CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

Lecture 7: Data Mining

Announcements

- Lisa's OH this week:
  - Saturday 3-4.30pm, Lathrop Tech Lounge
- Midterm next Tuesday, during class time
  - If you have a conflict, please tell us by the end of today

Midterm Logistics

- 70 minutes during class time (May 2, 1.30-2.40pm, be on time!)
- Very similar to homework questions
  - Write and read spreadsheet functions
  - Write and read SQL queries
  - Write python code to solve data problems
  - No pandas will be covered
- No notes allowed

Recap of Last Week

- Introduction to Python, an all-purpose programming language
- Covered Python fundamentals
  - Types, Variables
  - Data structures
  - Functions
  - Loading data
  - Plotting
- Correlation & Causation
- Privacy

Learning Goals for Today

- Python Fundamentals Continued
  - Learn how to build nested data structures
  - Know when to use which data structure

- Introduction to Data Mining
  - Understand the intuition, goals and techniques of data mining
  - Learn about frequent itemsets and association rules

Question 1:

You want to keep track of your friends' food preferences, so you can surprise them for dinner! Each friend can like multiple food items.

What data structure(s) would you use to store this information?

Answer:
A dictionary with the names of friends as keys,
and lists of food items as values.
→ *List in Dictionary*

Question 2:

You want to keep track of the customers of your business. For each customer, you would like to save: Customer ID, First name, Last name, Number of purchases.

You would like to look up your customers by their Customer ID.

What data structure(s) would you use to store this information?

Answer:
Outer dict has customer ID as key, and the inner dict as value. Inner dict has other customer atttributes as keys, and corresponding values.
→ *Dictionary in Dictionary*

Let’s see it in code!

Please download lecture_7.zip from the course website and open

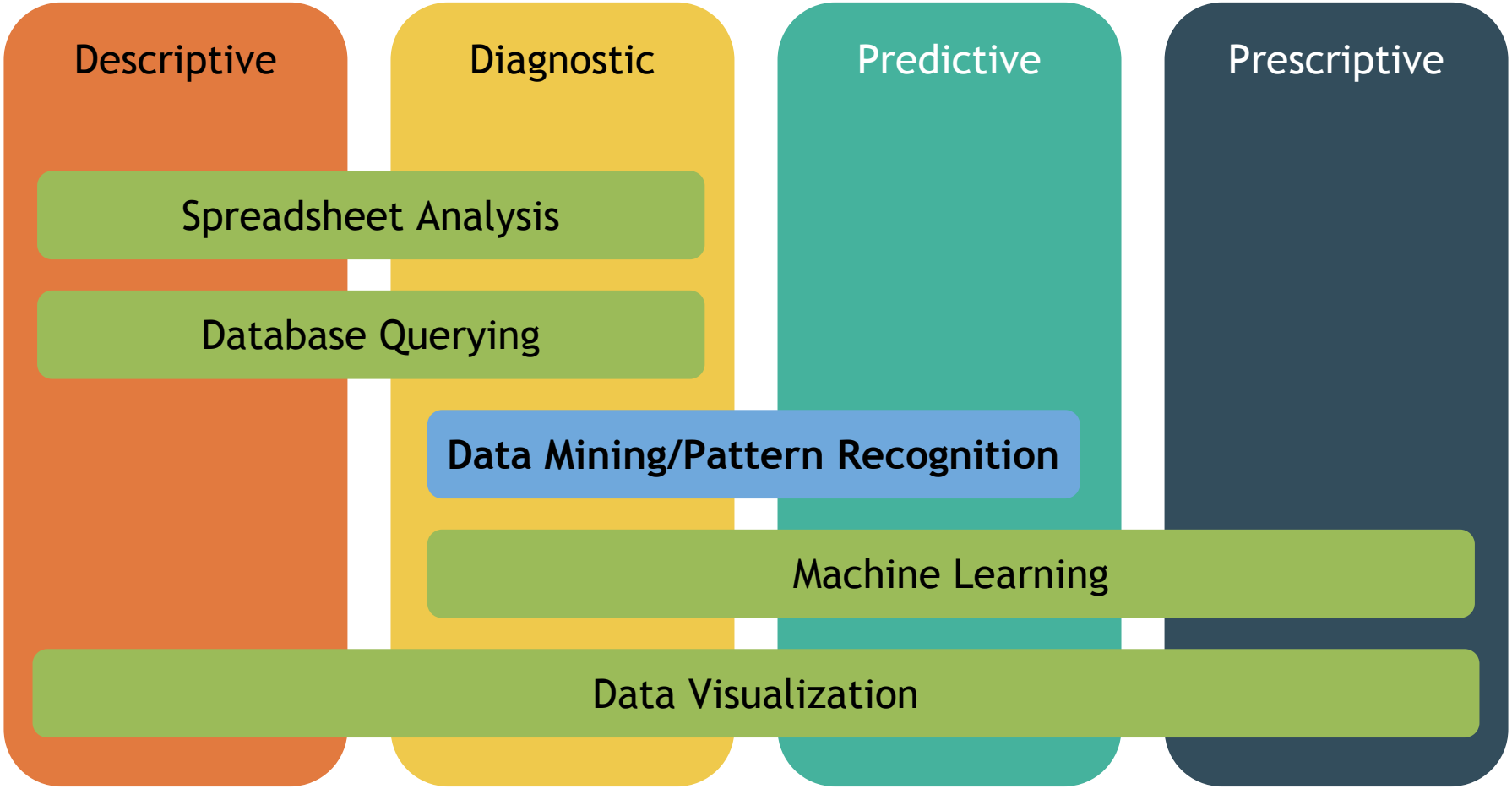
Lecture_7_More_Python_Examples.ipynb

Additional Python Resources

- Book: Python in a Nutshell by A. Martelli
  - A more in-depth introduction to Python
  - Review of programming concepts (e.g. object-oriented programming etc.)

