עידן מוסרי 305555179
עמית זוהר 204389043

מבוא ללמידה וניתוח של מידע רב – עבודה 2

① a) עלינו למזער את:

$$\lambda\|w\|^2+\sum_{i=1}^{m}\left[\ell^h(w,(x_i,y_i))\right]^2$$

נשקול בעיה שקולה:

$$\text{minimize } \lambda\|w\|^2+\sum_{i=1}^{m}\xi_i^2$$

$$\text{s.t. } \forall i,\ y_i\langle w,x_i\rangle \ge 1-\xi_i \text{ and } \xi_i\ge 0$$

הבעיות שקולות כיוון ש:

$$\ell^h(w,(x_i,y_i)) = \max\{0,1-y_i\langle w,x_i\rangle\}$$

ועבור $y_i\langle w,x_i\rangle \ge 1$: $\min \xi_i^2 = 0 = \ell^h(w,(x_i,y_i))$

ועבור $y_i\langle w,x_i\rangle < 1$: $\min \xi_i^2 = 1-y_i\langle w,x_i\rangle = \ell^h(w,(x_i,y_i))$

① b) נמיר לבעיית Quadratic program:

$$\text{minimize}_{z\in\mathbb{R}^n} \frac{1}{2}z^T\cdot H\cdot z+\langle u,z\rangle \quad \text{st } Az\ge v$$

נשקול:

$$z=(w(1),w(2),\dots,w(d),\xi_1,\xi_2,\dots,\xi_m)^T$$

$$H = 2(\underbrace{\lambda,\lambda,\dots,\lambda}_{d\text{ times}},\underbrace{1,1,\dots,1}_{m\text{ times}})\cdot \text{Identity matrix}_{(d+m)\times(d+m)}$$

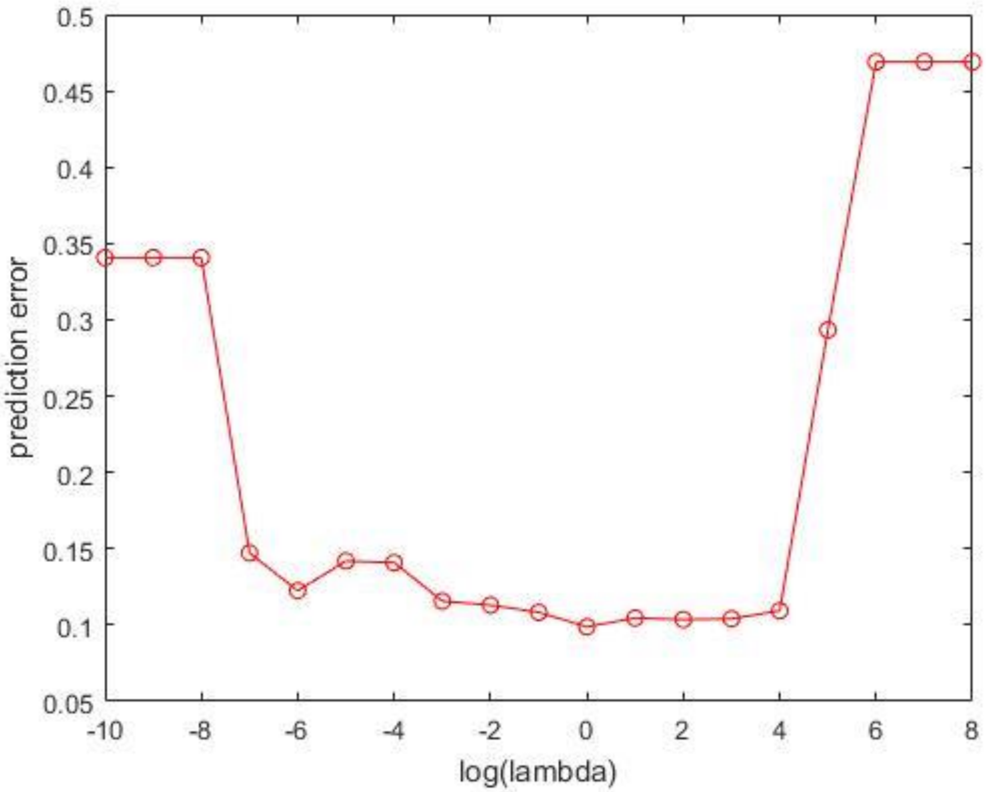
$$u=(\underbrace{0,0,\dots 0}_{d+m\text{ times}})^T \qquad v=(\underbrace{1,1,\dots,1}_{m\text{ times}},\underbrace{0,0,\dots 0}_{m\text{ times}})^T$$

כאשר:

$$A=\begin{bmatrix} \begin{matrix} y_1x_1\\ y_2x_2\\ \vdots\\ y_mx_m\end{matrix} & \begin{bmatrix} Id_{m\times m}\end{bmatrix} \\ \begin{bmatrix} Zeros_{m\times d}\end{bmatrix} & \begin{bmatrix} Id_{m\times m}\end{bmatrix}\end{bmatrix}_{2m\times(d+m)}$$

