CS246: Mining Massive Data Sets **Winter 2013**

Final

These questions require thought, but do not require long answers. Be as concise as possible.

You have **three** hours to complete this final. The exam has 22 pages and the total is 180 points so that you can pace yourself (1 point per minute).

You may use your notes, text and any material provided to you from the class, but may not use any outside resources. You may use your portable electronic devices, but any form of wired/wireless communication is prohibited, including connecting to the internet.

SCPD students who are taking this exam remotely should also follow the above rules and must be proctored by a SCPD approved proctor. You can take exam anytime during the week before the appointed deadline. The exam must be taken in a single continuous 3 hour time interval. Finished exams should be emailed to cs246.mmds@gmail.com by **Thursday 3/21/2013 5:00pm PST**. We will not accept any late exams under any circumstances.

Name and SUNetID: _____

I acknowledge and accept the Honor Code.

Signature: _____

Question	Total Points	Points
1	20	
2	8	
3	15	
4	10	
5	6	
6	10	
7	10	
8	5	
9	16	
10	15	
11	15	
12	15	
13	10	
14	10	
15	15	
Total	**180**	

1 MapReduce [20 points]

`DISTINCT` operator using Map-Reduce.

The `DISTINCT(X)` operator is used to return only distinct (unique) values for datatype (or column) `X` in the entire dataset .

As an example, for the following table `A`:

A.ID	A.ZIPCODE	A.AGE
1	12345	30
2	12345	40
3	78910	10
4	78910	10
5	78910	20

```
DISTINCT(A.ID) = (1, 2, 3, 4, 5)
DISTINCT(A.ZIPCODE) = (12345, 78910)
DISTINCT(A.AGE) = (30, 40, 10, 20)
```

(a) [**4 points**] Implement the `DISTINCT(X)` operator using Map-Reduce. Provide the algorithm pseudocode. You should use only one Map-Reduce stage, *i.e.* the algorithm should make only one pass over the data.

★ **SOLUTION:** The solution exploits MapReduce's ability to group keys together to remove duplicates. Use a mapper to emit X from each record as the key. The reducer simply emits the keys.

Pseudo code:

```
map(key, record):
    output (value, null) from column X of each input row
reduce(key, records):
    output key
```

(b) [**4 points**] The `SHUFFLE` operator takes a dataset as input and randomly re-orders it.

Hint: Assume that we have a function `rand(m)` *that is capable of outputting a random integer between* $[1, m]$.

Implement the `SHUFFLE` operator using Map-Reduce. Provide the algorithm pseudocode.

★ **SOLUTION:** All the mapper does is output the record as the value along with a random key. In other words, each record is sent to a random reducer. The reducer emits the values.

Pseudo code:

```
map(key, record):
    rand(m) = pick a random integer in [1, m]
    output (rand(n), record
reduce(key, records):
    for record in records:
        output record
```

(c) [4 points] What is the communication cost (in terms of total data flow on the network between mappers and reducers) for following query using Map-Reduce:

```
Get DISTINCT(A.ID from A WHERE A.AGE > 30 )
```

The dataset A has 1000M rows, and 400M of these rows have `A.AGE <= 30`.
`DISTINCT(A.ID)` has 1M elements. A tuple emitted from any mapper is 1 KB in size.

Hint: Use selection and distinct operations.

★ **SOLUTION:** Let, p = 1000M, r = 400M, q = 1M
There will be 2 jobs and the output of WHERE is chained to DISTINCT :
WHERE emits (p - r) tuples from the mapper
DISTINCT emits (p - r) tuples from the mapper
Total = 2(p - r) = 2(600M) = 1200M * 1KB = 1.12 TB (approx)

Total = (p-r) = 600M * 1KB, if the values are filtered in the mapper.

(d) [**4 points**] Consider the checkout counter at a large supermarket chain. For each item sold, it generates a record of the form `[ProductId, Supplier, Price]`. Here, `ProductId` is the unique identifier of a product, `Supplier` is the supplier name of the product and `Price` is the sales price for the item. Assume that the supermarket chain has accumulated many terabytes of data over a period of several months.

The CEO wants a list of suppliers, listing for each supplier the average sales price of items provided by the supplier. How would you organize the computation using the Map-Reduce computation model?

