_n z_n (\frac{-y_n z_n exp(y_n(Az_n+B))}{(1+exp(y_n(Az_n+B)))^2}) & -y_n z_n (\frac{-y_n exp(y_n(Az_n+B))}{(1+exp(y_n(Az_n+B)))^2}) \\ -y_n (\frac{-y_n z_n exp(y_n(Az_n+B))}{(1+exp(y_n(Az_n+B)))^2}) & -y_n (\frac{-y_n exp(y_n(Az_n+B))}{(1+exp(y_n(Az_n+B)))^2}) \end{bmatrix}$

   $= \sum_{n=1}^{N} \begin{bmatrix} -y_n z_n (\frac{y_n z_n}{(1+exp(y_n(Az_n+B)))^2} - \frac{y_n z_n}{1+exp(y_n(Az_n+B))}) & -y_n z_n (\frac{y_n}{(1+exp(y_n(Az_n+B)))^2} - \frac{y_n}{1+exp(y_n(Az_n+B))}) \\ -y_n (\frac{y_n z_n}{(1+exp(y_n(Az_n+B)))^2} - \frac{y_n z_n}{1+exp(y_n(Az_n+B))}) & -y_n (\frac{y_n}{(1+exp(y_n(Az_n+B)))^2} - \frac{y_n}{1+exp(y_n(Az_n+B))}) \end{bmatrix}$

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

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

3. Gaussian kernel $K(\mathbf{x}, \mathbf{x'}) = \exp(-\gamma ||\mathbf{x} - \mathbf{x'}||^2)$

   When $\gamma \to \infty \Rightarrow K(\mathbf{x}, \mathbf{x'}) = I$

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

4. $e_0 = \frac{1}{M}\sum_{m=1}^{M} \tilde{y}_m^2$

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

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

   $= \frac{M}{2}(s_t + e_0 + e_t)$

5. $E_{test}(\sum_{t=0}^{T} \alpha_t g_t) = \frac{1}{M}\sum_{m=1}^{M} (\sum_{t=0}^{T} \alpha_t g_t(\tilde{x}_m) - \tilde{y}_m)^2$

   $= \frac{1}{M}\sum_{m=1}^{M} (\sum_{t=0}^{T} \alpha_t g_t(\tilde{x}_m))^2 - \frac{2}{M}\sum_{m=1}^{M} (\sum_{t=0}^{T} \alpha_t g_t(\tilde{x}_m))\tilde{y}_m + \tilde{y}_m^2$

   $\Rightarrow \frac{\delta E_{test}}{\delta \alpha_t} = \frac{2\alpha_t}{M}\sum_{m=1}^{M} (g_t(\tilde{x}_m))^2 - \frac{2}{M}\sum_{m=1}^{M} g_t(\tilde{x}_m)\tilde{y}_m^2$

$= 2\alpha_t s_t - s_t + e_0 + e_t$

We can compute gradient as $\nabla E_{test}(\alpha) = [\frac{\delta E_{test}}{\delta \alpha_0} ... \frac{\delta E_{test}}{\delta \alpha_T}]$, then we can use gradient descent to find $\min_{\alpha 0, ... \alpha_T} E_{test}(\sum_{t=0}^{T} \alpha_t g_t)$

6. Let two generated examples at time $t = (a_t, 2a_t - a_t^2), (b_t, 2b_t - b_t^2)$
$h_t(x) = \frac{(2a_t - a_t^2) - (2b_t - b_t^2)}{a_t - b_t} x - \frac{(2a_t - a_t^2) - (2b_t - b_t^2)}{a_t - b_t} a_t + 2a_t - a_t^2$
$h_t(x) = (2 - a_t - b_t)x - (2 - a_t - b_t)a_t + 2a_t - a_t^2$
$h_t(x) = (2 - a_t - b_t)x + a_t b_t$
$\bar{g}(x) = (2 - E(a_t) - E(b_t))x + E(a_t b_t)$
$\bar{g}(x) = (2 - \frac{1}{2} - \frac{1}{2})x + \frac{1}{4}$
$\bar{g}(x) = x + \frac{1}{4}$

7. $u_+^{(2)} = u_+^{(1)} / \sqrt{\frac{0.87}{0.13}}$
$u_-^{(2)} = u_-^{(1)} * \sqrt{\frac{0.87}{0.13}}$
$u_+^{(2)} / u_-^{(2)} = (\frac{1}{N} / \sqrt{\frac{0.87}{0.13}}) / (\frac{1}{N} * \sqrt{\frac{0.87}{0.13}})$
$u_+^{(2)} / u_-^{(2)} = \frac{0.13}{0.87} = 0.15$

8. Each dimensions has $\{0, 1 ..., 6\} \cdot \{+1, -1\} = 12$ decision stumps.
$d = 2$ has $12 \cdot 2 = 24$ decision stumps.