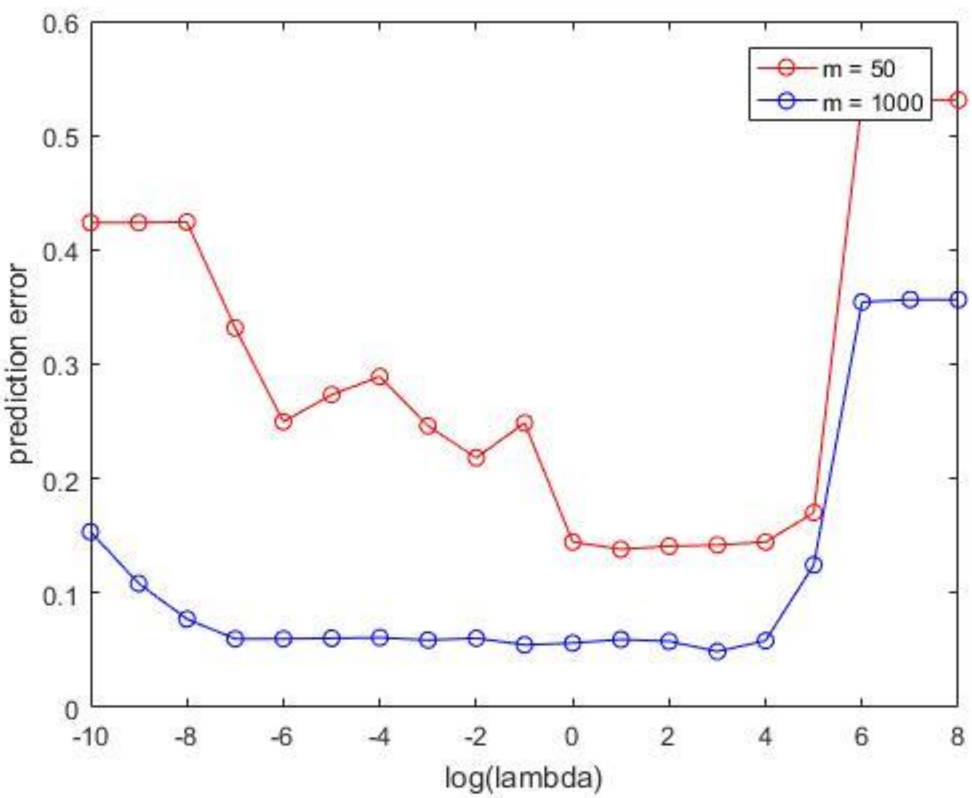
$$y_ix_i=(y_ix_i(1),y_ix_i(2),\dots,y_ix_i(d))$$

3. a.



0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
prediction error
-10
-8
-6
-4
-2
0
2
4
6
8
log(lambda)

3. b.



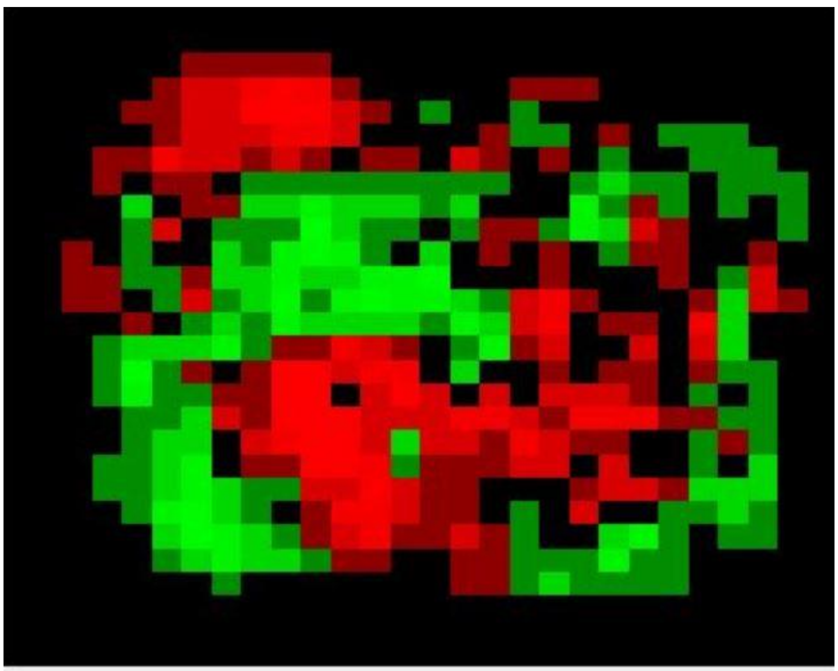
m = 50
m = 1000
0.6
0.5
0.4
0.3
0.2
0.1
0
prediction error
-10
-8
-6
-4
-2
0
2
4
6
8
log(lambda)

3. c. *Similarities: In both lines, the error is relatively low when 0 <= log(lambda) <= 4*

Differences: When m is 50, the error is high when lambda is very small or very large. In contrast, when m is 1000, the error is large only when lambda is large, but when lambda is small the error is still relatively low.

