COSC 290 Discrete Structures

Lecture 2: Sets

Prof. Michael Hay

Friday, Sep. 1, 2017

Colgate University

Plan for today

1. Sets
2. Computer science connection: frequent itemset mining
3. Reasoning about sets

Sets

Sets

A *set* is an unordered collection of objects.

- $Fruits := \{banana, apple, pear\}$
- Membership: $apple \in Fruits$ is true
- Cardinality: $|Fruits| = 3$
- Defining a set by...
  - Enumeration:

$$SingleDigitOdds := \{1, 3, 5, 7, 9\}$$

  - Abstraction:

$$SingleDigitOdds := \{x \in \mathbb{Z} : 0 \leq x \leq 10 \text{ and } x \bmod 2 = 1\}$$

Polling questions

Rules of the game.

- *(before class, you prepare yourself by reading the textbook and completing any problem sets)*
- I ask a question.
- You first answer it by yourself... no talking!
- Then discuss in groups of 3-4 students.
- Answer the question a second time.
- I will ask someone to answer and we will discuss.

Why?

Poll everywhere

On a device of your choice, go to `pollev.com/cosc290`

Set equality

Let A and B be sets. A and B are equal, denoted $A = B$, if A and B have exactly the same elements.

A little more formally, $A = B$ if every $x \in A$ is also an element of B **and** if every $y \in B$ is also an element of A.

Poll

$$R := \{1+1, 2+2, 3+3, 4+4\}$$
$$S := \{8, 4, 8, 2, 6, 4\}$$
$$T := \{2, 4, 6, 8\}$$

Which sets are equal?

a) R and S only
b) S and T only
c) R and T only
d) R, S and T
e) I'm lost.

Poll

$$R := \{2x : x \in \mathbb{Z}^{>0} \text{ and } x < 10\}$$
$$S := \{x \in \mathbb{Z}^{>0} : x \bmod 2 = 0 \text{ and } x < 10\}$$
$$T := \{2, 4, 6, 8\}$$

Which sets are equal? Choose the best answer.

a) R and S only
b) S and T only
c) R and T only
d) R, S and T
e) I'm lost.... what the heck is $\mathbb{Z}^{>0}$?

Set operations

$A = \{1, 3, 5, 7\}$ and $B = \{1, 2, 3, 4\}$ and let universe $\mathcal{U}$ be single digit positive integers.

- Union: $A \cup B =$?
- Intersection: $A \cap B =$?
- Difference: $A - B =$?
- Complement: $\sim A =$?
  (Note: complement always defined with respect to universe $\mathcal{U}$)

Venn diagram on the board.

Computer science connection: frequent itemset mining

Computer science connections

What: Connections between abstract concepts of this course and practical/interesting/fun problems that come up in computer science.

Why?

- help reinforce your understanding of concepts
- help you see value in learning these concepts

You are *not* expected to memorize the details of these examples.

Frequent Itemset Mining

Your internship at @WalmartLabs: analyze data on customer purchases.

Specifically, find all frequent itemsets. A frequent itemset is a collection of items that are frequently purchased together (by at least 1% of customers, for example).

Why might this be useful?

Let's work together to (a) formalize the problem mathematically, and (b) design an algorithm that solves this problem.

Input

Data on consumer purchases.

Representation: A list of n transactions $I_1, \ldots, I_n$ where each transaction I_i is represented as a *set* of items purchased (why a set?).

Example:

$$I_1 = \{\text{soy milk}, \text{coffee}\}$$
$$I_2 = \{\text{milk}, \text{orange juice}, \text{cocoa puffs}\}$$
$$\ldots$$
$$I_n = \{\text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk}\}$$

Support for an itemset

Suppose we have a particular itemset in mind, say $J := \{\text{coffee}, \text{soy milk}\}$.

We want to know the support for the itemset: the number of transactions in which the items in J were purchased together.

Example: suppose we have $n = 5$ transactions.

$$I_1 = \{\text{soy milk}, \text{coffee}\}$$
$$I_2 = \{\text{milk}, \text{orange juice}, \text{cocoa puffs}\}$$
$$I_3 = \{\text{soy milk}, \text{sugar}, \text{coffee}\}$$
$$I_4 = \{\text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk}\}$$
$$I_5 = \{\text{coffee}, \text{orange juice}\}$$

The support for J is 3.

Poll

You are given an itemset J and a transaction I_i. For example J might be $J := \{\text{tofu}, \text{coffee}\}$ and transaction I_i might be $I_i := \{\text{apples}, \text{tofu}, \text{bananas}, \text{coffee}\}$.

Which of the following expressions evaluates to true if the items in J were purchased together in transaction I_i? Choose an answer that will apply for any J and I_i and not just the specific example.

