CPSC 304 – Administrative notes November 27 and 28, 2024

- Project:
  - Milestone 4: Project implementation – due November 29
    - You cannot change your code after this point!
  - Milestone 5: Group demo – week of December 2
    - Sign up for demos NOW – see Piazza – due end of Friday the 29th
    - A small number of you have had your project TA changed – check Canvas messages (groups 14, 34, 45, 56, 84)
  - Milestone 6: Individual Assessment – Due November 29
- Tutorials: basically project group work time/office hours
  - No tutorials the last week of class (so the TAs have more time to grade/do demos)
- Regular office hours end with the end of classes
  - Additional office hours will be posted
- Final exam: December 16 at 12pm! Osborne A

Lectures for the rest of the term

- Today we're covering data mining
- Friday, Nov 29/Tuesday Dec 3: Data mining ethics, beyond 304, research, and office hours (totally optional, nothing on final)
- Wednesday, Dec 4/Thursday Dec 5: a data mining exercise (no new material) and Review
- Friday, Dec 5: Office hours (the Wednesday/Friday section has one more lecture than Tuesday/Thursday, hence the lack of parity)

What topics would you like to see in review?

Please, give us your feedback!

- Through the Canvas Course Evaluation page. Direct link: https://seoi.ubc.ca/evaluations

- Click in when you are done:
  A. Done
  B. Working on it

TAs to evaluate – only evaluate TAs you interact with!

Tutorials:

Day	Time	TAs
Mon	13:00-14:00	Stellar & Bryan
Tue	14:00-15:00	Jack & Kate
Wed	12:00-13:00	Jessie & Linh
Wed	13:00-14:00	Pranjali & Christian
Wed	16:00-17:00	Stellar
Wed	18:00-19:00	Linh & Yikang
Thu	10:00-11:00	Pranjali
Thu	11:00-12:00	Bryan & Abdel
Fri	12:00-13:00	Jessie
Fri	13:00-14:00	Jessie & Yikang
Fri	14:00-15:00	Christian & Justine
Fri	15:00-16:00	Abdel

Lectures:

Day	TAs
Tue	Jianhao & Kate
Wed	Jack & Jianhao
Thu	Boqiao (Steven) & Jianhao
Fri	Boqiao (Steven) & Justine

CPSC 304 – August 8, 2024
Administrative notes

- Use tutorial times as open office hours
- Check your iClicker syncronization
- Milestone 5 is due August 9
- Milestone 6 is due August 9

- Final Exam
  - LIFE 2201 - Aug 16 at 8:30-11:00 am
  - bring your IDs, laptops, chargers and handwritten cheatsheets
    (one double-sided sheet per person)

How to study?

- Do the practice exercises without looking at the answer key
- Read over your notes and slides
- Look over the in-class exercises, tutorial questions, iclickers, and lecture recordings

Administrative Notes
March 27, 2024

- In-class Exercise DW 1 is due @10pm
- No class on April 1 (Easter)
- Upcoming project milestones:
    - Milestone 4: Implementation (Apr. 5). THIS IS WHEN YOUR CODE IS DUE.
    - Milestone 5: Demo
    - Milestone 6: Individual & peer evaluations
- Final exam: April 16 @7:00pm

Administrative Notes November 29 & 30, 2023

- Tutorials this week: project office hours
- Next week: no tutorials
- Reminder: standard office hours end this week
    - Special office hours coming before the final
- Upcoming project milestones:
    - Milestone 4: Implementation (Dec. 1). THIS IS WHEN YOUR CODE IS DUE.
    - Milestone 5: Demo (Dec 4-7)
        - Sign up by the 1st!!!
    - Milestone 6: Individual & peer evaluations (Dec. 1)
- Final exam: December 14 @8:30am.

CPSC 304 – December 6, 2022 Administrative Notes

- No Tutorial this week
- Final exam: Sunday, December 11 @3:30pm: Osborne A
  - Final exam office hours are listed on the office hour page. The page is still undergoing changes so check back regularly.
  - The final is cumulative
  - It will cover everything from the beginning of the term until the end of today other than PHP & JDBC.

