HW2 - Data Mining - due by 10 am 08 February 2019

Homework Policy:

1. You can discuss your HW with at most one other group. However, discussions are restricted to oral only. Written similarity is regarded as plagiarism. If you group discusses with other groups, you must acknowledge the discussion in your written solution.
2. Homework submission: one person in the group submitting the homework is ok, but all the names must appear on the paper of the written homework.
3. The written homework must be submitted by the due date. Please turn in a hard copy in class. The hard copy will be graded.
4. This list may be updated further.

Problem 1 (10 points) (Exercise 6.3.1 MMDS book) Here is a collection of twelve baskets. Each contains three of the six items 1 through 6.

$$\begin{array}{cccc} \{1,2,3\} & \{2,3,4\} & \{3,4,5\} & \{4,5,6\} \\ \{1,3,5\} & \{2,4,6\} & \{1,3,4\} & \{2,4,5\} \\ \{3,5,6\} & \{1,2,4\} & \{2,3,5\} & \{3,4,6\} \end{array} \tag{1}$$

Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set $\{i, j\}$ is hashed to bucket $i \times j \mod 11$.

(a) [5 points] Which buckets are frequent?

(b) [5points] Which pairs are counted on the second pass of the PCY Algorithm?

Problem 2 (10 points) (Exercise 6.3.4 MMDS book) Suppose we perform the PCY Algorithm to find frequent pairs, with market-basket data meeting the following specifications:

1. The support threshold is $10,000$.
2. There are one million items, represented by the integers $0, 1, ..., 999999$.
3. There are $250,000$ frequent items, that is, items that occur $10,000$ times or more.
4. There are one million pairs that occur $10,000$ times or more.
5. There are P pairs that occur exactly once and consist of two frequent items.
6. No other pairs occur at all.
7. Integers are always represented by 4 bytes.

8. When we hash pairs, they distribute among buckets randomly, but as evenly as possible; i.e., you may assume that each bucket gets exactly its fair share of the P pairs that occur once.

Suppose there are S bytes of main memory. In order to run the PCY Algorithm successfully, the number of buckets must be sufficiently large that most buckets are not frequent. In addition, on the second pass, there must be enough room to count all the candidate pairs. As a function of S, what is the largest value of P for which we can successfully run the PCY Algorithm on this data?

Problem 3 (5 points) Given three 2×2 matrices:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix} \tag{2}$$

Compute $\mathrm{Tr}(CAB)$ and $\nabla_A \mathrm{Tr}(CAB)$.

Problem 4 (10 points) Let A and B be an $n \times m$ and an $m \times n$ matrices, respectively.

(a) [5 points] Prove that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

(b) [5 points] Prove that $\nabla_A \mathrm{Tr}(AB) = B^T$.

Problem 5 (10 points) In this problem, we complete our toy example given in class. Given four points $(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, -1), (\begin{bmatrix} 2 \\ 1 \end{bmatrix}, -1), (\begin{bmatrix} 3 \\ 4 \end{bmatrix}, 1), (\begin{bmatrix} 4 \\ 3 \end{bmatrix}, 1)$ and four lines:

$$\begin{aligned} &(L_1) : w_1 + w_2 - 4 = 0 \qquad &&(L_2) : w_1 + w_2 - 5 = 0 \\ &(L_3) : w_1 + w_2 - 6 = 0 \qquad &&(L_4) : w_1 + 2w_2 - 6 = 0 \end{aligned} \tag{3}$$

(a) [5 points] Compute $\gamma(L_1), \gamma(L_2), \gamma(L_3), \gamma(L_4)$.

(b) From the computation in part (a), you would observe that $\gamma(L_2)$ has maximum margin. Show that for any separating line $(L) aw_1 + bw_2 + c = 0$ of the four points, $\gamma(L_2) \geq \gamma(L)$.

Problem 6 (5 points) (Exercise 12.3.3 MMDS book) The following training set obeys the rule that the positive examples all have vectors whose components have an odd sum, while the sum is even for the negative examples.

$$\begin{aligned} &(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, 1) \quad (\begin{bmatrix} 3 \\ 4 \end{bmatrix}, 1) \quad (\begin{bmatrix} 5 \\ 2 \end{bmatrix}, 1) \\ &(\begin{bmatrix} 2 \\ 4 \end{bmatrix}, -1) \quad (\begin{bmatrix} 3 \\ 1 \end{bmatrix}, -1) \quad (\begin{bmatrix} 7 \\ 3 \end{bmatrix}, -1) \end{aligned} \tag{4}$$

Suggest a starting vector $\mathbf{w}$ and constant b that classifies at least three of the points correctly and starting with the suggested vector, run (batch) gradient descent for two steps, with $C = 0.1$ and learning rate $\eta = 0.2$. Recall that the cost function here is:

$$J(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + C \sum_{i=1}^{6} \max(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) \tag{5}$$