★ **SOLUTION:** SELECT AVG(Price) FROM DATA GROUP BY Supplier

Pseudo code :

```
map(key, record):
    output [ record(SUPPLIER), record(PRICE) ]
reduce(SUPPLIER, list of PRICE):
        emit [ SUPPLIER, AVG(PRICE) ]
```

For the following questions give short explanations of your answers.

(e) [1 point] True or false: Each mapper/reducer must generate the same number of output key/value pairs as it receives on the input.

★ **SOLUTION:** False. Mappers and reducers may generate any number of key/value pairs (including zero).

(f) [1 point] True or false: The output type of keys/values of mappers/reducers must be of the same type as their input.

★ **SOLUTION:** False. Mapper may produce key/value pairs of any type.

(g) [1 point] True or false: The input to reducers is grouped by key.

★ **SOLUTION:** True. Reducers input key/value pairs is grouped by the key.

(h) [1 point] True or false: It is possible to start reducers while some mappers are still running.

★ **SOLUTION:** False. Reducer's input is grouped by the key. The last mapper could theoretically produce key already consumed by running reducer.

2 Distance Measures [8 points]

Calculate the following distance measures between the two vectors,

$$v_1 = [0, 1, 1, 0, 0, 0, 1]$$
$$v_2 = [1, 0, 1, 0, 1, 0, 0]$$

(a) [2 points] Jaccard distance:

★ **SOLUTION:** $|v1 \cup v2| = 1$
$|v1 \cap v2| = 5$

Jaccard distance $= 1 - \frac{|v1 \cup v2|}{|v1 \cap v2|}$ = 1 - 1/5 = 4/5

(b) [2 points] Cosine distance (result in the form of $\arccos(x)$ is acceptable):

★ **SOLUTION:** $v1 \cdot v2 = 1$

$\|v1\| = \sqrt{3}$
$\|v2\| = \sqrt{3}$

Cosine distance $= arccos\left(\frac{1}{3}\right)$

(c) [4 points] For any two binary vectors Jaccard distance is always greater or equal than the Cosine distance. Argue why the statement is true or give a counter example.

3 Shingling [15 points]

(a) [**5 points**] Consider two documents A and B. Each document's number of token is $\mathcal{O}(n)$. What is the runtime complexity of computing A and B's k-shingle resemblance (using Jaccard similarity)? Assume that comparison of two k-shingles to assess their equivalence is $\mathcal{O}(k)$. Express your answer in terms of n and k.

★ **SOLUTION:** Assuming $n >> k$,
Time to create shingles $= O(n)$
Time to find intersection (using brute force algo) $= O(kn^2)$
Time to find union (using the intersection) $= O(n)$
Total time $= (kn^2)$

(b) [**5 points**] Given two document A and B, if their 3-shingle resemblance is 1, does that mean that A and B are identical? If so, prove it. If not, give a counterexample.

★ **SOLUTION:** NO. Example: $A = abab$, $B = baba$

(c) [**5 points**] Is there an $n \geq 1$ such that the following statement is true: Two documents A and B with n-shingle resemblance equal to 1 must be identical. If so, provide the smallest such n and show why. If not state how you can construct counterexamples for each n.

★ **SOLUTION:** NO. Example: Each document has $n + 1$ tokens. A has alternating tokens a, b starting with a while B has alternating tokens a, b starting with b.

4 Minhashing [10 points]

Suppose we wish to find similar sets, and we do so by minhashing the sets 10 times and then applying locality-sensitive hashing using 5 bands of 2 rows (minhash values) each.

(a) [**5 points**] If two sets had Jaccard similarity $\frac{1}{2}$, what is the probability that they will be identified in the locality-sensitive hashing as candidates (*i.e.* they hash at least once to the same bucket)? You may assume that there are no coincidences, where two unequal values hash to the same bucket. A correct expression is sufficient: you need not give the actual number.

★ **SOLUTION:** $1 - (1 - 0.5^2)^5 = 0.76$

(b) [**5 points**] For what Jaccard similarity is the probability of a pair of sets being a candidate exactly $\frac{1}{2}$? An expression involving radicals (*i.e.* square roots, etc.) is sufficient: you need not give the actual number.