A) $J \in I_i$

B) $J \subset I_i$

C) $J \subseteq I_i$

D) $I_i \subseteq J$

E) None of the above / More than one of the above

Computing the itemset support

Input: Itemset J and a list of transactions $I_1, I_2, \ldots, I_n$
Output: Support s of itemset J.

```
1: s = 0
2: for i = 1 to n do
3:     if J ⊆ I_i then
4:         s = s + 1
5: return s
```


Finding all frequent itemsets

We know how to compute the support for a *given* itemset J. Itemset J is considered *frequent* if its support is at least $0.01 \times n$.

We want to find *all* frequent itemsets.

Poll

Suppose we have a collection of n transactions, $I_1, \ldots, I_n$. Example:

$$I_1 = \{\text{soy milk}, \text{coffee}\}$$
$$I_2 = \{\text{milk}, \text{orange juice}, \text{cocoa puffs}\}$$
$$\ldots$$
$$I_n = \{\text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk}\}$$

Let U represent the set of all items purchased in at least one transaction. Which of the following is a correct definition for U?

A) $U \coloneqq I_1 \cup I_2 \cup \cdots \cup I_n$
B) $U \coloneqq |I_1| + |I_2| + \cdots + |I_n|$
C) $U \coloneqq I_1 \cap I_2 \cap \cdots \cap I_n$
D) $U \coloneqq \{I_1, I_2, \ldots, I_n\}$
E) None of the above / More than one of the above

Powerset

The powerset of a set U denotes the set of all subsets of U.

Notation: We will use $\mathcal{P}(U)$ to denote the powerset of a set U.

What is the powerset of $\{1, 2, 3\}$? In other words, what is $\mathcal{P}(\{1, 2, 3\})$?

Poll

Suppose we have a collection of n transactions, $I_1, \ldots, I_n$. Example:

$$I_1 = \{\text{soy milk}, \text{coffee}\}$$
$$\ldots$$
$$I_n = \{\text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk}\}$$

Suppose that J is a frequent itemset. Consider this statement:

$$J \in \mathcal{P}(I_1 \cup I_2 \cup \cdots \cup I_n)$$

Choose the best answer:

A) This statement must be true.
B) This statement may be true.
C) This statement must be false.
D) This statement is not well defined.
E) My brain just exploded from too much notation.

Finding all frequent itemsets

Input: A list of transactions $I_1, I_2, \ldots, I_n$
Output: A set of all frequent itemsets.

```
1: FrequentItemSets = {}
2: U = I_1 ∪ I_2 ∪ ··· ∪ I_n                ▷ U is the set of all items
3: for all J ∈ P(U) do                      ▷ for each subset of items, check if its frequent
4:     s = 0
5:     for i = 1 to n do
6:         if J ⊆ I_i then
7:             s = s + 1
8:     if s > 0.01 × n then                 ▷ at least 1% of transactions
9:         FrequentItemSets = FrequentItemSets ∪ {J}
10: return FrequentItemSets
```


Efficiency considerations

Avoid checking *every* set in the powerset!

Useful property: If J is not frequent, then any J' that is a superset of J won't be frequent either.

Example, suppose

$$J \coloneqq \{\text{organic tofu}, \text{meat lover's frozen pizza}\}$$

and

$$J' \coloneqq \{\text{organic tofu}, \text{meat lover's frozen pizza}, \text{toothpaste}\}$$

If J is not frequent, then J' cannot be frequent either.

Towards a more efficient algorithm

Idea: Proceed in rounds

1. find frequent items (itemsets of size one)
2. find frequent pairs (itemsets of size two)
3. find frequent triples ...
4. ...

Only check itemsets whose subsets were found to be frequent in previous round.

Frequent itemset mining in the real world

Fun fact: Walmart found that when a hurricane/storm is forecast, people stock up on...

- Bottled water
- Flashlights
- Batteries
- Pop tarts
- Beer

Source: `http://abcnews.go.com/US/hurricanes/hurricane-irene-pop-tarts-top-list-hurricane-purchases/story?id=14393602`

Reasoning about sets

Poll

Let S and T be two sets with $|S| = m$ and $|T| = n$ and suppose we know that $m < n$. What is the smallest cardinality for $A \cup B$?

In other words, $|A \cup B|$ must be at least...

a) 0
b) m
c) n
d) $n + m$
e) $n \times m$

Poll

Let S and T be two sets with $|S| = m$ and $|T| = n$ and suppose we know that $m < n$. What is the smallest cardinality for $A - B$?

a) 0
b) m
c) n
d) $n + m$
e) $n \times m$

Poll

Let S and T be two sets with $|S| = m$ and $|T| = n$ and suppose we know that $m < n$. What is the largest cardinality for $A \cup B$?

a) 0
b) m
c) n
d) $n + m$
e) $n \times m$

Poll

Let S and T be two sets with $|S| = m$ and $|T| = n$ and suppose we know that $m < n$. What is the largest cardinality for $A \cap B$?

a) 0
b) m
c) n
d) $n + m$
e) $n \times m$

Problem set 2

(If time) Go over the problems on problem set 2