Applied Data Analysis (CS401)

ada

Lecture 1
Intro to ADA
20 Sep 2023

EPFL

Robert West

dlab

Important websites

http://ada.epfl.ch

Your main entry point. All materials linked from there.

https://go.epfl.ch/ada2023-ed

Main communication channel. Sign in with your EPFL email address (or simply access via Moodle).

https://github.com/epfl-ada/2023

Used for homework, project, and final exam.

ABOUT ME

- Born in Ingolstadt, Bavaria, Germany

- Education:

School	Location	Degree	# seasons
TUM	Germany	Diplom	4
McGill	Canada	MS	2
Stanford	USA	PhD	1

- Assistant professor at EPFL since Dec. '16
- Heading Data Science Lab

dlab
011011o

ABOUT ME

- Born in Ingolstadt, Bavaria, Germany

- Education:

School	Location	Degree	# seasons
TUM	Germany	Diplom	4
McGill	Canada	MS	2
Stanford	USA	PhD	1

- Assistant professor at EPFL since Dec. '16
- Heading Data Science Lab

dlab
011011o

⇒ Call me Bob

bob
10010

My path toward data science

– Born in Ingolstadt, Bavaria, Germany

MY RESEARCH

Develop neat algorithms

Distill raw data into insights into people & the world

Give back to the real world (⇒ applications)

DATA SCIENCE

Leverage large datasets

Address real-world issues

Draw meaningful conclusions from "found data" (a.k.a. observational studies)

When necessary, generate my own data (human computation, crowdsourcing)

BUZZWORDS

Machine learning

Data mining

Social & information network analysis

Computational social science

Natural language processing

Human computation, crowdsourcing

Information retrieval

Data analysis

“... the process of **inspecting, cleaning, transforming, and modeling data** with the goal of **discovering useful information**, suggesting conclusions, and supporting decision-making.”

“Data analysis has multiple facets and approaches, encompassing **diverse techniques** under a variety of names, **in different** business, science, and social science **domains.**”

Applied data analysis

- This course is about **breadth,** not depth
- *"What methods, principles, and tools are out there?"*, rather than *"How can I become an expert in deep learning for computer vision applied to images of cats?"*
- Data science is a fast-paced, shifting field
- Obsessing on one tool or technique won't pay off in a few years
- Be ready to explore and keep learning on your own
- Goal of this class: Enable you to conduct a full-fledged data science project from start to finish
- That said, depth matters, too…

Complementary courses:
Machine learning
NLP
DIS
Data viz

Let's abbreviate this course as **Ada,** not A-D-A, in honor of Ada Lovelace, "the world's first computer programmer."

https://en.wikipedia.org/wiki/Ada_Lovelace

Place
Ada Lovelace

Syllabus

- Handling data
  - “Slicing and dicing”: obtaining, preparing, juggling data
- Visualizing data
  - Exploration of data, communication of results
- Describing data
  - How to support (and be suspicious of) claims about data
- Regression analysis for disentangling data
  - How to disentangle datasets with correlated variables
- Causal analysis of observational data
  - How to deal with “found data”
  - Correlation != causation

Syllabus (cont'd)

- Learning from data
  - Supervised learning
  - Unsupervised learning
  - Applied aspects of machine learning
- Handling specific types of data
  - Handling text data
  - Handling network data
- Scaling to massive data

Grading

- 30% **Homework assignments (2)**
  - Involving skills required from data scientists
  - Groups of 5 students (may switch groups after Homework 1)
  - Homework of 2017, 2018, 2019, 2020, 2021, 2022
- 30% **Final exam** (date TBD)
  - Mini data analysis project
  - Done on laptop, individually, on campus
  - Final exams of 2017, 2018, 2019, 2020, 2021, 2022
- 25% **Project** (more details soon)
  - Your own freestyle data analysis
  - Done in groups of 5 students (same as for homework)
  - Milestones spread throughout the semester
  - Projects of 2017, 2018, 2019, 2020, 2021, 2022
- 15% **Quizzes**
  - Weekly, online (on Moodle), 5 questions in 10 minutes

Grading

- 30% **Homework assignments (2)**
  - Involving skills required from data scientists
  - Groups of 5 students (may switch groups after Homework 1)
  - Homework of 2017, 2018, 2019, 2020, 2021, 2022
- 30%
  -
  -
  -
- 25%
  -
  - Done in groups of 5 students (same as for homework)
  - Milestones spread throughout the semester
  - Projects of 2017, 2018, 2019, 2020, 2021, 2022
- 15% **Quizzes**
  - Weekly, online (on Moodle), 5 questions in 10 minutes

This class will be hard work, but it will get you a job.

Grading (cont'd)

- To obtain a meaningful grade distribution, scaling/shifting will be applied to each of {homework, project, exam, quizzes} before taking weighted average (standard practice at EPFL)
- While intermediate grades are a good indication of where you stand, remember there might be some wiggle
  - → Don't rely on intermediate grades to decide whether you can afford to skip the exam etc.

Meeting logistics: Lectures

- **Wednesdays 8:15–10:00**
- If you want to see it live, come to class! (No live streaming)
- Lectures are also recorded and made available after class

Meeting logistics: Lab sessions

- **Fridays 13:15–14:45**
- In person only:
  - BCH 2201 (next to UNIL)
  - CE 1 106
- Labs are complementary to lectures, not simply more detail on same
- You solve exercises that we make available the day before, can ask questions and get help from assistants
- In certain weeks: homework/project office hours (probably on Zoom, in parallel to exercises)
- Weekly online quizzes on Moodle (see next slide)

Weekly quizzes

- Held online on Moodle
- 5 questions, to be answered within 10 minutes of starting
- During first 15 minutes (13:15–13:30) of each Fri lab session
- Week 1 (this week): no quiz
- Week 2 (next week): Quiz 1: no real questions, just to let you get familiar with the setup
- Week 3: Quiz 2: the first quiz that counts
- Quiz i (held in week $i + 1$) is about lecture material of week i
- Goal:
  - Engage continuously with course material
  - Think (not just find right slide)
- No one is expected to get everything right
- Your two lowest-scoring quizzes from entire semester won't count

Project

- We'll provide a number of datasets
- You need to form and pitch a crisp project idea
- Free to combine with other datasets (at your own risk)
- Goal: not a loose collection of results – tell a story with the data!
  - Data stories of 2017, 2018, 2019, 2020, 2021, 2022
  - Nice example data story

Homework and projects: GitHub

- De-facto standard for managing and sharing code
- All students in this class need a GitHub account
- Homework and project submissions done via GitHub

Main communication channel: ed

- Class forum, available via Moodle
- Also accessible directly, outside of Moodle: https://go.epfl.ch/ada2023-ed
  (sign in using the same email address as for Moodle)
- Central place to ask all class-related questions
- Mandatory! We'll send important announcements on Ed only
- Help each other (without cheating, of course)

Watch-at-home videos

- Throughout the semester, we'll release videos with supplemental information; e.g.,
  - Intro to lab sessions (already available!)
  - Project instructions
  - Homework 1 postmortem
  - Homework 2 postmortem

General note on communication

- Multiple platforms used in ADA for various tasks (as in real life): Ed, GitHub, Google docs, ADA website
- To avoid confusion,
  - familiarize yourself with communication guidelines
  - all materials will be linked from the website as a central point of entry: https://ada.epfl.ch
  - all discussions will take place on Ed

Group registration

- Must form teams within 2 weeks, starting now (in time for release of Homework 1)!
- Get started immediately to find 4 teammates
- By Fri 6 Oct 23:59, complete the registration form (to be done by each team member individually): https://forms.gle/kEX3asGKPKRFBhN2A
- Can change team after Homework 1 (but try to avoid it)

Prerequisites

Basics of

- probabilities and stats
- databases
- programming
  - You won’t survive if you can’t program
  - Homework, exam: Python required
  - Project: up to you, but we support only Python
  - Brush up your Python skills (many great online courses out there)

Python environments

- Homeworks and exams to be done as Jupyter Notebooks
- You will submit a pre-executed .ipynb file
  - We don't care how you produce it
  - Option 1: local Python installation (e.g., Anaconda + JupyterLab)
  - Option 2: Google Colab = notebook hosted by Google
  - Option 3: noto = notebook hosted by EPFL
- To get started: come to Friday's lab session ("Exercise 0")
- "Homework 0": do it yourself at home after lab session (optional, not graded)
- Doing Homework 0 is the best way of making sure you're set up correctly for later homework, project, exam

Python++

jupyter

scikits learn

machine learning in Python

Spark

Lightning-Fast Cluster Computing

seaborn

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

POLLING TIME

- **"What is your prior experience with Python?"**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Instructor

Bob West

Teaching assistants (TAs) = PhD students

Akhil Arora

Manoel Horta

Saibo Geng

Mohammad Masani

Silin Gao

Tim Davidson

Beatriz Borges

Sepideh Mamooler

Marija Šakota

Aoxiang Fan

Martin Josifoski

Terry Zhang

Student assistants (SAs) = MS students

Nearchos Potamitis

Berkay Döner

Riccardo Brioschi

Ali Waseem

Henrique da Silva

Hugo May

Elif Balcioglu

Lucas Burget

Boukil Farouk

Emanuele Nevali

Tommaso Martorella

Federico Di Gennaro

Pau Romeu Llordella

- Help each other on Ed
- Participate actively in classes and labs
- Give us **feedback**

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec1-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What would you like the instructor to (not) wear next time?
- …

Questions?

Commercial break

ADA students: sharp like teeth!

What is data science?

MENU

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Why now?

I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!
I WISH I HAD BECOME A DATA SCIENTIST BACK IN 2023…
KDNUGGETS.COM

“Data science”

- All science is (or should be) based on data, per definitionem
- So how is “data science” different from plain old “science”?

