ML Technique Hw2 r06922114 陳啟元

1. $\nabla F(A,B) = \begin{bmatrix} \frac{\partial F}{\partial A} = \frac{1}{N}\sum \frac{1}{1+\exp()} \exp()\,(-y_n \cdot z_n) \\ \frac{\partial F}{\partial B} = \frac{1}{N}\sum \frac{1}{1+\exp()} \exp()\cdot(-y_n) \end{bmatrix}$

$= \begin{bmatrix} \frac{1}{N}\sum p_n\,(-y_n z_n) \\ \frac{1}{N}\sum p_n\cdot(-y_n) \end{bmatrix}$

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

(i) $\frac{\partial^2 F}{\partial A^2} = \frac{1}{N}\sum(-y_n \cdot z_n)\frac{\partial P_n}{\partial A}$

and $\frac{\partial P_n}{\partial A} = \frac{-y_n z_n \exp()\,[1+\exp()-\exp()]}{[1+\exp()]^2}$

$= -y_n z_n \cdot P_n \left(\frac{1+\exp()}{1+\exp()} - \frac{\exp()}{1+\exp()}\right) = -y_n z_n P_n (1-P_n)$.

so $\frac{\partial^2 F}{\partial A^2} = \frac{1}{N}\sum(-y_n \cdot z_n)^2 P_n(1-P_n) = \frac{1}{N}\sum z_n^2 P_n(1-P_n)$.

(ii) $\frac{\partial^2 F}{\partial A \partial B} = \frac{1}{N}\sum(-y_n z_n)\frac{\partial P_n}{\partial B}$

and $\frac{\partial P_n}{\partial B} = \frac{-y_n \exp()\,[1+\exp()-\exp()]}{[1+\exp()]^2} = -y_n P_n(1-P_n)$

so $\frac{\partial^2 F}{\partial A \partial B} = \frac{1}{N}\sum(-y_n)^2 z_n P_n(1-P_n) = \frac{1}{N}\sum z_n P_n(1-P_n)$

(iii) $\frac{\partial^2 F}{\partial B \partial A} = \frac{\partial^2 F}{\partial A \partial B}$

(iv) $\frac{\partial^2 F}{\partial B^2} = \frac{1}{N}\sum(-y_n)\frac{\partial P_n}{\partial B} = \frac{1}{N}\sum P_n(1-P_n)$.

3. $e^{-\gamma\|x-x'\|^2} \to 0$ as $\gamma \to \infty$ if $x \neq x'$

$e^{-\gamma\|x-x'\|^2} = 1$ if $x = x'$

So $K = I_N$

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

4. $e_t = \frac{1}{M} \sum_m (g_t(\tilde{x}_m) - \tilde{y}_m)^2 = \frac{1}{M}\left[\sum g_t^2(\tilde{x}_m) - 2\sum g_t(\tilde{x}_m)\tilde{y}_m + \sum (g_0(\tilde{x}_m) - \tilde{y}_m)^2\right]$

$= S_t - \frac{2}{M} \sum g_t(\tilde{x}_m)\tilde{y}_m + e_0$

$\Leftrightarrow \frac{(S_t - e_t + e_0)}{2} = \sum g_t(\tilde{x}_m)\tilde{y}_m$

5

$E_{test}(\sum_t \alpha_t g_t) = \frac{1}{M} \sum_m (\sum_t \alpha_t g_t(\tilde{x}_m) - \tilde{y}_m)^2$

To min this least square error, it's equivalent to solve the problem

$$\underbrace{\begin{bmatrix} g_0(\tilde{x}_1) & g_1(\tilde{x}_1) & \cdots & g_T(\tilde{x}_1) \\ \vdots & & & \vdots \\ g_0(\tilde{x}_M) & \cdots & \cdots & g_T(\tilde{x}_M) \end{bmatrix}_{M\times(T+1)}}_{G} \underbrace{\begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_T \end{bmatrix}_{(T+1)\times 1}}_{\alpha} = \underbrace{\begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_M \end{bmatrix}_{M\times 1}}_{y}$$

In linear algebra we know that the α minimizing E_{test} satisfies

$G^T G \alpha = G^T y$

if G is linearly independent. $G^T G$ is invertible, $\alpha = (G^T G)^{-1} G^T y$

if not, α have infinite many solu.

r06922114 陳政元.

6 $W_1 = \dfrac{(2x_2-x_2^2)-(2x_1-x_1^2)}{x_2-x_1} = 2-x_1-x_2$

We know that $w_1x_1+w_0 = 2x_1-x_1^2 \Leftrightarrow w_0 = 2x_1-x_1^2-(2-x_1-x_2)x_1 = x_1x_2$

By WLLN, $\frac{1}{T}\sum g(x) \xrightarrow{P} E(g(x))$

$\bar{g}(x) = E(w_1x+w_0) = xE(2-x_1-x_2)+E(x_1x_2) = x+\frac{1}{4}$

$\left(E(x_1)=E(x_2)=\frac{1}{2}\ ,\ \because x_1, x_2\ \text{iid}\ \therefore E(x_1x_2)=E(x_1)E(x_2)\right)$

7 $u_+^{(2)} = u_+^{(1)} \times$ incorrect rate

$u_-^{(2)} = u_-^{(1)} \times$ correct rate

In Adaboost Algorithm $u_1 = [\frac{1}{N} \cdots \frac{1}{N}]$

$\therefore u_+^{(2)} / u_-^{(2)} = \frac{1}{N}\times 13\% \,/\, \frac{1}{N}\times 87\% = \frac{13}{87}$

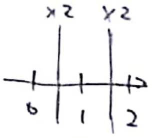
8 Excluding g s.t. all $g(x)=1$ or all $g(x)=-1$,

$|g_{s,i,\theta}| = |S| \times d \times M = 2dM$.