★ **SOLUTION:** $\sqrt{1 - (0.5)^{0.2}}$

5 Random Hyperplanes [6 points]

Consider the following vectors in a 7-dimensional space.

$$\begin{array}{rcl} a & = & [\ 1, \quad 0, \ -2, \ 1, \ -3, \quad 0, \ 0\ ] \\ b & = & [\ 2, \quad 0, \ -3, \ 0, \ -2, \quad 0, \ 2\ ] \\ c & = & [\ 1, \ -1, \quad 0, \ 1, \quad 2, \ -2, \ 1\ ]. \end{array}$$

Suppose we use the random hyperplane method to compute sketches for the vectors, using the following 3 "random" hyperplanes:

$$\begin{array}{rcl} x & = & [\quad 1, \ 1, \quad 1, \quad 1, \quad 1, \quad 1, \quad 1\ ] \\ y & = & [\ -1, \ 1, \ -1, \quad 1, \ -1, \quad 1, \ -1\ ] \\ z & = & [\quad 1, \ 1, \quad 1, \ -1, \ -1, \ -1, \ -1\ ]. \end{array}$$

(a) [3 points]: Compute the sketches of each of the three vectors using these "random" hyperplanes.

Vector	Sketch
a	
b	
c	

(b) [3 points]: Estimate the angle between the vectors using their sketches.

Vectors	Angle
a,b	
b,c	
a,c	

★ **SOLUTION:** (5 minutes) Take the dot product of each of a,b,c with each of the normal vectors. If the dot product is positive, the corresponding component of the sketch has a 1, and if it is negative, the sketch component is -1.

Vector	Sketch
a	-1,1,1
b	-1,1,-1
c	1,-1,-1

The estimate of the angle between two vectors is computed from their sketches, by multiplying 180 degrees by the fraction of the positions in which the two sketches do not agree.

Vectors	Angle
a,b	60
b,c	120
a,c	180

6 Market Baskets [10 points]

A market-basket data set has 1 million baskets, each of which has 10 items that are chosen from a set of 100,000 different items (not all of which necessarily appear in even one basket). Suppose the support threshold is 100.

(Note: For partial credit, you should at least briefly explain your reasoning below.)

(a) [5 points]: What are the minimum and maximum numbers of frequent items? Explain your reasoning.

★ **SOLUTION:** (5 minutes) Minimum number of frequent items: 1;
Maximum number of frequent items: $100,000$ (Got from $1,000,000 * 10/100$).

(b) [5 points]: What are the minimum and maximum numbers of frequent pairs? Explain your reasoning.

★ **SOLUTION:** (3 minutes) Minimum number of frequent pairs: 0;
Maximum number of frequent pairs: $450,000$ (Got from $10 * 9/2 * 1,000,000/100$)

7 Counting Pairs of Items [10 points]

One method taught in class to represent the counts for the pair of items is to use a triangular matrix of dimension $n \times n$, where n is the number of items. To recap how this method works, we iterate over each pair of items, say $\{i, j\}$, and increment the count stored in the position $(i-1)(n-i/2)+j-i$ of the triangular matrix.

Now suppose we instead use a binary search tree to store the counts, where each node is now a quintuple $(i, j, c, \text{leftChild}, \text{rightChild})$. In this notation c is the count corresponding to item pair $\{i, j\}$, while leftChild and rightChild are pointers to children nodes.

Also suppose that there are n items, and p distinct pairs that actually appear in the data.

Assume all integers and pointers are 4 bytes each.

(a) [5 points] What is the memory requirement (as a function of n and p) when using the binary search tree?

★ **SOLUTION:** Each node needs $5 \times 4 = 20$ bytes. So, in total, we need $20p$ bytes.

(b) [5 points] Under what circumstances does it save space to use the above binary-search tree rather than a triangular matrix?

★ **SOLUTION:** The triangular matrix uses $4 \times \sum_{i=1}^{n} i = 2n(n-1)$ bytes, so it is more efficient to use the binary tree approach when $20p < 2n(n-1) \iff p < \frac{n(n-1)}{10}$.

8 Clustering [5 points]

Let us consider a one dimentional space ($\mathbb{R}$ for example). We wish to perform a hierarchical clustering of the points 1, 4, 9, 16, and 25.

Show what happens at each step until there are two clusters, and give these two clusters.