Data volume explodes

"Between the dawn of civilization and 2003, we only created **five exabytes** of information; now [in 2010] we're creating that amount **every two days.**"

Eric Schmidt, Google (2010)

Data variety explodes

Text (indexed Web pages, email), **networks** (Web graph, Google+, knowledge graph), **images, maps, logs** (search logs, server logs, GPS logs), **speech,** …

Needed: A method to the madness

- Scientific method 1.0:

  - Focused on “Model the data”

  - Scientist has hypothesis prior to analyzing the data

- Scientific method 2.0:

  - Systematic cycle (see diagram)

  - “Explore the data” becomes increasingly important

  - **Data as a first-class citizen**

Scientist 2.0

“A data scientist is someone who can obtain, scrub, explore, model, and interpret data, blending hacking, statistics, and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

Hilary Mason, chief scientist at bit.ly

Josh Wills, Data Scientist at Slack

data
is the new oil
we need to find it,
extract it, refine it,
distribute it and
monetize it.
David Buckingham

More data often beats better algorithms

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

21st-century politics

Most Americans say political debate in the U.S. has become less respectful, fact-based, substantive

% who say over the last several years the tone and nature of political debate in this country has become ...

% who say Donald Trump has changed the tone and nature of political debate in the U.S. ...

Note: No answer responses not shown.
Source: Survey of U.S. adults conducted April 29-May 13, 2019.

PEW RESEARCH CENTER

We ask: Do these subjective impressions reflect the true state of US political discourse?

ADA will teach you the tools to answer such questions using data (see next slides)

https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/06/PP_2019.06.19_Political-Discourse_FINAL.pdf

Syllabus, revisited

- **Handling data**
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

Quotebank

Data: https://github.com/epfl-dlab/Quotebank
Web interface: https://quotebank.dlab.tools/

Syllabus, revisited

- Handling data
- **Visualizing data**
- Describing data
- Regression analysis for dise
- Causal analysis of observati
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

Syllabus, revisited

- Handling data
- Visualizing data
- **Describing data**
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

"Is the effect real, or could it have been produced by chance?"

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- **Regression analysis for disentangling data**
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

$$y_t = \boldsymbol{\alpha_0} + \boldsymbol{\beta_0}\, t + \boldsymbol{\alpha}\, i_t + \boldsymbol{\beta}\, i_t\, t + \varepsilon_t$$

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- **Causal analysis of observational data**
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

"What caused the observed increase in negativity?"

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- **Learning from data**
- Handling text data
- Handling network data
- Scaling to massive data

Quotebank

Quotes were attributed to speakers by a machine learning algorithm

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- **Handling text data**
- Handling network data
- Scaling to massive data

Research question ("Did political discourse become more negative?") is a question about language == text

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- **Handling network data**
- Scaling to massive data

In follow-up work we ask: "Who speaks about whom in what way?" → Construct "who-mentions-whom" network

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- **Scaling to massive data**

Quotebank

Extracting 235 million quotes from 127 million news articles (about 1 terabyte of raw text) spanning 12 years requires big-data tools (e.g., Spark)

Curious to learn more?

Full paper available at https://www.nature.com/articles/s41598-023-36839-1

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

United States politicians' tone became more negative with 2016 primary campaigns

Jonathan Külz[1], Andreas Spitz[2], Ahmad Abu-Akel[3], Stephan Günnemann[1] & Robert West[4]

There is a widespread belief that the tone of political debate in the US has become more negative recently, in particular when Donald Trump entered politics. At the same time, there is disagreement as to whether Trump changed or merely continued previous trends. To date, data-driven evidence regarding these questions is scarce, partly due to the difficulty of obtaining a comprehensive, longitudinal record of politicians' utterances. Here we apply psycholinguistic tools to a novel, comprehensive corpus of 24 million quotes from online news attributed to 18,627 US politicians in order to analyze how the tone of US politicians' language as reported in online media evolved between 2008 and 2020. We show that, whereas the frequency of negative emotion words had decreased continuously during Obama's tenure, it suddenly and lastingly increased with the 2016 primary campaigns, by 1.6 pre-campaign standard deviations, or 8% of the pre-campaign mean, in a pattern that emerges across parties. The effect size drops by 40% when omitting Trump's quotes, and by 50% when averaging over speakers rather than quotes, implying that prominent speakers, and Trump in particular, have disproportionately, though not exclusively, contributed to the rise in negative language. This work provides the first large-scale data-driven evidence of a drastic shift toward a more negative political tone following Trump's campaign start as a catalyst. The findings have important implications for the debate about the state of US politics.

A vast majority of Americans—85% in a representative survey by the Pew Research Center[1]—have the impression that "the tone and nature of political debate in the United States has become more negative in recent years". Many see a cause in Donald Trump, who a majority (55%) think "has changed the tone and nature of political debate [...] for the worse", whereas only 24% think he "has changed it for the better"[1]. The purpose of the present article

TODO before Friday's lab session

- Sign up for Ed here and familiarize yourself with it
- If you're not on GitHub yet, sign up for GitHub
- Start looking for 4 teammates
  - You may use "Group formation" category on Ed
- Check out Google Colab and noto (to see if you want to use either of them)
- Check out Exercise 0 here (in prep for Fri lab session)
-

Any feedback? -- Let us know!

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec1-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more details?
- What would you like the instructor to wear next time?
- …

Applied Data Analysis (CS401)

ada

Lecture 2
Handling data
27 Sep 2023

EPFL

Robert West

dlab

Announcements

- **Register** your teams (5 people) here by Fri 6 Oct
  - Each team member must individually complete the form!
- **Project milestone P1** to be released this Fri, due Fri 13 Oct
- **First quiz** ("Q1"*) to be held in this Friday's lab session
  - Test run, to make sure everything works smoothly
  - Won't count towards grade
- **Friday's lab session:**
  - Intro to Pandas (very important for Homework H1 and rest of course)

* Remember: Q*i* is always about material from week *i*.

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec2-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Is it nicer to follow the lecture online or offline?
- …

Cooking with data

Part 2:
Data sources

Part 1:
Data models

Part 3:
Data wrangling

Cooking with data

Part 2:
Data sources

Part 1:
Data models

Part 3:
Data wrangling

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community
Recent cha
Upload file

Tools
What links
Related ch
Special pag
Permanent
Page inform
Cite this pa
Wikidata ite

Print/expor
Download
Printable ve

In other projects
Wikimedia Commons

Languages

Data model

From Wikipedia, the free encyclopedia

A **data model** (or **datamodel**)[1][2][3][4][5] is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities. For instance, a data model may specify that the data element representing a car be composed of a number of other elements which, in turn, represent the color and size of the car and define its owner.

The term **data model** can refer to two distinct but closely related concepts. Sometimes it refers to an abstract formalization of the objects and relationships found in a particular application domain: for example the customers, products, and orders found in a manufacturing organization. At other times it refers to the set of concepts used in defining such formalizations: for example concepts such as entities, attributes, relations, or

Bob's definition: A data model specifies how you think about the world

4 Topics
4.1 Data architecture
4.2 Data modeling
4.3 Data properties

functional specification to aid a computer software make-or-buy decision. The figure is an example of the interaction between process and data models.[6]

Q: “How do you think about the world?”
A: “See my entity–relationship diagram!”

Q: “How do you think about the world?”
A: “See my entity–relationship diagram!”

How to store my data on a computer?

Q1: “How should I store my data on a computer?”
Q2: “How do I think about the world?”

- “The world is simple: one type of entity, all with the same attributes”
  → **Flat model**

```
66.249.65.107 - -  08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 1117  "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
```

- “The world contains many types of entities, connected by relationships”
  → **Relational model**

- “The world is a hierarchy of entities”
  → **Document model**

- “The world is a complex network of entities”
  → **Network model**

Flat model

- Example: log files; e.g., Apache web server (`httpd`)
  - Entities = requests from clients to server

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"

111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1" 200
10801
"http://www.google.com/search?q=in+love+with+ada+lovelace+what+to+do&ie=ut
f-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7)
Gecko/20070914 Firefox/2.0.0.7"
```

- Another common format: CSV (“comma-separated vector”)

Q1: “How should I store my data on a computer?”
Q2: “How do I think about the world?”

- “The world is simple: one type of entity, all with the same attributes”
  → **Flat model**

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
```

- “The world contains many types of entities, connected by relationships”
  → **Relational model**

- “The world is a hierarchy of entities”
  → **Document model**

- “The world is a complex network of entities”
  → **Network model**

Relational model

- "The world contains many types of entities, connected by relationships"
- The relational model is ubiquitous:
  - MySQL, PostgreSQL, Oracle, DB2, SQLite, …
  - You use it many times every day
- Data represented as tables describing
  - entities,
  - relationships between entities

id	name
1	Bush
2	Trump
3	Obama

president	successor
1	3
3	2

Processing data in the relational model: SQL

- *Declarative* language for core data manipulations
- You think about what you want, not how to compute it

Imperative

```
//dogs = [{name: 'Fido', owner_id: 1}, {...}, ... ]
//owners = [{id: 1, name: 'Bob'}, {...}, ...]

var dogsWithOwners = []
var dog, owner

for(var di=0; di < dogs.length; di++) {
  dog = dogs[di]

  for(var oi=0; oi < owners.length; oi++) {
    owner = owners[oi]
    if (owner && dog.owner_id == owner.id) {
      dogsWithOwners.push({
        dog: dog,
        owner: owner
      })
    }
  }}
}
```

Declarative

```
SELECT * from dogs
INNER JOIN owners
WHERE dogs.owner_id = owners.id
```

From: http://latentflip.com/imperative-vs-declarative/

SQL

```
SELECT * from dogs
INNER JOIN owners
WHERE dogs.owner_id = owners.id
```

- You should know basics of SQL
- Need a refresher? → Watch/do online tutorials!
- Key operations:
  - Select (!), update, delete
  - Unique keys
  - Joins (inner, left outer, right outer, full)
  - Sorting
  - Aggregation (group by, count, min, max, avg, etc.)