How to study

- Do the practice exercises without looking at the answer key
- Read over your notes
- Look over the lecture recordings

Before we do our last topic, course evaluations

- Through the Canvas Course Evaluation page. Direct link: https://seoi.ubc.ca/evaluations
- Click in when you are done (pick an answer, any answer)

Data Mining: KDD Process, Frequent Itemsets, Association Rules, Frequent-Pattern Trees

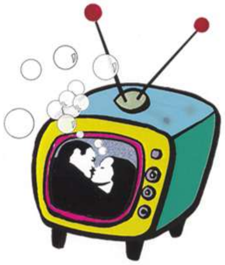
Text:

Chapter 26

Other references:

Data Mining: Concepts and Techniques, by Han and Kimber, Second Edition (Chapters 1 & 5)

Databases: the continuing saga



When last we left databases…

- We had decided they were great things
- We knew how to conceptually model them in ER diagrams
- We knew how to logically model them in the relational model
- We knew how to normalize our database relations
- We could write queries in different languages
- We'd processed things so people could analyze them

Now: What do we do with all that data?

Learning Goals

- Define the term *knowledge discovery*.
- Explain the general steps involved in the *knowledge discovery in databases* (KDD) process
- Comment on the benefits and challenges that data mining has when dealing with imperfect data quality, especially in large datasets (e.g., data mining can point out anomalies (outliers), optimize the use of human time, detect patterns in data (including patterns that are there just by chance).
- Explain the value of finding frequent itemsets and association rules. Provide some real-world examples of their use (e.g., retailing, biology).
- Explain the purpose of association rules.
- Apply the Apriori Algorithm and compute frequent itemsets and association rules (by hand, for a small dataset).

A Definition of Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns (business intelligence).

Understandable: can interpret and comprehend the patterns.

What about exceptions to patterns? (Outliers or anomalies)

Characteristics of Data Mining

- Key Characteristics
  - Large, multidimensional datasets
  - Efficient algorithms to “discover” knowledge
- What’s the connection with database systems?
  - Managing the data
    - Extract, Transform, and Load
    - There may be many distributed, heterogeneous sources.
  - Large numbers of records; many dimensions
  - Heavy emphasis on I/Os
  - Query evaluation and optimization considerations

Characteristics of Data Mining (cont.)

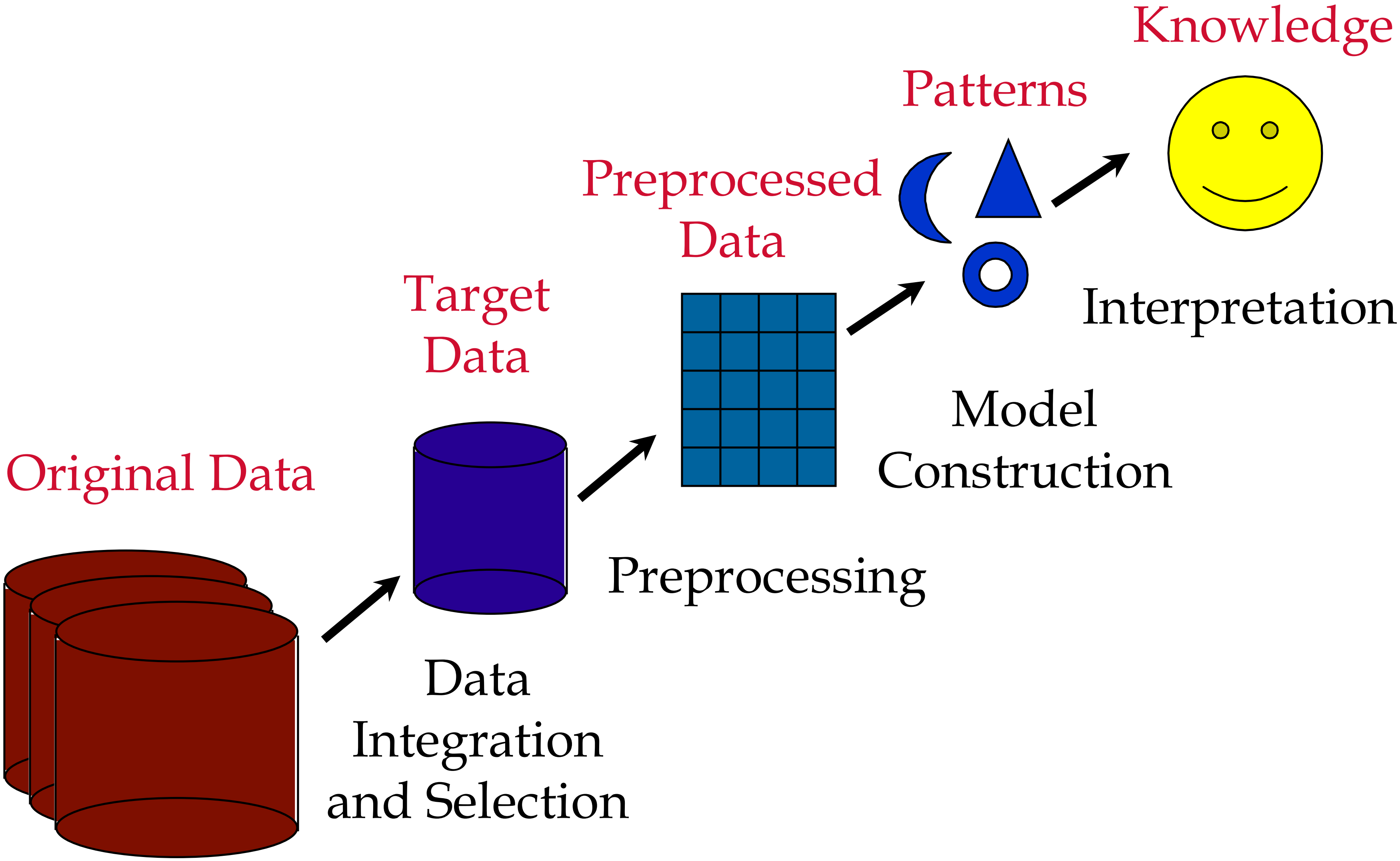
- Where does the data come from?
  - Databases, flat files, Excel, numerous applications, …
  - Data warehouses: Large, read-mostly, summarized data, that’s often downloaded from OLTP (transactional) systems
  - Sensor networks; cameras; satellites
  - Web logs
  - Transaction data (e.g., supermarkets)