- Tutorialspoint (free)
  - More Python tutorials than you ever need
  - Good as a reference while you write programs
  - https://www.tutorialspoint.com/python/

Tools & Techniques



Data Mining

- “Looking for patterns in data”
- Contrast data mining to:
  - ***Data(base) operations:*** -Executing specific operations or queries over data
  - ***Machine learning:*** **-**Using data to make inferences or predictions

Fill your baskets!

Imagine you walk into *Safebuy* Supermarket.

Enter the items you want to buy here:bit.ly/my_basket

Fill your baskets!

Now imagine you are the CEO of Safebuy and you are trying to increase sales by positioning items together that are frequently bought together.

Go to http://bit.ly/basket_stats to see what your customers' baskets look like.

Discuss with a partner how you would go about finding patterns.

Remember this from lecture 1? This should be funnier now :D

The Tale of Beer and Diapers

Data Mining on Market Basket Data

Patterns we will look at:

1. Frequent Itemsets
    - Sets of items that occur together frequently in transactions
2. Association Rules
    - When Set1 of items appear in a transaction, Set2 often occurs in the same transaction

This type of analysis originated with retail data, but can be applied on other types of data as well.

Market basket is not just market basket

Come up with at least 3 examples where this type of analysis might be useful.

For each example, write down what the items are, and what the transaction is.

For example:

Items: Groceries, Transaction: grocery cart

Items: Words, Transaction: document

Market-Basket Data

Items	**Transaction**
Groceries	Grocery cart
Online goods	Online shopping cart
College courses	Student transcript
students	Party
Symptoms	Patient
Menu items	Customer selection
Words	Document
Movies	Netflix user

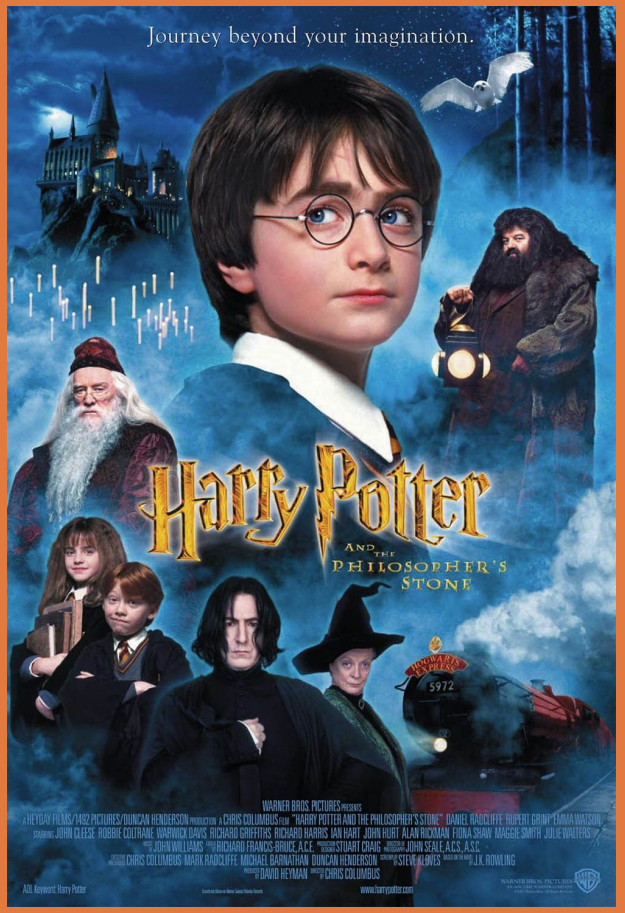
Which symptoms often appear together in patients?

Sore throat and coughing

Which words often appear together?

“Data” and “Mining”

Which movies are often watched together?


Journey beyond your imagination.
Harry Potter
AND THE PHILOSOPHER'S STONE


LEGENDEN OM
NARNIA
LØVEN, HEKSA OG KLESSKAPET
EN FILMBILDEBOK

You get the idea, it's what you already do intuitively.

Let's make it more rigorous!

Frequent Itemsets

Sets of items that occur together frequently in transactions.

Definition:

- *Support*: # transactions containing set / total # transactions
  - Look for sets with support > support-threshold
- *Support-threshold*: Min support to count as a “frequent itemset”.
  - Parameter value can change based on application, risk etc.

Frequent Itemsets

Intuition:

- *Support*: How much evidence/support for this particular itemset can you find in the data?

- *Support-threshold*: What is the minimum amount of support required to call an itemset a “frequent itemset”?

Let’s code it up!

Open Lecture_7_Data_Mining_Starter_Code.ipynb

Association Rules

- Set1 → Set2: When Set1 occurs in a transaction, Set2 often occurs in the same transaction.
- The rule is an implication.

Definitions:

- *Confidence*: # transactions with both Set1 & Set2 / total # transactions with Set1
- *Support*: # transactions containing Set1 / total # transactions.
  - Set1 should be a frequent itemset
  - Look for Set1 with support > *support-threshold*

Frequent Itemsets vs Association Rules

- Customers who buy eggs and butter are likely to buy milk as well
- Frequent itemset of 3: eggs, butter, milk
- Association rule: (eggs, butter) → milk

Differences:

➢ Frequent itemsets: no directionality encoded in sets.

➢ Association rules: implication has directionality. Set1 → Set2 does not mean Set2 → Set1.

Summary

- Python data structures review + more examples
- Introduction to Data Mining
- Market Basket Analysis
  - Frequent Itemsets
  - Association Rules

Please take 2 minutes to give us feedback here:
http://bit.ly/cs102_feedback