POLLING TIME

- **“Have you worked with SQL joins?”**
- Scan QR code or go to https://web.speakup.info/room/join/66626

SQL implementations

etc.

```
#!/usr/bin/python

import MySQLdb

# Open database connection
db = MySQLdb.connect("localhost","testuser","test123","TESTDB" )

# prepare a cursor object using cursor() method
cursor = db.cursor()

sql = "SELECT * FROM EMPLOYEE \
       WHERE INCOME > '%d'" % (1000)
try:
   # Execute the SQL command
   cursor.execute(sql)
   # Fetch all the rows in a list of lists.
   results = cursor.fetchall()
   for row in results:
      fname = row[0]
      lname = row[1]
      age = row[2]
      sex = row[3]
      income = row[4]
      # Now print fetched result
      print "fname=%s,lname=%s,age=%d,sex=%s,income=%d" % \
             (fname, lname, age, sex, income )
except:
   print "Error: unable to fecth data"

# disconnect from server
db.close()
```

SQL and “SQL”

- The declarative-programming principles of SQL are widespread, even where it’s less obvious

“SQL”: Pandas (Python library)

- Similar to SQL (declarative), with additional elements of functional programming (map(), filter(), etc.)
- SQL “table” ⟷ Pandas “DataFrame”
- Need intro? Come to Friday’s lab session!

Pandas vs. SQL

+ Pandas is lightweight and fast

+ Natively Python, i.e., full SQL expressiveness plus the expressiveness of Python, especially for function evaluation

+ Integration with plotting functions like Matplotlib

- In Pandas, tables must fit into memory

- No post-load indexing functionality: indices are built when a table is created

- No transactions, journaling, etc. (matters for parallel applications)

- Large, complex joins are slower

“SQL”: Unix command line

Goal: Find top 5 URLs visited most frequently by users between 18 and 35 years old

user_id	age
User145	33
User24	15
User5	76
…	

user_id	url
User2	ada.epfl.ch
User244	facebook.com
…	

```
cat users.txt \
| awk ‘$2 >= 18 && $2 <= 25’ \
| join -1 1 -2 1 url_visits.txt - \
| cut -f 4 \
| sort \
| uniq -c \
| sort -k 1,1 -n -r \
| head -n 5
```

Q1: “How should I store my data on a computer?”
Q2: “How do I think about the world?”

- “The world is simple: one type of entity, all with the same attributes”
  → **Flat model**

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
```

- “The world contains many types of entities, connected by relationships”
  → **Relational model**

- “The world is a hierarchy of entities”
  → **Document model**

- “The world is a complex network of entities”
  → **Network model**

Document model

- "The world is a hierarchy of entities"
- XML format:

```
<contact>
   <id>656</id>
   <firstname>Chuck</firstname>
   <lastname>Smith</lastname>
   <phone>(123) 555-0178</phone>
   <phone>(890) 555-0133</phone>
   <address>
     <street>Rue de l'Ale 8</street>
     <city>Lausanne</city>
     <zip>1007</zip>
     <country>CH</country>
   </address>
</contact>
```

- JSON format:

```
contact: {
   id: 656,
   firstname: "Chuck",
   lastname: "Smith",
   phones: ["(123) 555-0178",
            "(890) 555-0133"],
   address: {
     street: "Rue de l'Ale 8",
     city: "Lausanne",
     zip: 1007,
     country: "CH"
   }
}
```

- Document model

```
<contact>
  <id>656</id>
  <firstname>Chuck</firstname>
  <lastname>Smith</lastname>
  <phone>(123) 555-0178</phone>
  <phone>(890) 555-0133</phone>
  <address>
    <street>Rue de l'Ale 8</street>
    <city>Lausanne</city>
    <zip>1007</zip>
    <country>CH</country>
  </address>
</contact>
```

Think for a minute:

If we want to use a relational DB (e.g., MySQL) instead of XML, how can we store several phone numbers for the same person?

(Feel free to discuss with your neighbor.)

Solution to “Think for a minute”

- Document model

```
<contact>
  <id>656</id>
  <firstname>Chuck</firstname>
  <lastname>Smith</lastname>
  <phone>(123) 555-0178</phone>
  <phone>(890) 555-0133</phone>
  <address>
    <street>Rue de l’Ale 8</street>
    <city>Lausanne</city>
    <zip>1007</zip>
    <country>CH</country>
  </address>
</contact>
```

- Same in relational model

id	first name	...
656	Chuck	...
...	...	...

id	phone
656	(123) 555-0178
656	(890) 555-0133
...	...

Processing XML and JSON

- Document structure = tree
- Processing via tree traversal (depth- or breadth-first search)
- Or use proper query language, such as XQuery or jq

Q1: “How should I store my data on a computer?”
Q2: “How do I think about the world?”

- “The world is simple: one type of entity, all with the same attributes”
  → **Flat model**

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
```

- “The world contains many types of entities, connected by relationships”
  → **Relational model**

- “The world is a hierarchy of entities”
  → **Document model**

- “The world is a complex network of entities”
  → **Network model**

Network model

- “The world is a complex network of entities”

“How should I store my data on a computer?” —A word (or two) on binary formats

- “Parsing” = converting strings (as stored in text files) to data types used by computer programs (e.g., int, float, boolean, array, list)
- Possibly expensive, but can be avoided by using binary formats: store data to disk “as is”, without first converting to strings
- Modern binary formats support nested structures, various levels of schema enforcement, compression, etc.
- Python pickle, Java Serializable, Protocol Buffers (Google), Avro (supports schema evolution), Parquet (column-oriented), etc.

→ Consider converting to a binary format at the beginning of your processing pipeline (especially when using “big data”)

Commercial break

ADA: more than just rocket science!

Cooking with data

Part 2:
Data sources

Part 1:
Data models

Part 3:
Data wrangling

Data sources at Web companies

Examples from Facebook

- Application databases
- Web server logs
- Client-side event logs
- API server logs
- Ad server logs
- Search server logs
- Advertisement landing page content
- Wikipedia
- Images and video

Structured data (with clear schema)

Semi-structured data (“self-describing” structure; CSV etc.)

Unstructured data

Another example: Wikipedia

- 300+ languages
- 42 million entities
- Mind-boggling richness of data

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikiquote
Wikivoyage

Languages
Afrikaans
Alemannisch
ᎣᏍᏏᏲ
العربية
Aragonés
Asturianu
Azərbaycanca
تۆرکجه
বাংলা

San Francisco

From Wikipedia, the free encyclopedia
(Redirected from San Francisco, California)

Coordinates: 37°47′N 122°25′W

link

This article is about the city and county in California. For other uses, see San Francisco (disambiguation).

San Francisco (initials **SF**[17]) (/ˌsæn frənˈsɪskoʊ/, Spanish for Saint Francis; Spanish: [san fraˈsisko]), officially the **City and County of San Francisco**, is the cultural, commercial, and financial center of Northern California. The consolidated city-county covers an area of about 47.9 square miles (124 km^2)[18] at the north end of the San Francisco Peninsula in the San Francisco Bay Area. It is the fourth-most populous city in California, and the 13th-most populous in the United States, with a 2016 census-estimated population of 870,887.[13] The population is projected to reach 1 million by 2033.[19]

San Francisco was founded on June 29, 1776, when colonists from Spain established Presidio of San Francisco at the Golden Gate and Mission San Francisco de Asís a few miles away, all named for St. Francis of Assisi.[1] The California Gold Rush of 1849 brought rapid growth, making it the largest city on the West Coast at the time. San Francisco became a consolidated city-county in 1856.[20] After three-quarters of the city was destroyed by the 1906 earthquake and fire,[21] San Francisco was quickly rebuilt, hosting the Panama-Pacific International Exposition nine years later. In World War II, San Francisco was a major port of embarkation for service members shipping out to the Pacific Theater.[22] It then became the birthplace of the United Nations in 1945.[23][24][25] After the war, the confluence of returning servicemen, massive immigration, liberalizing attitudes, along with the rise of the "hippie" counterculture, the Sexual Revolution, the Peace Movement growing from opposition to United States involvement in the Vietnam War, and other factors led to the Summer of Love and the gay rights movement, cementing San Francisco as a center of liberal activism in the United States. Politically, the city votes strongly along liberal Democratic Party lines.

A popular tourist destination,[26] San Francisco is known for its cool summers, fog, steep rolling hills, eclectic mix of architecture, and landmarks, including the Golden Gate Bridge, cable cars, the former Alcatraz Federal Penitentiary, Fisherman's Wharf, and its Chinatown district. San Francisco is also the headquarters of five major banking institutions and various other companies such as Levi Strauss & Co., Gap Inc., Fitbit, Salesforce.com, Dropbox, Reddit, Square, Inc., Dolby, Airbnb, Weebly, Pacific Gas and Electric Company, Yelp, Pinterest, Twitter, Uber, Lyft, Mozilla, Wikimedia Fo
home to number of educational and cultural institutions, such as the Univer
Museum, the San Francisco Museum of Modern Art, and the California Aca

San Francisco has several nicknames, including "The City by the Bay", "Go
City", and as well as older ones like "The City that Knows How", "Baghdad
City".[17] As of 2017, San Francisco is ranked high on world liveability ranki

San Francisco, California
Consolidated city-county
City and County of San Francisco

San Francisco and the Golden Gate Bridge from Marin Headlands

Flag Seal

Learning resources from Wikiversity

Places adjacent to San Francisco	[show]
V·T·E City and County of San Francisco	[show]
Articles relating to the City and County of San Francisco	[show]

Authority control: WorldCat Identities · VIAF: 143700861 · LCCN: n79018452 · ISNI: 0000 0004 0461 8991 · GND: 4051520-5 · SUDOC: 040776433 · NDL: 00628542

Categories: San Francisco | 1850 establishments in California | California counties | Cities in the San Francisco Bay Area | Consolidated city-counties in the United States | Counties in the San Francisco Bay Area | County seats in California | Hudson's Bay Company trading posts | Incorporated cities and towns in California | Populated coastal places in California | Populated places established in 1776 | Port cities and towns of the West Coast of the United States | Spanish mission settlements in North America

Contents [hide]

1 History
2 Geography
2.1 Cityscape
2.1.1 Neighborhoods
2.2 Climate
3 Demographics
3.1 Race, ethnicity, religion and languages
3.2 Education, households, and income
3.2.1 Homelessness
4 Economy

Location of San Francisco in California
Coordinates: 37°47′N 122°25′W

Country	United States
State	California

Wikipedia

How to work with Wikipedia?

- REST (cf. later) API
- XML dumps with wiki markup, SQL database dumps
- Issues: Unicode, size, recency, etc.
- To make your life easier:
  (1) Find projects on GitHub to help you
  (2) Use more structured versions (p.t.o.)