Exercise: what is some data that that would be interesting to mine?

Knowledge Discovery and Data Mining (KDD) process

- Data pre-processing (may take 60%+ of the effort)
  - Data selection: Identify target datasets and relevant fields
  - Extraction and Integration
  - Data cleaning
    - Remove noise and outliers
    - Data transformation
    - Create common units
    - Generate new fields
- Data Mining
  - Model Construction
- Evaluation
  - Model Validation
  - Take Action: Execute Business Strategy

The KDD Process



Market Analysis and Management

- Where are the data sources for market analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaints, demographics, surveys
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
    - Higher income groups
    - Lower income groups
    - BMW owners
    - New baby in the household
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.

Market Analysis and Management (cont.)

- Cross-market analysis
  - Associations or correlations between product sales
    - Amazon.com: People who bought X also bought Y.
  - Predictions based on association information (e.g., Recommender Systems)
    - Suppose you watch some movies.
    - You rank/rate those movies (e.g., You liked “Titanic” and “Slumdog Millionaire”; but, you didn’t like “Apollo 13” and “Terminator 2”).
    - People who rated movies similarly to the way you did, share a common profile or “cluster”.
      - So, let’s find other points (people) in this cluster, and then find out what other movies they liked (and that you haven’t seen yet).
      - Chances are that you’ll like their choices, too; but, if not, we can further refine the cluster, and establish a narrower profile. Iterate.
      - Compare this to the old way of “film critics” deciding which are the “best” movies for you to see.

Market Analysis and Management (cont.)

- Customer Profiling
  - Data mining can tell you what types of customers buy what products (e.g., via clustering or classification).
- Identifying Customer Requirements
  - Identify the best products for different customers.
  - Try to predict the factors that will attract new customers, or retain current customers.
- Summary Information
  - Statistical summaries (e.g., central tendency and variation)

Fraud Detection and Management

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Examples
  - Auto insurance: Detect a group of people who stage accidents to collect on insurance
  - Money laundering: Detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network—FINCEN)
  - Medical insurance: Detect professional "patients" and rings of doctors and rings of references
  - Telephone calls: Detect patterns that deviate from an expected norm.
  - Retail shrink: Analysts estimate that 38% of retail shrink is due to dishonest employees.

Data Mining Techniques

- Supervised Learning
  - Classification and regression
    - Classify into pre-defined groups, based on previous information
  - Some examples
    - Handwritten postal codes on envelopes
    - Insurance or medical expert systems (decision trees)
      - If (16 ≤ age ≤ 25 & car_type = (“sports car” or “Hummer”) then:
        - risk_status = HIGH
      - If (symptomA = … and symptomB = …) then …

Data Mining Techniques (cont.)