9. $K_{ds}(\mathbf{x}, \mathbf{x}') = \sum_{t=1}^{|\mathcal{G}|} g_t(\mathbf{x}) g_t(\mathbf{x}')$
$= \sum_{t=1}^{|\mathcal{G}|} s_t \cdot sign(x_{i_t} - \theta_t) \cdot s_t \cdot sign(x'_{i_t} - \theta_t)$
$= \sum_{t=1}^{|\mathcal{G}|} sign(x_{i_t} - \theta_t) \cdot sign(x'_{i_t} - \theta_t)$
$= 2 \sum_{i=1}^{d} \sum_{m=0}^{M} sign(x_i - m) \cdot sign(x'_i - m)$
$= 2 \sum_{i=1}^{d} M - 2|x_i - x'_i|$

10. $K_{hi}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \min(x_i, x'_i)$
$\min(x_i, x'_i) = \sum_{m=0}^{M} \frac{sign(x_i - m) + 1}{2} \frac{sign(x'_i - m) + 1}{2}$
$= \sum_{m=0}^{M} h_{(1,i,m)}(x_i) h_{(1,i,m)}(x'_i)$
$\Rightarrow \phi_{hi} =$ select all decision stumps with $s = 1$
$\Rightarrow q_t = \begin{cases} 1 & s = 1 \\ 0 & s = -1 \end{cases}$

11.

E_{in}

$\lambda\backslash\gamma$	0.001	1	1000
32	**0.0**	**0.0**	**0.0**
2	**0.0**	**0.0**	**0.0**
0.125	**0.0**	0.03	0.2425

12.

E_{out}

$\lambda\backslash\gamma$	0.001	1	1000
32	0.45	0.45	0.45
2	0.44	0.44	0.44
0.125	0.46	0.45	**0.39**

13.

λ	0.01	0.1	1	10	100
E_{in}	0.3175	0.3175	0.3175	0.32	**0.3125**

14.

λ	0.01	0.1	1	10	100
E_{out}	**0.36**	**0.36**	**0.36**	0.37	0.39

15.

λ	0.01	0.1	1	10	100
E_{in}	0.3175	0.3175	0.32	0.3225	**0.3125**

$\lambda = 1, 10$ 的 E_{in} 比 P13 的高一點點，其他都一樣。

16.

λ	0.01	0.1	1	10	100
E_{out}	**0.36**	0.37	0.37	0.38	0.39

$\lambda = 0.1, 1, 10$ 的 E_{out} 比 P14 的高一點點，其他都一樣。

17. Soft-Margin SVM $g_{svm}(\mathbf{x}) = sign(\sum_{\text{SV indices n}} \alpha_n y_n (K_1(x_n, \mathbf{x}) + \kappa) + b)$

    By KKT condition, we know optimal α satisfy $\sum \alpha_n y_n = 0$

    $\Rightarrow g_{svm}(\mathbf{x}) = sign(\sum_{\text{SV indices n}} \alpha_n y_n (K_1(x_n, \mathbf{x}) + \kappa) + b)$

    $= sign(\sum_{\text{SV indices n}} \alpha_n y_n K_1(x_n, \mathbf{x}) + \kappa \sum_{\text{SV indices n}} \alpha_n y_n + b)$

    $= sign(\sum_{\text{SV indices n}} \alpha_n y_n K_1(x_n, \mathbf{x}) + b)$

    K_1 and K_2 have same optimal g_{svm} and α.

18. Let $K_1(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$, α is optimal solution of K_1

    Largange function of K_2:

    $\mathcal{L}(b, \mathbf{w}, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\Phi(x_n) + \sum_{m=1}^{M} \alpha_m y_m r(x_n) + \sum_{m=1}^{M} \alpha_m y_m r(x_m) + b))$

    $\frac{\delta \mathcal{L}}{\delta \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \Phi(x_n) = 0$(By KKT condition, $\mathbf{w} = \sum \alpha_n y_n z_n$)

    $\Rightarrow \alpha$ is still optimal solution of K_2