Wikidata

- “Database version” of Wikipedia
- {fr:Suisse, de:Schweiz, it:Svizzera, en:Switzerland, …} → Q39

English Not logged in

WIKIDATA

Item Discussion Read View history Search Wikidata

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Switzerland (Q39)

edit

federal republic in Western Europe

Swiss Confederation | CH | SUI | Suisse | Schweiz | Svizzera |

In more languages Configure

Language	Label	Description	Also known as
English	Switzerland	federal republic in Western Europe	Swiss Confederation CH SUI Suisse Schweiz Svizzera
German	Schweiz	Staat in Mitteleuropa	Schweizerische Eidgenossensc… Eidgenossenschaft CH SUI
Swiss German	Schwyz	No description defined	
French	Suisse	pays d'Europe	Confédération helvétique Confédération suisse CH SUI

All entered languages

Statements

instance of	sovereign state 1 reference	edit
	country start time 12 September 1848 *Gregorian*	edit

Wikidata

- “Database version” of Wikipedia
- {fr:Suisse, de:Schweiz, it:Svizzera, en:Switzerland, …} → Q39
- Both API access and full database dumps
- Available as
  - JSON (document model)
  - RDF (network model)

Wikipedia pageview logs

Crawling and processing webpages: HTML

Plenty of bulk-downloadable HTML data:

- Common Crawl dataset, about 1.82 billion web pages -- huge!
- (… but less than 0.1% of Google's Web crawl, as of 2015)
- 145 TB, hosted on Amazon S3, also available for download

… but if you need a specific website: use a **crawler/"spider"**: Apache Nutch, Storm, Heritrix 3, Scrapy, etc. (or simply wget…)

Useful HTML tools

Requests http://docs.python-requests.org/en/master/
An elegant and simple HTTP library for Python

Scrapy https://scrapy.org/
An open-source framework to build Web crawlers

Beautiful Soup http://www.crummy.com/software/BeautifulSoup/
A Python API for handling real HTML

Plain ol' /regular/expres*ion/s…

Schema.org: microformats for Web pages

- Nuggets of structured information embedded in (semantically) unstructured HTML

Text as rendered by browser:

Avatar
Director: James Cameron (born August 16, 1954)
...

HTML under the hood:

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span>
(born <time itemprop="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

Web services

- Most large web sites today actively discourage “screen-scraping” to get their content
- Instead: Web service APIs, for interoperable machine-to-machine interaction over a network
- The preferred way to get data from online sources
- Most common framework: REST
  - You request a URL from the server via HTTP
  - The server responds with a text file (e.g., JSON, XML, plain text)

REST example

```
{
  "user": {
    "name": "Jane",
    "gender": "female",
    "location": {
      "href":
"http://www.example.org/us
/ny/new_york",
      "text": "New York"
    }
  }
}
```

← This resource is a description of a user named Jane

- Requested by sending GET request for the resource's URL, e.g., via curl:
  `curl http://www.example.org/users/jane/`
- If users need to modify the resource, they GET it, modify it, and PUT it back
- The href to the location resource allows savvy clients to get more information with another simple GET request
- Implication: Clients cannot be too "thin"; need to understand resource formats!

Joe Zawinul

I said, “What’s that?” and he said, “That’s **jazz**.” “How do you write that?” And he spelt it out. Somehow I saw my name in there, and I liked this word.

Robert West

I said, “What’s that?” and he said, “That’s **REST**.” “How do you write that?” And he spelt it out. Somehow I saw my name in there, and I liked this word.

[link]

Cooking with data

Part 2:
Data sources

Part 1:
Data models

Part 3:
Data wrangling

Working with raw data sucks

Data comes in all shapes and sizes

– CSV files, PDFs, SQL dumps, .jpg, …

Different files have different formatting

– Empty string or space instead of NULL, extra header rows, character encoding, …

“Dirty” data

– Unwanted anomalies, duplicates

Raw data without thinking: A recipe for disaster!

What is data wrangling?

- **Goal:** extract and standardize the raw data
  - Combine multiple data sources
  - Clean data anomalies
- **Strategy:** Combine automation with interactive visualizations to aid in cleaning

Wrangling takes **between 50% and 80% of your time…**

[Source]

Types of data problems

- Missing data
- Incorrect data
- Inconsistent representations of the same data
- About 75% of data problems require human intervention (e.g., experts, crowdsourcing, etc.)
- Tradeoff between cleaning data vs. over-sanitizing data

link

“Dirty data” horror stories

“Dear Idiot” letter

17,000 men are pregnant

As the crow flies

CHF 10,000 compute-cluster bill

[Source]

Diagnosing data problems

- Visualizations and basic stats can convey issues in “raw” data
- Different representations highlight different types of issues:
  - Outliers often stand out in the right kind of plot
  - Missing data will cause gaps or zero values in the right kind of plot
- Becomes increasingly difficult as data gets larger
  - Sampling to the rescue!

Facebook graph

Matrix view (1)

Automatic permutation of rows and columns to highlight patterns of connectivity

Matrix view (2)

Rows and columns sorted in the order in which data was retrieved via the Facebook API

Can you guess what's going on here?

Viz at scale? Careful!

Dealing with missing data

U.S. census counts of people working as ‘‘farm laborers’’; values from 1890 are **missing due to records being burned in a fire**

- Set values to zero?
- Interpolate based on existing data?
- Omit missing data?

Knowledge about domain and data collection should drive your choice!

Inconsistent data: “My name is Willy”

First name	Last name
Willy	NULL
...	...

Experiments on Pattern-based Rel

Willy Yap and Timothy Baldwin
NICTA Victoria Research Laboratory
Department of Computer Science and Software E
University of Melbourne
willy@csse.unimelb.edu.au, tim@csse.unir

Before you start analyzing your data

- “Do I have **missing data**?” “If data were missing, how could I know?”
- “Do I have **corrupted data**?” (May arise from measurement errors, processing bugs, etc.)
- **Parse/transform data** into appropriate format for your specific analysis (see “Part 1: Data models”)
- Don’t be surprised if you need to come back to this stage!

Desiderata

It's always ideal if you can put your hands on the **code/documentation about the dataset** you are analyzing (provenance)

It's always ideal if the provided **data format is nicely parseable** (otherwise you need regexes, or maybe even pay humans)

Highly non-parseable data

Entire NY Times archive (since 1851) digitized as of 2015

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec2-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Is it nicer to follow the lecture online or offline?
- …

Example from Bob's research (AD 2013)

Q: What do

Our method: I

Consenting IE users, 18 mo

URL	Referrer	Tin
yahoo.com?q=the+onion	yahoo.com	128
theonion.com	yahoo.com?q=the+onion	128
theonion.com/Area-Man-Sad		
bing.com		
bing.com?q=feijoada+recipe		
allrecipes.com/tasty-feijoada		
food.com/best-feijoada-recipe		

Ingredients Edit and Save

Original recipe makes 8 servings Cha

- 1 (12 ounce) package dry black beans, soaked overnight
- 1 1/2 cups chopped onion, divided
- 1/2 cup green onions, chopped
- 1 clove garlic, chopped
- 2 smoked ham hocks
- 8 ounces diced ham
- 1/2 pound thickly sliced bacon, diced

(optional)

1/4 cup chopped (optional)

Sodium 299 mg 12%

* Percent Daily Values are based on a 2,000 calorie diet.

See More

powered by esha RESEARCH

12

- Find Amazon add-to-cart events heuristically in logs:

  Referrer: http://www.amazon.com/Forks-Over-Knives-Plant-Based-Health/dp/1615190457

  URL: http://www.amazon.com/gp/cart/view-upsell.html?...
- Get product info for product id using Amazon API
- Consider all add-to-cart events for category "Diets & Weight Loss"

- Collaborated with clinician at Washington Hospital Center, Washington, D.C.
- Data: All CHF admissions to emergency department for time period of our browsing logs

Paper with results and plots

Applied Data Analysis (CS401)

ada

Lecture 3
Visualizing data
4 Oct 2023

EPFL

Robert West

dlab

Announcements

- You must find 4 team mates and register by this Fri 6 Oct 23:59
  - Every student must register individually
  - Still looking for a team or for team members? Use Ed! If need be, talk to Sepideh [TA]
- Project milestone 1 was released last Fri; due Fri 13 Oct 23:59
- Homework 1 released Fri 13 Oct (due two weeks later)
- Friday's lab session:
  - 13:15–13:30 (on Moodle): Quiz 2 (the first one that counts)
  - 13:30–14:30: Project milestone P1 office hours (Zoom)
  - In parallel: Exercise on data visualization and working with data from the Web

Announcements

Want to contribute to a (fun!) research study?

https://go.epfl.ch/wikispeedia_LLM_ada_doc

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec3-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What's your favorite color, baby?
- …

Uses for data visualization

Support reasoning about information **(analysis)**

- Finding relationships
- Discover structure
- Quantifying values and influences
- Should be part of the data analysis cycle ➡

Inform and persuade others **(communication)**

- Capture attention, engage
- Tell a story visually
- Can focus on certain aspects and omit others

Make it easier to evaluate potential courses of action **(decision-making)**

An unconventional example

"Garden of Eden": 8 lettuces, each of which is enclosed in its own airtight plexiglas box and represents a major city. The concentration of ozone in each box is controlled in real-time to reflect the current pollution level in the city.

Static viz

Great for data exploration, developed throughout the last few centuries…

Interactive viz

More and more common when delivering the results (and also during exploration). New frameworks are the key enabler.

Want to learn more?

Dedicated course:

COM-480: Data Visualization

Today’s lecture

- Part 1: Navigating the chart landscape
- Part 2: Principles and best practices
- Part 3: A (small) selection of use cases for data visualization

Part 1
Navigating the chart landscape

Chart selection

One variable: histograms

Histograms can tell you a lot about a single variable, discrete or continuous

Easy to recognize skewed distributions!