Including g we excluded before, we have $2dM+2$

9. Excluding g s.t. all $g(x)=1$ or all $g(x)=-1$

For jth feature, its corresponding $|g_{s,j,\theta}| = 2M$ (e.g.



$2\times2=4$)

where the first M kinds of $g_{s,j,\theta}$ are $\xleftarrow{-1}|\xrightarrow{1}$ i.e. $s=1$ — case (i)

the remaining are $\xleftarrow{1}|\xrightarrow{-1}$ i.e. $s=-1$ — case (ii)

In case (i):
$$\begin{cases} g_{1,j,\theta}(x)\cdot g_{1,j,\theta}(x') = -1\ , & x_j \le \theta \le x'_j \text{ or } x'_j \le \theta \le x_j \\ g_{1,j,\theta}(x)\cdot g_{1,j,\theta}(x') = 1\ , & \text{otherwise} \end{cases}$$

In case (ii): same as case (i)

So for jth feature, its inner product of $\overrightarrow{g_{s,j,\theta}}$ is two times of

$1[M-|x'_j-x_j|]+(-1)(|x'_j-x_j|) = (M-2|x'_j-x_j|)\times 2$

Inner product of $\overrightarrow{g_{s,i,\theta}}\ \overset{i=1\ldots d}{\vee}$ is $2+\sum_{i=1}^{d}(M-2|x'_i-x_i|)\times 2$ (Include what we excluded before)

16. Excluding all g s.t. $g(x)=1$ or $g(x)=-1$

For jth feature, the first M kinds of h_t are $\overset{0}{\leftarrow}|\overset{1}{\rightarrow}$ — case (i)

the remaining are $\overset{1}{\leftarrow}|\overset{0}{\rightarrow}$ — case (ii)

In case (i): $\begin{cases} h_t(x)\cdot h_t(x') = 1, & \theta \le x \le x' \text{ or } \theta \le x' \le x \\ h_t(x)\cdot h_t(x') = 0, & \text{otherwise} \end{cases}$

In case (ii): $\begin{cases} h_t(x)\cdot h_t(x') = 1, & x \le x' \le \theta \text{ or } x' \le x \le \theta \\ h_t(x)\cdot h_t(x') = 0, & \text{otherwise} \end{cases}$

We notice that in case (i), inner product of $\overrightarrow{h_t}$ is $\min(x, x')$

So for each feature, the first M $g_t = 1$, g_t is corresponding to h_t

the remaining $g_t = 0$

Including g we excluded before, and set their $g_t = 0$.

17. $\min_\alpha \frac{1}{2}\sum_n\sum_m \alpha_n \alpha_m y_n y_m (K(x_n, x_m) + k) - \sum_n \alpha_n$

$= \min_\alpha \left(\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m K(x_n,x_m) - \sum_n \alpha_n + \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m k\right)$

(We know that $\sum_n \alpha_n y_n = 0$, so $\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m k$

$= \frac{1}{2}\sum_n \alpha_n y_n k \underbrace{\sum_m \alpha_m y_m}_{=0} = 0$)

$= \min_\alpha \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m K(x_n, x_m) - \sum_n \alpha_n$

$w = \sum \alpha_n y_n z_n$, $b = y_s - \sum_{SV} \alpha_n y_n K(x_n, x_s)$ depends on α_n is hence the same.

18. $\min_\alpha \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m (K(x_n,x_m) + r(x) + r(x')) - \sum_n \alpha_n$

$= \min_\alpha \left(\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m K(x_n,x_m) - \sum_n \alpha_n + \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m (r(x_n) + r(x_m))\right)$

($\because \sum_n \alpha_n y_n = 0$, so $\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m (r(x_n)+r(x_m))$

$= \frac{1}{2}\sum_n \alpha_n y_n r(x_n) \underbrace{\sum_m \alpha_m y_m}_{=0} + \frac{1}{2}\underbrace{\sum_n \alpha_n y_n}_{=0} \sum_m \alpha_m y_m r(x_m) = 0$)

$= \min_\alpha \left(\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m K(x_n,x_m) - \sum_n \alpha_n\right)$

MLT HW2

r06922114 陳啟元

11

min $E_{in}=0$

r	λ	E_{in}	E_{out}
32	0.001	0	0.45
32	1	0	0.45
32	1000	0	0.45
2	0.001	0	0.44
2	1	0	0.44
2	1000	0	0.44
0.125	0.001	0	0.46
0.125	1	0.03	0.45
0.125	1000	0.2425	0.39

12. $r=0.125$ $\lambda=1000$ has min $E_{out}=0.39$

13.

λ	E_{in}	E_{out}
0.01	0.3175	0.36
0.1	0.3175	0.36
1	0.3175	0.36
10	0.32	0.37
100	0.3125	0.39

$\lambda=100$ has min $E_{in}=0.3125$

14. $\lambda=0.01$ or 0.1 or 1 has min $E_{out}=0.36$

15.

λ	E_{in}	E_{out}
0.01	0.3175	0.36
0.1	0.3125	0.37
1	0.32	0.37
10	0.3225	0.37
100	0.3125	0.39

$\lambda=0.1$ has min $E_{in}=0.3125$

Since square error is just a upper bound for 0/1 error

$E_{in}(g)$ will roughly smaller than $E_{in}(G)$ but not always.

16. $\lambda=0.01$ has min $E_{out}=0.36$

Since training data in problem 13 is the true data where we boootstrap from, $E_{out}(g)=\bar{g}\approx E_{out}(G)$ but in theory, g is more sensitive to data that's why we use G.