Your answer should be a table with a row for each step; the row should contain the members of the new cluster formed, and its centroid. More specifically, if you are merging a cluster $C_1 = \{x, y, z\}$ of centroid c_1 with a cluster $C_2 = \{p, q\}$ of centroid c_2, you should report $\{x, y, z, p, q\}$ in the table, as well as the centroid obtained with these 5 points).

★ **SOLUTION:** Here is the table we get. The two final clusters are $\{1, 4, 9\}$ and $\{16, 25\}$.

Step	**Cluster members**	**Centroid**
1	$\{1, 4\}$	2.5
2	$\{1, 4, 9\}$	4.67
3	$\{16, 25\}$	20.5

Table 1: Clustering steps.

9 Singular Value Decomposition [16 points]

$$M = \begin{array}{c} \\ \text{Diane} \\ \text{Ethan} \\ \text{Frank} \\ \text{Grace} \\ \text{Hank} \\ \text{Ingrid} \\ \text{Joe} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{MMDS} & \text{Machine Learning} & \text{Data Structures} & \text{Dynamics} & \text{Mechanics} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

Figure 1: Ratings of classes by students

Matrix M represents the ratings of Engineering courses taken by Stanford students. Each row of M represents the given student's set of ratings and the columns of M represent the classes. The labels for students and classes are provided along each row and column. An entry M_{ij} in M represents the student i's rating for class j. Now, the SVD decomposition of matrix M is found to be as follows:

$$USV^T = \begin{bmatrix} .18 & 0 \\ .36 & 0 \\ .18 & 0 \\ .90 & 0 \\ 0 & .53 \\ 0 & .80 \\ 0 & .27 \end{bmatrix} \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

Figure 2: SVD Decomposition of matrix M

It's pretty clear that there are two concepts of classes here: the first three are Computer Science classes while the last two are Mechanical Engineering classes.

Suppose a new student named Tony has the following reviews: 4 for Machine Learning, 5 for Data Structures, and 2 for Dynamics. This gives a representation of Tony in the class space as [0 4 5 2 0].

(a) [4 points] What is the representation of Tony in the concept space?

★ **SOLUTION:** Given that the class space for Tony, we can represent it as a vector, $a = [0, 4, 5, 2, 0]$. Then the mapping to concept space is $aV = [5.220, 1.420]$

(b) [4 points] Explain what this representation predicts about how much Tony would like the two remaining classes that he has not reviewed (MMDS and Mechanics).

★ **SOLUTION:** It shows that he prefers MMDS rather than Mechanics, since he has higher preferences toward CS classes.

Another student named Bruce has the following reviews: 5 for MMDS, 2 for Machine Learning, 4 for Dynamics, and 5 for Mechanics.

(c) [4 points] What is the representation of Bruce in the concept space?

★ **SOLUTION:** Similarly, we can represent Bruce's review as a vector, $b =$ [5, 2, 0, 4, 5], and obtain $bV =$ [4.06, 6.39]

(d) [4 points] Using similarity defined as $\text{sim}(a, b) = \frac{a \cdot b}{||a|| \cdot ||b||}$ where $||a||$ is the L_2 norm, calculate the cosine similarity of the two users using their concept space vectors.

★ **SOLUTION:** The similarity between Tony and Bruce is 0.739, which is simply obtained by setting the values in the right vectors.

10 Recommender Systems [15 points]

Consider a database containing information about movies: genre, director and decade of release. We also have information about which users have seen each movie. The rating for a user on a movie is either 0 or 1.

Here is a summary of the database:

Movie	Release decade	Genre	Director	Total number of ratings
A	1970s	Humor	D_1	40
B	2010s	Humor	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Humor	D_3	1

Table 2: Summary of the movie database.

Consider user U_1 is interested in the time period 2000s, the director D_2 and the genre Humor We have some existing recommender system R that recommended the movie B to user U_1.

The recommender system R could be one or more of the following options:

1. User-user collaborative filtering
2. Item-item collaborative filtering
3. Content-based recommender system

(a) [5 points] Given the above dataset, which one(s) do you think R could be? (If more than one option is possible, you need to state them all.) Explain your answer.

★ **SOLUTION:** R has to be either U-U or I-I collaborative filtering. It cannot be content based because if it were, movie C would have been predicted instead of B. Also B is the most popular movie. So that also indicates it could have been collaborative.