Smoothed histogram (a.k.a. kernel density estimate)

One variable: box plots

Two variables: scatter plots

Scatter plots quickly expose the relationships between two variables

Two variables: line plots

If relationship is functional (for instance, after binning and aggregating)

> 2 variables: scatter plot matrix

> 2 variables: stacked plots

Here: 3 variables: stack index, height, color

Stack variable and color variables categorical,
height variable continuous:

Color variable categorical,
stack and height variables continuous:

Dimensionality reduction

- For example, **PCA:** allows visualization of high-dimensional continuous data in 2D using principal components
- The principal components are the strongest (highest variation) dimensions in the dataset, and are orthogonal

One dataset, visualized 25 ways

http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways

"You must help the data focus and get to the point. Otherwise, it just ends up rambling about what it had for breakfast this morning and how the coffee wasn't hot enough."

Part 2
Principles and best practices

Instructive coffee table books by Edward

[another great coffee table book]

Perception of magnitudes

(134, 134, 134)

(144, 144, 144)

Which is brighter?

Just noticeable difference (JND)

- Weber's law:

$$\frac{\Delta I}{I} = k,$$

- I: intensity; ΔI: increase from I to notice a difference; k: constant
- Required increase ΔI depends on original intensity I
- Most continuous variations in stimuli are perceived in discrete (multiplicative) steps

Compare area of circles

Compare area of circles

Perception of magnitudes

Most accurate

Least accurate

Position

Length

Slope

Angle

Area

Volume

Color hue-saturation-density

Cleveland & McGill (1984) *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*

Choose your axes wisely!

Time series of pageviews of Wikipedia article about "Coronavirus"

(linear y-axis)

(logarithmic y-axis)

[link]

Commercial break

Choose your axes wisely: Visualizing heavy-tailed distributions

Linear x-axis

Logarithmic x-axis

Heavy-tailed data: power laws

- Probability of x: $p(x) = Cx^{-\alpha}$,
  - Very large values are rare, “but not very rare”
  - Body size vs. city size
  - Many natural phenomena are power laws (e.g., # of friends)
  - For dealing with them, need to know some tricks
  - E.g., for small α, mean & var = ∞ → use median!
  - E.g., straight line on log-log axes:
    - $y = C\,x^{-\alpha} \leftrightarrow \underline{\log(y)} = \underline{\log(C)} - \alpha\,\underline{\log(x)}$

Heavy-tailed data: power laws

- Complementary cumulative distribution function (CCDF):
  $P(x) := \Pr\{X >= x\}$

POLLING TIME

- "What shape does the CCDF of a power law have when plotted on log-log axes?"
- Scan QR code or go to https://web.speakup.info/room/join/66626

Heavy-tailed data: power laws

- Complementary cumulative distribution function (CCDF):
  P(x) := Pr{X >= x}
- CCDF of power law is also a power law (with exponent α – 1)
  $$P(x) = C \int_x^\infty x'^{-\alpha}\,\mathrm{d}x' = \frac{C}{\alpha - 1} x^{-(\alpha-1)}.$$
- CCDF plot is monotonically decreasing (even without binning)

Answer fast: which time series has a higher mean value?

Use consistent axes!

Answer fast: which time series has a higher mean value?

Label your axes!

THINK FOR A MINUTE:

How could we show details of both time series without using different y-axes?

(Feel free to discuss with your neighbor.)

Guinness
Paulaner
Pints consumed
50.00
20.00
5.00
2.00
0.50
0.20
0.05
Jan 2015
Jan 2016
Jan 2017
Jan 2018
Jan 2019
Day

Show data uncertainty!

Which beer is more popular, Guinness or Paulaner?

Consider using small multiples!

Use colors consistently!

[link]

Use colors wisely!

Choose colors based on the information you want to convey

Sequential

Colors can be ordered from low to high

Diverging

Two sequential schemes extended out from a critical midpoint value

Categorical

Lots of contrast between each adjacent color

Number of data classes: 3
how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
photocopy safe
3-class BuGn
HEX
EXPORT
#e5f5f9
#99d8c9
#2ca25f
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency

Use colorblind-safe palettes!

- Remember: 10% of males have some form of colorblindness

	Original	Simulation: Protan	Simulation: Deutan	Simulation: Tritan		Hue	for Photoshop, Illustrator, Freehand, etc. C,M,Y,K (%)	for Photoshop, Illustrator, Freehand, etc. R,G,B (0-255)	for Word, Power Point, Canvas, etc. R,G,B (%)
1					Black	– °	(0,0,0,100)	(0,0,0)	(0,0,0)
2					Orange	41°	(0,50,100,0)	(230,159,0)	(90,60,0)
3					Sky Blue	202°	(80,0,0,0)	(86,180,233)	(35,70,90)
4					bluish Green	164°	(97,0,75,0)	(0,158,115)	(0,60,50)
5					Yellow	56°	(10,5,90,0)	(240,228,66)	(95,90,25)
6					Blue	202°	(100,50,0,0)	(0,114,178)	(0,45,70)
7					Vermilion	27°	(0,80,100,0)	(213,94,0)	(80,40,0)
8					reddish Purple	326°	(10,70,0,0)	(204,121,167)	(80,60,70)

Use data ink wisely! Avoid chart junk!

Graphical excellence gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
-Edward Tufte

Toilet exercise: Which principles and best practices do these graphics violate?

Courtesy of viz.wtf

Part 3
A (small) selection of use cases for data visualization

Use case: Presenting scientific results

Use case:
Data wrangling

Multimodal data

- Two or more distinct peaks in a histogram often **suggest 2 or more distinct populations** of samples.
- But don't guess! Explore further by using, e.g., color and a histogram of multiple populations (p.t.o.).

Use case:
Data wrangling

Multimodal data

Explore further by using, e.g., color and a histogram of multiple populations

Use case:
Data wrangling

Weird data

- Maintain a **theory of what the data should look like**.
- Some data is **very hard to explain**.
- Never just blink it away!
- First, assume a bug. Try to fix it.
- If not a bug: you might have made an interesting discovery!
- Some of science's most important findings were made by not ignoring weird data, but dwelling on it!

[link]

Use case: Journalism

NY Times interactive visualizations (recession/recovery 2014)

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

And 2014 “the year in interactive storytelling”

http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html?_r=0

NY Times graphics are a great source of best practices in viz (except for when they’re not…)

Use case: Educating the public

Hans Rosling:

200 countries, 200 years, 4 minutes

https://www.youtube.com/watch?v=jbkSRLYSojo

Use case: Give new perspectives

Charles Joseph Minard 1869 Napoleon's march

According to Tufte: "*It may well be the best statistical graphic ever drawn.*"

5 variables: army size, location, dates, direction, temperature during retreat

Tools

(remaining slides for your personal perusal)

Interactive toolkits: D3

Without doubt, the most widely used interactive visualization framework is **D3**.

Note from the authors: *D3 is intentionally a low-level system. During the early design of D3, we even referred to it as a "visualization kernel" rather than a "toolkit" or "framework"*

Interactive toolkits: Vega

Vega is a “visualization grammar” developed on top of D3.js

It specifies graphics in JSON format.

vega

Vega is a *visualization grammar*, a declarative format for creating, saving, and sharing interactive visualization designs.

Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?

Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?

Bokeh: another interactive viz library

Bokeh is an independent Viz library focused more heavily on big data visualization. Has both Python and Scala bindings.

Visualizing maps: Folium

More in tomorrow's lab session!

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec3-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What's your favorite color, baby?
- …

> 2 variables: parallel-coord. plots

Color, x, y

Color variable is categorical, others arbitrary

> 2 variables: radar charts

- Similar to parallel-coord. plots
- Doesn't pretend that x axis has meaningful order
- Also good for periodic data

Heavy-tailed data: power laws

- Smart trick for plotting CCDF of any distribution:
  - x-axis: data sorted in ascending order
  - y-axis: (n:1)/n (where n is number of data points)

Interactive chart design: simplifying

- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.

Use structure!

Gestalt psychology principles (1912)

A case for ugly visualizations

People instinctively gravitate to attractive visualizations, and they have a better chance of getting on the cover of a journal.

But does this **conflict with the goals of visualization?**

- Rapid exploration
- Focus on most important details
- Easy and fast to develop and customize

Guide your audience!

Applied Data Analysis (CS401)

ada

Lecture 4
Describing data
11 Oct 2023

EPFL

Robert West

dlab

Announcements

- Project milestone P1 due this Fri 13 Oct 23:59
  - Remember: we won’t answer questions in final 24h
- Homework H1 to be released Fri 13 Oct, due Fri 27 Oct
- Friday’s lab session:
  - From now on: one single room: BCH2201
  - Exercises on topic of this lecture (describing data)
  - Quiz 3
- Next week’s Wed lecture: held on Zoom due to travel
  - You can watch a live stream in Rolex Learning Center or watch from home

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec4-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo? / Où est Charlie?
- …

Overview of today's lecture

- Part 1: Descriptive statistics
- Part 2: Quantifying uncertainty
- Part 3: Relating two variables

ADA won't cover the basics of stats!

You know these things from prerequisite courses

But stats is a key ingredient of data analysis

Today: some highlights and common pitfalls

Part 1
Descriptive statistics

Descriptive statistics

```
baseball.describe()
```

	year	stint	g	ab	r
count	100.00000	100.000000	100.000000	100.000000	100.00000
mean	2006.92000	1.130000	52.380000	136.540000	18.69000
std	0.27266	0.337998	48.031299	181.936853	27.77496
min	2006.00000	1.000000	1.000000	0.000000	0.00000
25%	2007.00000	1.000000	9.500000	2.000000	0.00000
50%	2007.00000	1.000000	33.000000	40.500000	2.00000
75%	2007.00000	1.000000	83.250000	243.750000	33.25000
max	2007.00000	2.000000	155.000000	586.000000	107.00000

Means: micro- vs. macro-average

Robust statistics

x

A statistic is said to be robust if it is not sensitive to **extreme values**

```
baseball.describe()
```

	year	stint	g	ab	r
count	100.00000	100.000000	100.000000	100.000000	100.00000
mean	2006.92000	1.130000	52.380000	136.540000	18.69000
std	0.27266	0.337998	48.031299	181.936853	27.77496
min	2006.00000	1.000000	1.000000	0.000000	0.00000
25%	2007.00000	1.000000	9.500000	2.000000	0.00000
50%	2007.00000	1.000000	33.000000	40.500000	2.00000
75%	2007.00000	1.000000	83.250000	243.750000	33.25000
max	2007.00000	2.000000	155.000000	586.000000	107.00000

Min, max, mean, std are not robust

Median, quartiles (and others) are robust

Check these Wikipedia pages

Heavy-tailed distributions

- Some distributions are all about the extreme values
- E.g., power laws (see last lecture): $f(x) = ax^{-k}$
  - Very very large values are rare, “but not very rare”
  - Body size vs. city size
  - For k <= 3: infinite variance
  - For k <= 2: infinite variance, infinite mean
  - Don’t report (arithmetic) mean/variance for power-law-distributed data!
  - Use robust statistics (e.g., median, quantiles, etc.) or geometric mean (p.t.o.)