- Unsupervised Learning
  - Clustering (many types of clustering algorithms exist)
  - Automatic discovery of clusters or commonalities
- Dependency Modeling
  - Associations, correlations, summarization, causality
- Outliers (Deviation Detection)
  - Which observations appear to be “out of place” or suspicious?
- Trend Analysis
  - How to spot trends, and how to spot trend changes

Association Rules

- One type of data mining techniques is Association Rules
- An example rule is “people who buy diapers tend to buy beer”
- This is useful for stores because they can improve stock
- They’ve also been used in many areas, including medical diagnoses, protein sequence composition, health insurance claim analysis and census data

Here’s the plan

- Stores keep track of all the items that people bought at a time
- By looking at all of the different purchases, we can figure out which items were bought at the same time
- Then we can figure out which one was the “cause” and which one was the “effect”

Let’s look at some sample data

- Each row is a *transaction* – one person’s grocery order
- So in T2 the person bought Sushi and Bread

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

- Now we need to decide whether there are any items that people tend to buy when they buy other items. We refer to this as a rule (e.g., diapers → beer is one rule (not supported by this data!)).

Support

- Informally: support measures if items appear together a lot of times
- Formally: A rule X→Y holds with support *sup* if sup% of transactions contain *X* **AND** *Y*.

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

- For example, {Chicken, Ramen, Milk} occurs with 3/7= 42% support

Support question

What is the support of Sushi → Bread (express as a fraction)?

(Reminder: a rule X→Y holds with support *sup* if sup% of transactions contain *X* **AND** *Y*.)

A. 3/7

B. 3/4

C. 4/7

D. None of the above

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

Support question

What is the support of Sushi → Bread (express as a fraction)?

(Reminder: a rule X→Y holds with support *sup* if sup% of transactions contain *X* **AND** *Y*.)

[✓ A. 3/7]

B. 3/4

C. 4/7

D. None of the above

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

Confidence

Informally: confidence measures which items suggest the others will be there, too.

Formally: A rule X→Y holds with *confidence* conf% if conf% of transactions that contain X also contain Y

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

Ramen → Bread [conf = 1/3 = 33%]

Ramen, Chicken → Milk [conf = 3/3 = 100%]

Confidence question

What is the confidence of Sushi → Chicken (express as a fraction)?

(Reminder: A rule X→Y holds with *confidence* conf% if conf% of transactions that contain X also contain Y)

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

A. 3/7

B. 3/4

C. 3/5

D. None of the above

Confidence question

What is the confidence of Sushi → Chicken (express as a fraction)?

(Reminder: A rule X→Y holds with *confidence* conf% if conf% of transactions that contain X also contain Y)

A. 3/7

[✓ B. 3/4]

C. 3/5

D. None of the above

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

So when is a rule valid?

A rule is valid if its support is above a given threshold (minimum support) and its confidence is over another given threshold (minimum confidence).

A frequent itemset is a set of items that has at least minimum support

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

In this example, {chicken, milk, ramen} is a frequent itemset if the minimum support is less than 3/7.

Support clicker question

What is the support of Apple → Corn (express as a fraction – no need for math)?

(Reminder: a rule X→Y holds with support *sup* if sup% of transactions contain *X* **AND** *Y*.)

A. 1/4

[✓ B. 2/4]

C. 3/4

D. 4/4

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

The Apriori algorithm key idea

Calculating association rules on terabytes of data can be sloooowww. The slowest part is *counting the support*.

The Apriori algorithm speeds things up based on the observation that each subset of a frequent itemset must *also* be a frequent itemset

For example, since rice only appears one time, it can’t appear two or more times with anything else.

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori exercise, part the first

Start by finding the support of all itemsets of size 1

Support: {apple} = 2/4
{corn}
{dates}
{rice}
{tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

What is the support for corn?

a. 1/4

b. 2/4

c. 3/4

[✓ d. 4/4]

Apriori exercise, part the first

Start by finding the support of all itemsets of size 1

Support:

{apple} = 2/4
{corn} = 4/4
{dates} = 3/4
{rice} = 1/4
{tuna} = 3/4

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori exercise, part the first

Start by finding the support of all itemsets of size 1

Support:
{apple} = 2/4
{corn} = 4/4
{dates} = 3/4
{rice} = 1/4
{tuna} = 3/4

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

If the minimum support is 50%, can any itemset containing rice *ever* be a frequent itemset?

A. Yes

[✓ B. No]

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}
{apple, dates}
{apple, rice}
{apple, tuna}
{corn, dates}
{corn, rice}
{corn, tuna}
{dates, rice}
{dates, tuna}
{rice, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Because {rice} only occurs once, anything including {rice} can’t occur 2 or more times, so we can ignore itemsets including {rice}.