(b) [5 points] If some user U_2 wants to watch a movie, under what conditions can our recommender system R recommend U_2 a movie?

If R recommends a movie, how does it do it? If R cannot recommend a movie, give reasons as to why it can't. State any additional information R might want from U_2 for predicting a movie for this user, if required.

★ **SOLUTION:** Since U_2 is new user, it means we do not have any information about him. So R won't be able to recommend. But if user U_2 has watched any movie and if we have his interests, we can use this to recommend movies.

- User-user collaborative filtering: We can get U-I matrix with values in 0,1 from which user watched which movie and user Jaccard / Cosine similarity to get top similar users and then recommend movies watched by them. No recommendations on cold start.
- Item-item collaborative filtering: Same as above, just get similar items instead.

(c) [5 points] Item-item collaborative filtering is seen to work better than user-user because users have multiple tastes. But this also means that users like to be recommended a variety of movies.

Given the genre of each movie (there are 3 different genres in the dataset) and an item-item collaborative filtering recommender system that predicts k top-movies to a user (k can be an input to the recommender), suggest a way to find top 5 movies to a user such that the recommender will try to incorporate movies from different genres as well.

(Note: Explain in 3–5 lines maximum, no rigorous proof is required.)

★ **SOLUTION:** Several reasonable answers this question:

- Restrict the maximum number of items to be chosen from each category.
- Run the recommendation algorithm on the 3 subsets corresponding to each of the 3 categories, and merge the results.
- Predict the movies as usual, and then select a certain number of movies from each category in their order of appearance.
- Etc.

11 PageRank [15 points]

A directed graph G has the set of nodes $\{1, 2, 3, 4, 5, 6\}$ with the edges arranged as follows.

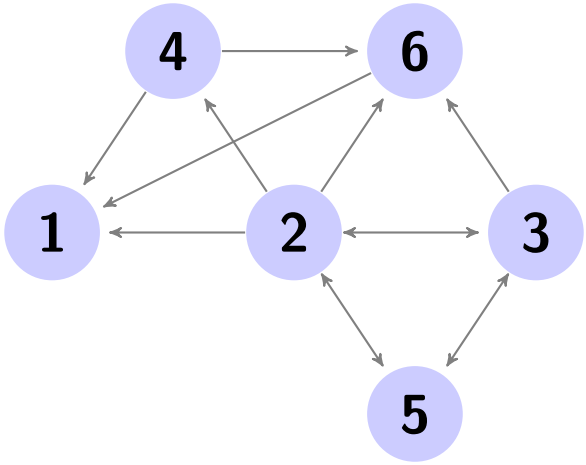

Figure 3: Graph G for Question 10.

(a) [5 points] Set up the PageRank equations, assuming $\beta = 0.8$ (jump probability $= 1-\beta$). Denote the PageRank of node a by $r(a)$.

★ **SOLUTION:**

$$r(1) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{2} \cdot r(4) + r(6) + \frac{1}{5} \cdot r(2)) + \frac{0.2}{6} \quad (1)$$

$$r(2) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{3} \cdot r(3) + \frac{1}{2} \cdot r(5)) + \frac{0.2}{6} \quad (2)$$

$$r(3) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{5} \cdot r(2) + \frac{1}{2} \cdot r(5)) + \frac{0.2}{6} \quad (3)$$

$$r(4) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{5} \cdot r(2)) + \frac{0.2}{6} \quad (4)$$

$$r(5) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{5} \cdot r(2) + \frac{1}{3} \cdot r(3)) + \frac{0.2}{6} \quad (5)$$

$$r(6) = 0.8(\frac{1}{6} \cdot r(1) + \frac{1}{5} \cdot r(2) + \frac{1}{3} \cdot r(3) + \frac{1}{2} \cdot r(4)) + \frac{0.2}{6} \quad (6)$$

(b) [5 points] Order nodes by PageRank, from lowest to highest.

(Note: No need to explicitly compute the scores. We are just asking for the ordering.)

★ **SOLUTION:** In descending order: 1,6,2,3,5,4

(c) [5 points] Set up the Hubs and Authorities equations on the graph G.