Generalized means [Wikipedia]

- Common trick: transform data into a different space (via function f), take mean there, then transform back into the original space (via f^{-1}):

$$f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$$

- $f(x) = x^2$, $f^{-1}(x) = \sqrt{x}$ "root mean square"
- $f(x) = x$, $f^{-1}(x) = x$ "arithmetic mean"
- $f(x) = \log(x)$, $f^{-1}(x) = \exp(x)$ "geometric mean"
- $f(x) = 1/x$, $f^{-1}(x) = 1/x$ "harmonic mean"

Distributions

link

Distributions

Some important distributions:

- **Normal:** see previous slides
- **Poisson:** the distribution of counts that occur at a certain “rate”; e.g., number of visits to a given website in a fixed time interval.
- **Exponential:** the interval between two such events.
- **Binomial/multinomial:** The number of “successes” (e.g., coin flips = heads) out of *n* trials.
- **Power-law/Zipf/Pareto/Yule:** e.g., frequencies of different terms in a document; city size

You should understand the distribution of your data before applying any model!

“Dear data, where are you from?”

- Visual inspection for ruling out certain distributions:
  e.g., when histogram/box plot is asymmetric (even for large sample size), the data cannot be normal

Box plot

(Smoothed) histogram

Quantile-quantile (QQ) plots

- Statistical tests:
  - Goodness-of-fit tests
  - Kolmogorov-Smirnoff test
  - Normality tests

Recognizing a power law

Log-log axes

Who likes Snickers better?

Part 2
Quantifying uncertainty

Who likes Snickers better?

- Most straightforward descriptive statistic to answer this question:
- Mean for each group (women, men)

Who likes Snickers better?

Who likes Snickers better?

Be certain to quantify your uncertainty!

- Finite samples introduce uncertainty
  - Even a complete dataset is a finite sample!
- Whenever you report a statistic, you need to quantify how certain you are in it!
- We will discuss two ways of quantifying uncertainty:
  (1) Hypothesis testing
  (2) Confidence intervals
- All plots should have error bars!

How to quantify uncertainty?
Approach 1:

Hypothesis testing

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

THINK FOR A MINUTE:

Which of these statements about p-values are true?

(Feel free to discuss with your neighbor.)

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

POLLING TIME

- **"Which of these statements about p-values are true?"**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Hypothesis testing: intro

Joseph B. Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP ("extrasensory perception"), i.e., they could guess the color of all 10 cards!

Q: Do you agree?

Hypothesis testing: intro

He called back the “psychic” subjects and had them do the same test again. They all failed.

He concluded that **the act of telling psychics that they have psychic abilities** causes them to lose them…

Hypothesis testing

- A huge subject; can take entire classes on it
- Many people don't like it
  - cf. Bayesian vs. frequentist debate (a.k.a. war)
- Need to understand basics even if you don't use it yourself
- Never use it without understanding exactly what you're doing

Commercial break

The logic of hypothesis testing

- Flip a coin 100 times; outcome: 40 heads; “Is the coin fair?”
- Null hypothesis: “yes”; alternative hypothesis: “no”
- “How likely would I be to see an outcome at least this extreme (i.e., ≤ 40 heads) if the null hypothesis were true (i.e., if the coin were fair, i.e., if we expect 50 heads)?”
- If this probability is large, the null hypothesis suffices to explain the data (and is thus not rejected)
- Otherwise, dig deeper in order to understand your data

The logic of hypothesis testing

- Idea: Gain (weak and indirect) support for a hypothesis H_A by **ruling out a null hypothesis H_0**

- by inspecting a **test-statistic:** a measurement made on the data that is likely to be large under H_A but small under H_0

Coin example

- Idea: Gain (weak and indirect) support for a hypothesis H_A by **ruling out a null hypothesis H_0**
  - H_0: “the coin is fair” (simplest hypothesis, cf. Occam’s razor)
  - H_A: “the coin is not fair (a.k.a. biased)”

- by inspecting a **test-statistic:** a measurement made on the data that is likely to be large under H_A but small under H_0
  - e.g., number of heads after 100 coin tosses (1-tailed)
  - e.g., abs(50 - number of heads after 100 coin tosses) (2-tailed)

Coin example (cont'd)

- **Null hypothesis H_0:** "the coin is fair", i.e., "probability of heads = 0.5"
- **Test statistic s:** abs(50 - # of empirically observed heads after 100 coin tosses)
- $Pr(S \mid H_0)$: probability distribution of test statistic, assuming that H_0 is true

- Decision rule: **reject H_0 if $Pr(S \geq s \mid H_0) < \alpha$,**
  i.e., if the probability of deviating from 50 heads at least as much as empirically observed is small
  - $Pr(S \geq s \mid H_0)$ = **"p-value"**
  - α = **"significance level"**
- α controls "false-rejection rate" (probability of rejecting H_0 although it is true)
  - You as the data analyst choose α (common values: 5%, 1%, 0.5%, 0.1%)
  - Higher $\alpha \rightarrow$ higher false-rejection rate

Selecting the right test

There are many statistical tests (see next slide)
Although they differ in their details, the basic logic is always the same (previous slides)

The right choice of test depends on multiple factors (here a selection):

- Question
- Data type (continuous vs. categorical; dimensionality; number of outcomes)
- Sample size
- When comparing two samples: same population or different populations?
- Parametric assumptions about distribution of test statistic under null hypothesis? (Less important for large sample sizes, due to central limit theorem)

Good news: **Plenty of advice available (p.t.o.)**

How Many Outcome Variables
What Type of Outcome
How Many Predictor Variables
What Type of Predictor?
If catecorial Predictor, How Many Categories?
If a Categorial Predictor, Same Participants or Different Used in Each Category?
Do Data Meet Assumptions for Parametric Tests?
One
Two Or More
Continuous
Categorical
Continuous
One
Two Or More
One
Two Or More
One
Two Or More
Categorical
Continuous
Categorical
Continuous
Both
Categorical
Continuous
Categorical
Continuous
Both
Categorical
Categorical
Both
Two
More Than Two
Different
Same
Different
Same
Different
Same
Both
Different
Different
Different
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Indepent t-test / Point-Biserial Correlation
Mann-Whitney Test
Dependent t-test
Wilcoxon Matched-Pairs Test
One Way Independent ANOVA
Kruskal-Wallis Test
One Way Repeated Measures ANOVA
Friedman's ANOVA
Pearson Correlation or Regression
Spearman Correlation or Kendall's Tau
Independent Factorial ANOVA / Multiple Regression
Factorial Repeated Measures ANOVA
Factorial Mixed ANOVA
Multiple Regression / Partial Correlation
Multiple Regression/ ANCOVA
Pearson Chi-Square or Likelihood Ratio
Logistic Regression or Biserial / Point-Biserial Correlation
Loglinear Analysis
Logistic Regression
MANOVA
Factorial MANOVA
MANCOVA

Remarks on p-values

- Widely used in all sciences
- They are widely misunderstood!
- Don’t use them if you don’t understand them!
- Large p means that even under a simple null hypothesis your data would be quite likely
- This tells you nothing about the alternative hypothesis

Remarks on p-values

- Historically, not meant as a method for formally deciding whether a hypothesis is true or not
- Rather, an informal tool for assessing a particular result
- Low p-value means: “The simple null hypothesis doesn’t explain the data, so keep looking for other explanations!”
- p = 0.05 means: if you repeat experiment 20 times, you’ll see extreme data even under null hypothesis → you might have “lucked out”
- Look at the effect size (“y-axis”), not just the p-value!

Remarks on p-values

- Important to understand what p-values are
- Maybe even more important to understand what they are not…
- Read this paper: [A Dirty Dozen: 12 P-Value Misconceptions]

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The

Alternative approach: Bayes factors

$$\frac{\text{Prob(Data, under } H_0)}{\text{Prob(Data, under } H_A)}$$

- See here
- Great (and amusing) explanation of difference between hypothesis-testing approach and Bayesian approach: Chapter 37 in MacKay's (free) book on "Information Theory, Inference, and Learning Algorithms"

How to quantify uncertainty?
Approach 2:

Confidence intervals

Confidence intervals: idea

- Confidence interval (CI)
  = a range of estimates for the parameter of interest (e.g., mean) that seems reasonable given the observed data
- Confidence level γ ⇒ "γ CI"
  (often γ = 95% ⇒ 95% CI)

Confidence intervals: definition

- μ: true value of parameter of interest
- m: empirical estimate of parameter of interest
- CIs and hypothesis testing are tightly connected:
  - γ CI contains those values μ_0 for which the null hypothesis "H_0: $\mu = \mu_0$" cannot be rejected at significance level 1−γ

How to compute confidence intervals?

- **Parametric** methods assume that the test statistic follows a known (typically Normal) distribution
  $\longrightarrow$ Need to verify that this is actually true! Ugh…

- **Non-parametric** methods make no assumptions about the distribution of the test statistic. They instead work by sampling the empirical data.
  $\longrightarrow$ Yay!

Confidence intervals: another view

- If we were to repeat the data collection $N \rightarrow \infty$ independent times, we'd obtain N estimates of μ: $m_1, \ldots, m_N$
- Average of m_i's will approach the true μ (by law of large numbers)
- For a fraction γ of the N repetitions, m_i lies within the γ CI around μ
- $\rightarrow$ May estimate CI from histogram of m_i's

Non-parametric CIs: bootstrap resampling

Error bars

- An important use case for CIs
- But be careful! Error bars can potentially represent many things:
  - Confidence intervals (CIs)
  - Standard deviation (std)
  - Standard error of the mean: std/sqrt(n)
- → Always ask, always tell what the CIs represent!