This can be explained from the trade-off in the value lambda mentioned in lecture 6 page 11: When lambda is small, we "pay" more for sample size, meaning high estimation error.
This is why when m=50, the error is high for small lambda values. When lambda is large, we pay more for the norm of w. This means that the minimization looks for w with a smaller norm, meaning large margin, which means a bigger hinge loss. This does not depend on the sample size, and this is why the error is relatively high in both cases when lambda is large.

3. d. *In the following heatmap, large positive values are represented by red color, and negative numbers with a large absolute value are represented by light green color. Numbers close to 0 are represented by black color.*



3. e. The green color roughly makes a 3 shape, whereas the red roughly makes a 5 shape (laid down, like the pictures in the sample). This makes sense because 3 was labeled as -1 and 5 was labeled as 1 in our code. For each pixel, if its value in the example is bigger than w's value, it will contribute towards a prediction as 5, and vice versa for 3.

Because of this, the dominant pixels in 5 are represented in red and the dominant pixels in 3 are represented in green.

4) a) $|w^{(t+1)}(i)| \le t$, נוכיח באינדוקציה על t שלכל i

<u>בסיס</u>: $t=0 \Rightarrow w^{(t+1)} = w^{(1)} = (0,0,0\ldots0) \Rightarrow |w^{(1)}(i)| = 0 \le t$

<u>הנחה</u>: באיטרציה ה-k, מתקיים: $\forall i \le d: \quad |w^{(k+1)}(i)| \le k$

<u>צעד</u>: יהי (x_j, y_j) הזוג שנבחר ע"י האלגו' באיטרציה ה-k+1.

נשים-לב כי לפי הגדרת הבעיה, לכל $i \le d$: $|y_j| \le 1, |x_j(i)| \le 1$.

לכן נקבל: $\forall i \le d: |w^{(k+2)}(i)| = |w^{(k+1)}(i) + y_j x_j(i)| \le |w^{(k+1)}(i)| + |y_j x_j(i)|$ (אי שוויון המשולש)

$(ה.א.) \le k + |y_j x_j(i)|$

$= k + |y_j| \cdot |x_j(i)| \le k+1$

(*) נשים-לב כי בכל צעד $w^{(t)}$ גדל או קטן בכמות של 1

4) b) ולכן לכל $i \le d$ $w^{(T)}(i)$ מס' שלם

נוכיח באינדוקציה על i שמתקיים $w^{(T)}(i) \ge 2^{i-1}$

<u>בסיס</u>: i=1. $w^{(T)}$ מפריד ולכן מתקיים:

(1) $y_1 \langle w^{(T)}, x_1 \rangle > 0$

(2) $y_1 \langle w^{(T)}, x_1 \rangle = w^{(T)}(1)$

$\Rightarrow w^{(T)}(1) > 0$

(לפי (*)) $\Rightarrow w^{(T)}(1) \ge 1 = 2^{i-1}$

<u>הנחה</u>: לכל $i \le k$ מתקיים: $w^{(T)}(i) \ge 2^{i-1}$

<u>צעד</u>: נפריד למקרים: (1) <u>k+1 זוגי</u>: $y_{k+1} \langle w^{(T)}, x_{k+1} \rangle = (-1)\left(\sum_{j=1}^{k} w^{(T)}(j) - w^{(T)}(k+1)\right)$

$= w^{(T)}(k+1) - \sum_{j=1}^{k} w^{(T)}(j) > 0$ (כי $w^{(T)}$ מפריד)

$\Rightarrow w^{(T)}(k+1) > \sum_{j=1}^{k} w^{(T)}(j) \underset{(ה.א.)}{\ge} \sum_{j=1}^{k} 2^{j-1} = 2^k - 1$

$(*) \Rightarrow w^{(T)}(k+1) \ge 2^k$

(2) <u>k+1 אי-זוגי</u>: $y_{k+1} \langle w^{(T)}, x_{k+1} \rangle = 1\left(-\sum_{j=1}^{k} w^{(T)}(j) + w^{(T)}(k+1)\right) = w^{(T)}(k+1) - \sum_{j=1}^{k} w^{(T)}(j) > 0$ (כי $w^{(T)}$ מפריד)

$\Rightarrow w^{(T)}(k+1) > \sum_{j=1}^{k} w^{(T)}(j) \underset{(ה.א.)}{\ge} \sum_{j=1}^{k} 2^{j-1} = 2^k - 1$

$(*) \Rightarrow w^{(T)}(k+1) \ge 2^k$

הוכחנו כי לכל $i \le d$, $w^{(T)}(i) \ge 2^{i-1}$ ולכן בפרט מתקיים:

$\forall i \le d: |w^{(T)}(i)| \ge 2^{i-1}$

4) c) ~~באיטרציה ה~~ לפי א', באיטרציה ה-t,

$w^{(t)}(d) \le t+1$

מצד שני, לפי ב' $w^{(T)}(d) \ge 2^{d-1}$

לכן יהיו לכל הפחות $2^{d-1}-1$ איטרציות