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}
{apple, dates}
~~{apple, rice}~~
{apple, tuna}
{corn, dates}
~~{corn, rice}~~
{corn, tuna}
~~{dates, rice}~~
{dates, tuna}
~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Because {rice} is not frequent, anything including {rice} is not frequent, so we can ignore itemsets including {rice}.

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}

{apple, dates}

~~{apple, rice}~~

{apple, tuna}

{corn, dates}

~~{corn, rice}~~

{corn, tuna}

~~{dates, rice}~~

{dates, tuna}

~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Exercise: count support for remaining itemsets.

Apriori round 2:
Find all frequent itemsets of size 2

Support for all possible itemsets of size 2:

{apple, corn} = 2/4
{apple, dates} = 2/4
~~{apple, rice}~~
{apple, tuna} = 1/4
{corn, dates} = 3/4
~~{corn, rice}~~
{corn, tuna} = 3/4
~~{dates, rice}~~
{dates, tuna} = 2/4
~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Exercise: what are the frequent itemsets of size 2?

Apriori round 2:
Find all frequent itemsets of size 2

All frequent itemsets of size 2:

{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori round 3:
Find all frequent itemsets of size 3

Given frequent itemsets of size 2

{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Without counting support, what are all the possible frequent itemsets of size 3?

(remember: in order for an itemset to be possibly frequent, all subsets of it must be frequent, e.g., {apple, corn, rice} is not a possible frequent itemset because {rice} is not a frequent itemset)

Apriori round 3:
Find all frequent itemsets of size 3

Given frequent itemsets of size 2

{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Without counting support, what are all the possible frequent itemsets of size 3?

A. {apple, corn, dates}

B. {apple, corn, dates}, {apple, corn, tuna}, {corn, dates, tuna}

C. {apple, corn, tuna}, {corn, dates, tuna}

[✓ D. None of the above]

If option A included {corn, dates, tuna} it would be correct.

Apriori round 3:
Find all frequent itemsets of size 3

Great! Now count support for the remaining itemsets

(Exercise):

{apple, corn, dates}
{corn, dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori round 3:
Find all frequent itemsets of size 3

Great! Now count support for the remaining itemsets (Exercise):

{apple, corn, dates} = 2/4
{corn, dates, tuna} = 2/4

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Since 2/4 = 50%, both are frequent

Apriori round 4:
Find all frequent itemsets of size 4

Exercise: given frequent itemsets of size 3 :
{apple, corn, dates}
{corn, dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Without counting support, what are the possible frequent itemsets of size 4?

[✓ A. Nothing]

B. {apple, corn, dates, tuna}

C. {apple, corn, dates, tuna}, {apple, corn, dates, rice}

Apriori example: done!

The whole list of frequent itemsets for this example is:

{apple}
{corn}
{dates}
{tuna}
{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori example: done!

Frequent itemsets

{apple}
{corn}
{dates}
{tuna}
{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

Itemsets we counted support for:

{apple}
{corn}
{dates}
{rice}
{tuna}
{apple, corn}
{apple, dates}
{apple, tuna}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

All possible itemsets:

{apple}
{corn}
{dates}
{rice}
{tuna}
{apple, corn}
{apple, dates}
{apple, rice}
{apple, tuna}
{corn, dates}
{corn, rice}
{corn, tuna}
{dates, rice}
{dates, tuna}
{rice, tuna}
{apple, corn, dates}
{apple, corn, rice}
{apple, corn, tuna}
{corn, dates, rice}
{corn, dates, tuna}
{dates, rice, tuna}
{apple, corn, dates, rice}
{apple, corn, dates, tuna}
{corn, dates, rice, tuna}
{apple, corn, dates, rice, tuna}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

F_1 = {{corn}, {dates}, {tuna}}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Reminder: F_1 = {{corn},
{dates}
{tuna}}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

{C_2 = {{corn, dates},
{corn, tuna},
{dates, tuna}}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

Support: {{corn, dates} =3/4
{corn, tuna} =3/4
{dates, tuna}=2/4

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

F_2 = {{corn, dates}, = 3/4
{corn, tuna} = 3/4
~~{dates, tuna}} = 2/4~~

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Reminder: F_2 = {{corn, dates}
{corn, tuna}}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

C_3 =

[✓ A. nothing]
B. {corn, dates, tuna}
C. None of the above

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
    1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
    2. Count support of items in C_{k+1}
    3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Answer = F1 ∪ F2 ∪ F_3 =
{{corn}, {dates}, {tuna},
{corn, dates},{corn, tuna}}