EE412 Foundation of Big Data Analytics, Fall 2018
HW4

Name: Park Jaeyoung

Student ID: 20170273

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

Answer to Problem 1

(a) Implement the gradient descent SVM algorithm described in textbook Ch. 12.3.4 by Python
As k-fold cross validation is used, running time takes 10 times more than just single task.

Selected C: 1
Selected eta: 10^-5
Accuracy: 0.81533

(b) Implement the parallel version of SVM as described in textbook in Ch. 12.3.6 using Spark. You may use the first approach where **w** and **b** are distributed to different mappers and updated in parallel before being averaged.
.
This takes about 1 hour, single train/test process takes about 5 min.
Selected C: 1
Selected eta: 10^-5
Accuracy: 0.74950

Answer to Problem 2

(a) Solve following problems in textbook.
* Exercises 4.4.1 and 4.4.2
I assume tail is number of zero at the end of bitwise expression, like example in textbook.
(a) (b) (c) tail lengths are in table below

x	2x+1 mod 32		tail length
3	7	00111	0
1	3	00011	0
4	9	01001	0
1	3	00011	0
5	11	01011	0
9	19	10011	0
2	5	00101	0
6	13	01101	0
5	11	01011	0

x	3x+7 mod 32		tail length
3	16	10000	4
1	10	01010	1
4	19	10011	0
1	10	01010	1
5	22	10110	1
9	2	00010	1
2	13	01101	0
6	25	11001	0
5	22	10110	1

x	4x mod 32		tail length
3	12	01100	2
1	4	00100	2
4	16	10000	4
1	4	00100	2
5	20	10100	2
9	4	00100	2
2	8	01000	3
6	24	11000	3
5	20	10100	2

(a) Estimated number of element: 1
(b) Estimated number of element: 16
(c) Estimated number of element: 16

If 2 can divide a (or a is even number), this affects on modular value. This limited possible modular value. Also if a and b is greater than or equal to 32, it would be inefficient. There is substitutable reminder of 32 in [0, 31].

*Exercise 4.5.3

i	Xi.element	Xi.value
1	3	2
2	1	3
3	4	2
4	1	2
5	3	1
6	4	1
7	2	2
8	1	1
9	2	1

*Exercise 4.6.1
(a) Real value: 3. Estimated value: 1 + 1 + 2/2 = 3
(b) Real value: 9. Estimated value: 1 + 1 + 2 + 4 + 4/2 = 10

(b) Implement the DGIM algorithm.
Implemented.

Answer to Problem 3

*Exercise 8.2.1

It would definitely sure that rent for n days and then buy the ski would be best algorithm.

Then, the skier would spend 10n + 100$. However, if skier just try ski one more time, rent one more day would be better and this case will spend 10(n+1) $

If a person goes for ski 10 or more days, then buying ski is off-line algorithm. In this case, skier spends 100$. Thus,

$$c \leq \frac{100}{10n+100}, \; c \leq \frac{10n+10}{10n+100}$$

While n = 9, c is maximum and the value is 10/19 or 0.526.

*Exercise 8.3.3.

Possible perfect matching: (1, c), (2, b), (3, d), (4, a) – so edge (1, a) and (3, b) must not selected.

So if (1, a) comes prior than (1, c) and (4, a) or (3, b) comes prior than (3, d) and (2, b) then matching cannot be perfect.

So by inclusion–exclusion principle, probability of greedy method being perfect matching can be calculated by inclusion–exclusion principle.

$$1 - \frac{1}{3} - \frac{1}{3} + \frac{1}{9} = \frac{4}{9}$$

Thus, those number of cases is 720 * 4/9 = 320.

*Exercise 8.3.3.

(a) By greedy algorithm, first 4 queries must match to advertisers.

Since both B and C bid on x and y, x and y can be matched either B or C in any possible ways.

(b) If queries are yyzzzz, then optimum off-line algorithm assigns 4 queries – first two y to B and third and fourth z to C. However, greedy algorithm can match first two y on C so that z cannot match to any of advertisers.