Multiple-hypothesis testing

- If you perform experiments over and over, you're bound to find something
- If you consider "at least one positive outcome" to be the manifestation of an underlying effect: Significance level must be adjusted down when performing multiple hypothesis tests!

P(detecting an effect when there is none) = $\alpha = 0.05$

P(detecting no effect when there is none) $= 1 - \alpha$

P(detecting no effect when there is none, on every experiment) $= (1 - \alpha)^k$

P(detecting an effect when there is none on <u>at least one</u> experiment) $= 1 - (1 - \alpha)^k$

$\alpha = 0.05$

"Familywise Error Rate"

Family-wise error rate corrections

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

Part 3
Relating two variables

Pearson's correlation coefficient

- "Amount of linear dependence"

- More general:
  - Rank correlation, e.g., Spearman's correlation coefficient
  - Mutual information

Correlation coefficients are tricky!

- http://guessthecorrelation.com/
- Correlation != causation (cf. lecture in 2 weeks)
  http://www.tylervigen.com/spurious-correlations

Anscombe's quartet

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

Anscombe's quartet

Illustrates the **importance of looking at a set of data graphically** before starting to analyze

Highlights the *inadequacy of basic statistical properties for describing realistic datasets*

More on Wikipedia

UC Berkeley gender bias (?)

Admission figures from 1973

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Simpson's paradox

When a trend appears in different groups of data **but disappears or reverses when these groups are combined** -- beware of aggregates!

In the previous example, **women tended to apply to competitive departments** with low rates of admission

Summary

- Understand your data with descriptive statistics
  - Choose the right stats based on type of distribution
- Be certain to quantify your uncertainty
  - Hypothesis testing
  - Confidence intervals (preferred!)
  - Careful when performing multiple tests (apply correction)
- Relating 2 variables to one another
  - Correlation != causation
  - Even tricker with >2 variables (→ next lecture!)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec4-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo? / Où est Charlie?
- …

Applied Data Analysis (CS401)

Lecture 5
Regression for disentangling data
18 Oct 2023

Robert West

Announcements

- Homework H1 due end of next week
  - Due Fri 27 Oct 23:59
- Project milestone P1 feedback to be released next week
- Final exam has been scheduled: Tue 16 Jan 2024, 15:15–18:15
- Friday's lab session:
  - Quiz 4
  - Exercise on regression analysis (in BCH 2201)
  - Homework office hours (on Zoom, in parallel to exercise)
- Indicative course feedback is being collected (until Sun 22 Oct)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Course eval (“indicative feedback”) open until 22 Oct

Go to https://isa.epfl.ch now!

Linear regression

Credits

- Much of the material in this lecture is based on Andrew Gelman and Jennifer Hill's great book "Data Analysis Using Regression and Multilevel/Hierarchical Models", available for free here
- For a neat and gentle written intro to linear regression, especially check out chapters 3 and 4

What you should already know about linear regression

POLLING TIME

- **“How familiar are you with linear regression?”**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Linear regression as you know it

- **Given:** n data points (X_i, y_i), where X_i is k-dimensional vector of predictors (a.k.a. features) of i-th data point, and y_i is scalar outcome
- **Goal:** find the optimal coefficient vector $\beta = (\beta_1, \ldots, \beta_k)$ for approximating the y_i's as a linear function of the X_i's:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

Scalar product (a.k.a. dot product) of 2 vectors

where ϵ_i are error terms that should be as small as possible

- X_{i1} usually the constant 1 (by def) $\Rightarrow \beta_1$ a constant intercept

Example with one predictor

$$y \approx \beta_1 + \beta_2 X$$

Linear regression as you know it

- **Given:** n data points (X_i, y_i), where X_i is k-dimensional vector of predictors (a.k.a. features), and y_i is scalar outcome, of i-th data point
- **Goal:** find the optimal coefficient vector $\beta = (\beta_1, \ldots, \beta_k)$ for approximating the y's as a linear function of the X's:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

where ϵ_i are error terms that should be as small as possible

- X_{i1} usually the constant 1 $\rightarrow \beta_1$ a constant intercept

Optimality criterion: least squares

$$y_i \quad = \quad X_i\beta + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

- Intuitively, want errors ϵ_i to be as small as possible
- Technically, want sum of squared errors as small as possible
  ⇔ find $\hat{\beta}$ such that we minimize

$$\sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2$$

Use cases of regression

- **Prediction:** use fitted model to estimate outcome *y* for a new *X* not seen during model fitting (if you've seen regression before, then probably in the context of prediction)
- **Descriptive data analysis:** compare average outcomes across subgroups of data (today!)
- **Causal modeling:** understand how outcome *y* changes when you manipulate predictors *X* (next lecture is about causality, although not primarily using regression)

Regression as comparison of average outcomes

Example with one binary predictor X_i

- X_i = mom_hs = "Did mother finish high school?" $\in$ {0, 1} (No, Yes)
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 78 + 12 · mom_hs + error

mean kid_score for moms who **didn't finish** high school: 78

mean kid_score for moms who **finished** high school: 78 + 12 = 90

One binary predictor X_i: Interpretation of fitted parameters β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: mean outcome for data points i with $X_i = 0$
- **Slope** β_2: difference in mean outcomes between data points with $X_i = 1$ and data points with $X_i = 0$
- Reason: means minimize least-squares criterion: $\sum^n_{i=1} (y_i - m)^2$ is minimized w.r.t. m when $-2 \sum^n_{i=1} (y_i - m) = 0$, i.e., when $m = (1/n) \sum^n_{i=1} y_i$

One binary predictor X_i: Interpretation of fitted parameters β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: mean outcome for data points i with $X_i = 0$
- **Slope** β_2: difference in mean outcomes between data points with $X_i = 1$ and data points with $X_i = 0$
- Reason: means minimize least-squares criterion:

$\sum^n_{i=1} (y$ [illegible] when

$-2 \sum^n_{i=1}$ [illegible] $= (1/n$ [illegible] y_i

So why not just compute the two means separately and then compare them?

What a mean monkey!

Example with one continuous predictor X_i

- X_i = mom_iq = mother's IQ score $\in$ [70, 140]
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 26 + 0.6 · mom_iq + error

estimated (hypothetical) mean kid_score for moms with IQ = 0: 26

estimated mean kid_score for moms with IQ = 100: 26 + 0.6 · 100 = 86

One continuous predictor X_i: Interpretation of fitted parameters β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: estimated mean outcome for data points i with $X_i = 0$
- **Slope** β_2: difference in estimated mean outcomes between data points whose X_i's differ by 1
- Why "estimated"? $\longrightarrow$ e.g.,
- NB: for binary predictor, we got "exact" instead of "estimated"

Example with multiple predictors

- (X_{i1} = 1 = constant)
- X_{i2} = mom_hs = "Did mother finish high school?" $\in$ {0, 1} (No, Yes)
- X_{i3} = mom_iq = mother's IQ score $\in$ [70, 140]
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error

Example with multiple predictors

kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error

kids of moms who **finished** high school:
intercept = 26 + 6 = 32
slope = 0.6

kids of moms who **didn't finish** high school:
intercept = 26
slope = 0.6

Example with interaction of predictors

- X_{i2} = mom_hs = "Did mother finish high school?" $\in \{\overset{\text{No}}{0}, \overset{\text{Yes}}{1}\}$
- X_{i3} = mom_iq = mother's IQ score $\in [70, 140]$
- y_i = kid_score = child's score on cognitive test $\in [0, 140]$

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i2} X_{i3} + \epsilon_i$$

kid_score = −11 + 51 · mom_hs + 1.1 · mom_iq − 0.5 · mom_hs · mom_iq + error

Example with interaction of predictors

kid_score = −11 + 51 · mom_hs + 1.1 · mom_iq − 0.5 · mom_hs · mom_iq + error

kids of moms who **didn't finish** high school:
intercept = −11
slope = 1.1

kids of moms who **finished** high school:
intercept = −11 + 51 = 40
slope = 1.1 − 0.5 = 0.6

So why not just compute the two means separately and then compare them?

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	avg kid_score 90	avg kid_score 90
Mom didn't finish high school	avg kid_score 78	avg kid_score 78

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	avg kid_score 90	avg kid_score 90
Mom didn't finish high school	avg kid_score 78	avg kid_score 78

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

THINK FOR A MINUTE:

What is the mean outcome for Mercedes-driving moms vs. for non-Mercedes-driving moms? Compare the two means! What does the comparison tell you about the link between Mercedes-driving and kid_score?

(Feel free to discuss with your neighbor.)

- Mean kid_score for Mercedes drivers: **0.99 · 90 + 0.01 · 78 ≈ 90**
- Mean kid_score for non-Mercedes drivers: **0.01 · 90 + 0.99 · 78 ≈ 78**
- But really driving Mercedes makes no difference (for fixed high-school predictor)!
- Root of evil: **correlation** between finishing high school and driving Mercedes
- **Regression** to the rescue: **kid_score = 78 + 12 · mom_hs + 0 · mercedes + error**

	Mercedes	No Mercedes
Mom finished high school	mean kid_score 90	mean kid_score 90
Mom didn't finish high school	mean kid_score 78	mean kid_score 78

	Mercedes	No Mercedes
Mom finished high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

Course eval (“indicative feedback”) open until 22 Oct

Go to https://isa.epfl.ch now!