★ **SOLUTION:**

$$h(1) = 0 \tag{7}$$
$$h(2) = a(3) + a(5) \tag{8}$$
$$h(3) = a(2) + a(5) + a(6) \tag{9}$$
$$h(4) = a(1) + a(6) \tag{10}$$
$$h(5) = a(2) + a(3) \tag{11}$$
$$h(6) = a(1) \tag{12}$$
$$a(1) = h(4) + h(6) + h(2) \tag{13}$$
$$a(2) = h(3) + h(5) \tag{14}$$
$$a(3) = h(2) + h(5) \tag{15}$$
$$a(4) = h(2) \tag{16}$$
$$a(5) = h(2) + h(3) \tag{17}$$
$$a(6) = h(2) + h(3) + h(4) \tag{18}$$

12 Machine Learning [15 points]

(a) [5 points] Given a dataset of n data points and fixed C (slack penalty), the SVM formulation of the form (let us call it Model A):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i \cdot (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \\ & \xi_i \geq 0, \quad \forall i = 1, \ldots, n. \end{aligned}$$

Now consider a different formulation wherein the slack penalty is divided by n. That is (let us call this Model B):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i \cdot (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \\ & \xi_i \geq 0, \quad \forall i = 1, \ldots, n. \end{aligned}$$

Now we fix the value of C and vary the size of the dataset. How does the separating hyperplane found by Model A differ from Model B when n is large, *i.e.* when we have a large dataset?

★ **SOLUTION:** Model B will be less affected by outliers as compared to Model A

(b) [2 points] Consider a dataset of points $x_1, ..., x_n$ with labels $y_1, ..., y_n \in \{-1, 1\}$, such that the data is separable. We run a soft-margin SVM and a hard-margin SVM, and in each case obtain parameters w and b. Pick one that is true:

- A: The resulting w and b are the same in the two cases, hence the boundaries are the same.
- B: The resulting w and b can be different in the two cases, but the boundaries are the same.

- C: The resulting w and b can be different, and the boundaries can be different.
- D: None of the above.

★ **SOLUTION:** C

(c) [3 points] Let $x_1 = (1, 0), x_2 = (2, 1)$ and $x_3 = (2, 2)$ be the three support vectors for a SVM, such that x_1 and x_2 are positive examples and x_3 is a negative example. If x denote the first coordinate and y the second coordinate, the optimal margin boundary is (pick one that is true):

- A: $y = x + \frac{1}{2}$,
- B: $y = -x + \frac{1}{2}$,
- C: $y = x - \frac{1}{2}$,
- D: $y = -x - \frac{1}{2}$.

★ **SOLUTION:** C

(d) [2 points] In the soft-margin SVM we have variables ξ_i. If for some i, we have $\xi = 0$, this indicates that the point x_i is (pick one that is true):

- A: Exactly on the decision boundary,
- B: A support vector,
- C: Correctly classified,
- D: Incorrectly classified.

★ **SOLUTION:** C

(e) [3 points] Give 2 advantages that decision trees have over other machine learning methods.

★ **SOLUTION:**

- Easier to classify categorical features.
- For a dataset similar to the following, decision tree would be better. All data in first quadrant and third quadrant are positive. All data in second and fourth quadrants are negative.

13 AMS 3rd Moment Calculation [10 points]

The method of Alon, Matias, and Szegedy (AMS) that we covered in class was described as a way to compute second moments (surprise number), that is, $\sum_i {m_i}^2$, where m_i is the number of occurrences of the i^{th} value in a stream. It can also be used to compute higher order moments. If we want to compute third moments, that is, $\sum_i {m_i}^3$, then the algorithm for doing this is:

1. Pick a random place in the stream, say a place holding value a.
2. Count the number of occurrences of a from that time forward. Say a has occurred k times.
3. Let the value of X be $n(3k^2 - 3k + 1)$, where n is the length of the stream.

Show that the expected value of X is $\sum_i {m_i}^3$.

Hint: $k^3 - (k-1)^3 = 3k^2 - 3k + 1$.

★ **SOLUTION:** (15 minutes)

$$E[f(X)] = \frac{1}{n}\sum_{t=1}^{n} n(3C_t^2 - 3C_t + 1) \tag{19}$$

$$= \frac{1}{n}\sum_{a}\sum_{k=1}^{m_a} n(3k^2 - 3k + 1) \tag{20}$$

$$= \frac{1}{n} n \sum_{a}(3\sum_{k=1}^{m_a} k^2 - 3\sum_{k=1}^{m_a} k + m_a) \tag{21}$$

$$= \sum_{a}(3\frac{m_a(m_a+1)(2m_a+1)}{6} - 3\frac{(1+m_a)m_a}{2} + m_a \tag{22}$$

$$= \sum_{a} \frac{2m_a^3 + 3m_a^2 + m_a}{2} - \frac{3m_a^2 + 3m_a}{2} + m_a \tag{23}$$

$$= \sum_{a} m_a^3 \tag{24}$$

An alternate approach with the trick that $k^3 - (k-1)^3 = 3k^2 - 3k + 1$:

$$E[f(X)] = \frac{1}{n}\sum_{t=1}^{n} n(3C_t^2 - 3C_t + 1) \tag{25}$$

$$= \frac{1}{n}\sum_{a}\sum_{k=1}^{m_a} n(3k^2 - 3k + 1) \tag{26}$$

$$= \frac{1}{n} n \sum_{a}\sum_{k=1}^{m_a} (k^3 - (k-1)^3) \tag{27}$$

$$= \sum_{a} (m_a^3 - (m_a - 1)^3 + (m_a - 1)^3 - (m_a - 2)^3 + \ldots + 2^3 - 1^3 + 1^3 - 0^3)) \tag{28}$$

$$= \sum_{a} m_a^3 \tag{29}$$

14 Streams: DGIM [10 points]

Suppose we are maintaining a count of 1s using the DGIM method. We represent a bucket by (i, t), where i is the number of 1s in the bucket and t is the bucket timestamp (time of the most recent 1). Consider that the current time is 200, window size is 60, and the current list of buckets is:

$$(16, 148)\ (8, 162)\ (8, 177)\ (4, 183)\ (2, 192)\ (1, 197)\ (1, 200).$$

At the next ten clocks, 201 through 210, the stream has 0101010101.

What will the sequence of buckets be at the end of these ten inputs?

★ **SOLUTION:** (5 min) There are 5 1s in the stream. Each one will update to windows to be:
(1)

$$\begin{aligned} & (16,148)(8,162)(8,177)(4,183)(2,192)(1,197)(1,200),(1,202) \\ => \quad & (16,148)(8,162)(8,177)(4,183)(2,192)(2,200),(1,202) \end{aligned}$$

(2)

$$(16,148)(8,162)(8,177)(4,183)(2,192)(2,200),(1,202),(1,204)$$

(3)

$$\begin{aligned} & (16,148)(8,162)(8,177)(4,183)(2,192)(2,200),(1,202),(1,204),(1,206) \\ => \quad & (16,148)(8,162)(8,177)(4,183)(2,192)(2,200),(2,204),(1,206) \\ => \quad & (16,148)(8,162)(8,177)(4,183)(4,200),(2,204),(1,206) \end{aligned}$$

(4) Windows Size is 60, so (16,148) should be dropped.

$$(16,148)(8,162)(8,177)(4,183)(4,200),(2,204),(1,206),(1,208) => \quad (8,162)(8,177)(4,183)(4,200),(2,2$$

(5)

$$\begin{aligned} & (8,162)(8,177)(4,183)(4,200),(2,204),(1,206),(1,208),(1,210) \\ => \quad & (8,162)(8,177)(4,183)(4,200),(2,204),(2,208),(1,210) \end{aligned}$$

15 Streams: Finding The Majority Element [15 points]

Consider a stream of length n. Each item in the stream is an integer. We know that there exists a majority element in the stream, *i.e.* one of the integers repeats itself at least $\frac{n}{2}$ times (we don't know anything about the other items, they may repeat or may be unique).

Give an algorithm to find the majority element. You can use only a constant amount of extra space.

Also provide a brief correctness argument (need not be a mathematical proof, 3–5 lines will suffice) for your algorithm.

★ **SOLUTION:** Consider the following algorithm:

```
num = 0, count = 0
while stream.hasNext() do
  curInt = stream.next()
  if count == 0 then
    num = curInt, count = 1
  else if stream.next() == num then
    count++
  else
    count−
  end if
end while
output num
```

The main idea behind the solution is that if you remove two distinct elements from the stream, the answer remains unchanged.