Quantifying uncertainty

Quantifying uncertainty

- Statistical software gives you more than just coefficients β:

```
Residuals:
    Min      1Q  Median      3Q     Max
-52.873 -12.663   2.404  11.356  49.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
mom.hs       5.95012    2.21181   2.690  0.00742 **
mom.iq       0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-Squared: 0.2141,     Adjusted R-squared: 0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

p-value: probability of estimating such an extreme coefficient if the true coefficient were zero (= null hypothesis)

Residuals and R^2

- **Residual** for data point *i*: estimation error on data point *i*:

$$r_i = y_i - X_i\hat{\beta}$$

- Mean of residuals = 0

  (total overestimation = total underestimation)
- Variance of residuals

  = avg squared distance of predicted value from observed value

  = "unexplained variance"
- Fraction of variance explained by the model:

$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

Variance of outcomes *y*

Residuals and R^2

- **Residual** for data point *i*: estimation error on data point *i*:

$$r_i = y_i - X_i\hat{\beta}$$

duals = 0

timation = total underestimation)

esiduals

avg squared distance of predicted value from observed value

= "unexplained variance"

Fraction of variance explained by the model:

$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

Variance of outcomes *y*

Coefficient of determination: R^2

$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

$R^2 = 0.147$

$R^2 = 0.865$

Coefficient of determination: R^2

Coefficient of determination: R^2

$R^2 = 0.67$ everywhere!

Assumptions made in regression modeling

Assumptions for regression modeling

1. Validity:
   a. Outcome measure should accurately reflect the phenomenon of interest
   b. Model should include all relevant predictors
   c. Model should generalize to cases to which it will be applied

Assumptions for regression modeling (2)

2. Linearity:

$$
\begin{aligned}
y_i &= X_i\beta + \epsilon_i \\
&= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n
\end{aligned}
$$

But very flexible: we require linearity in *predictors* (not necessarily in raw inputs); predictors can be arbitrary functions of raw inputs, e.g.,

- logarithms, polynomials, reciprocals, …
- interactions (i.e., products) of multiple inputs
- discretization of raw inputs, coded as indicator variables

Assumptions for regression modeling (3)

3. Independence of errors: no interaction between data points
4. Equal variance of errors
5. Normality (Gaussianity) of errors

} less important in practice

Transformations of predictors and outcomes

Transformations of predictors

- When we apply linear transformations to predictors, the model remains "equally good":
  - The fitted coefficients may change, but predicted outcomes and model fit (R^2) won't change
- For instance,

$$\begin{aligned} \text{earnings} &= -61000 + 51 \cdot \text{height (in millimeters)} + \text{error} \\ \text{earnings} &= -61000 + 81000000 \cdot \text{height (in miles)} + \text{error.} \end{aligned}$$

Mean-centering of predictors

- Compute the mean value of a predictor over all data points, and subtract it from each value of that predictor:
  $X_{ij} \leftarrow X_{ij} - \text{mean}(X_{1j}, \ldots, X_{nj})$
- $\Rightarrow$ the predictor X_{ij} now has mean 0

(hypothetical) mean kid_score for moms with IQ = 0: 26

mean kid_score for moms with mean IQ: 86

After mean-centering of predictors, …

… you have a convenient interpretation of coefficients β_j of main predictors (i.e., non-interaction predictors):

- $j = 1$ (i.e., intercept):
  - Estimated mean outcome when each predictor has its mean value
- $j > 1$:
  - Model w/o interactions: estimated mean increase in outcome y for each unit increase in X_{ij}
  - Model with interactions: estimated mean increase in outcome y for each unit increase in X_{ij} **when each other predictor has its mean value**

Standardization via z-scores

- First **mean-center** all predictors, then **divide them by their standard deviations:**
  $X_{ij} \leftarrow [X_{ij} - \text{mean}(X_{1j}, \ldots, X_{nj})] / \text{sd}(X_{1j}, \ldots, X_{nj})$
- Resulting values are called "**z-scores**"
- All predictors now have the same units:
  distance (in terms of standard deviations) from the mean
- This lets us compare coefficients for predictors with previously incomparable units of measurement, e.g., IQ score vs. earnings in Swiss francs vs. height in centimeters

Logarithmic outcomes

- **Practical:** makes sense if the outcome *y* follows a heavy-tailed distribution
- Only works for non-negative outcomes
- **Theoretical:** turns an additive model into a **multiplicative model:**

$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

Exponentiating both sides yields

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i \end{aligned}$$

Logarithmic outcomes: Interpreting coefficients

$$
\begin{aligned}
y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\
&= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i
\end{aligned}
$$

- An **additive** increase of 1 in predictor $X_{\cdot 1}$ is associated with a **multiplicative** increase of $B_1 := \exp(b_1)$ in the outcome
- If $b_1 \approx 0$, we can immediately interpret b_1 (without needing to exponentiate it first to get B_1!) as the **relative increase** in outcomes, since $\exp(b_1) \approx 1 + b_1$
- E.g., $b_1 = 0.05 \Rightarrow B_1 = \exp(b_1) \approx 1.05$
  $\Rightarrow$ "+1 in predictor $X_{\cdot 1}$" is associated with "+5% in outcome"

Going beyond linear regression for comparing means

Beyond linear regression: generalized linear models

- Logistic regression: binary outcomes
- Poisson regression: non-negative integer outcomes (e.g., counts)

Beyond comparing means; or, A taste of causality: “Difference in differences”

- Two groups: *P*, *S*
- At time 2, group *P* receives a **treatment**, group *S* doesn’t
- Question: Did the treatment have an **effect?** If so, how large was it?
- *P* and *S* don’t start out the same at time 1
- There is a temporal “baseline effect” even w/o treatment

Beyond comparing means; or, A taste of causality: “Difference in differences” (2)

- Elegant linear model with binary predictors:
  $$y_{it} = a + b \cdot \text{treated}_i + c \cdot \text{time2}_t + d \cdot (\text{treated}_i \cdot \text{time2}_t) + \text{error}_i$$
- d = treatment effect
- All of this with one single regression!
- You get quantification of uncertainty (significance) for free!

Beyond comparing means; or, A taste of causality: “Difference in differences” (2)

- Elegan
  predictors:
  $y_{it} = a + b \cdot \text{treated}_i + c \cdot \text{time2}_t$
  $+ d \cdot (\text{treated}_i \cdot \text{time2}_t) + \text{error}$
- d = treatment effect
- All of this with one single regression!
- You get quantification of uncertainty (significance) for free!

A bonanza of causality: Next lecture!

Summary

- Linear regression as a tool for comparing means across subgroups of data
- How? Read group means off from fitted coefficients
- Advantages over plain comparison of means “by hand”:
  - Accounting for correlations among predictors
  - Quantification of uncertainty (significance) “for free”
  - Additive vs. multiplicative model: all it takes is a log
- Caveat emptor:
  - Model must be appropriately specified, else nonsense results $\longrightarrow$ stay critical, run diagnostics (e.g., R^2, data viz)

Course eval (“indicative feedback”) open until 22 Oct

Go to https://isa.epfl.ch now!

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Credits

- Much of the material in this lecture is based on Andrew Gelman and Jennifer Hill's great book "Data Analysis Using Regression and Multilevel/Hierarchical Models", available for free here
- For a neat and gentle written intro to linear regression, especially check out chapters 3 and 4

Bonus: Logarithmic outcomes and predictors

Interpretation of coefficient of logarithmic predictor:

- **Multiplicative** increase by 1% in predictor $X_{\cdot 1}$ is associated with a **multiplicative** increase by b_1% in the outcome
- Why?
  - $\log(y) = a + b \log(X) \Rightarrow y = \exp(a) * X^b$
  - Multiplying X by a factor c multiplies y by a factor of c^b
  - $c^b \approx 1 + b*(c-1)$ for $c \approx 1$ (hint: Taylor approximation!)
  - Example when using $c = 1.01$ (i.e., increase by 1%):
    $b = 2 \Rightarrow$ increasing X by 1% increases y by 2%

Applied Data Analysis (CS401)

Lecture 6
Causal analysis of observational data
25 Oct 2023

Robert West

Announcements

- Homework H1 due on Fri 27 Oct 23:59
  - You can ask questions until Thu 23:59; we won't answer questions asked after Thu
- Project:
  - Milestone P1 feedback has been released
  - Milestone P2: get cracking once homework H1 is done!
  - Don't use ChatGPT (one student already penalized on P1)
- Friday's lab session:
  - Exercises on causal analysis of observational data
  - Quiz 5

Course evaluation

- Thanks for your feedback! -- We'll use it to improve the class further
- Most of you like the class and what you learn
- Some concerns: class too hands-off ("more code!"); class too hands-on ("more theory!"); exercise sessions too crowded (?); quiz difficulty, wording, timing

Quiz scores so far

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec6-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Dr. Bob's smoking cure

- I claim to have developed a medicine that helps you quit smoking
- I ask all smokers: "Do you want to try my medicine?"
- Smokers = {treated smokers}
  $\cup$ {untreated ("control") smokers}
- Fraction of successful quitters is higher in the treated group
- I conclude: "My medicine helps you quit smoking! Buy it!"
- Do you believe me?

Goals of this lecture

- Clarify difference between **experimental** and **observational** studies
- Highlight **pitfalls** of observational studies
- Give you **tools** for avoiding the pitfalls, allowing you to draw valid conclusions from "found data" (very useful for project!)
- Motivate you to **read** Rosenbaum's great book "Design of Observational Studies" (in particular Chapters 1, 2, and 3; or this book) and Pearl's eye-opening "Book of Why"

Dr. Bob's "experiment" as a causal diagram

Dr. Bob's "experiment" as a causal diagram

Ideal setting as a causal diagram

Randomized controlled experiments

- Two experimental conditions:
  - Treatment (e.g., medicine)
  - Control (e.g., placebo [fun fact])
- Assignment of participants to conditions is random
  - Probability of receiving treatment same for everyone
- Treatment and control groups are indistinguishable
  - E.g., determination to quit smoking is not systematically higher in the treated group

Randomized controlled experiment as a causal diagram

Randomized controlled experiment as a causal diagram

Randomized controlled experiment as a causal diagram

Limits of randomization

- Do seat belts save lives?
- Experiment:
  - Flip coin at birth to assign to treatment (always wear seat belt for entire life) or control (never wear seat belt)
  - Measure fraction of traffic deaths in each group
- Randomized experiments aren't always feasible
  - Unethical (see above), expensive, fundamentally impossible (e.g., do earthquakes decrease life spans?)
  - Most modern "big data" is "found data"
- Experiments may lead to unrealistic scenarios

Alternative: observational studies

- Fundamentally different from experiment:
  - Researcher can’t control who goes to which condition
  - Researcher is merely an observer, not a tinkerer
  - Much less problematic w.r.t. ethics, price, feasibility
  - Much more problematic w.r.t. validity of conclusions
- All advantages of randomized experiment are gone
  - Subjects self-select to be treated
  - Treatment assignment and response may be caused by same hidden correlate (a.k.a. confounders; e.g., motivation to quit smoking)

Example: seat belts revisited

- Recall: experiment infeasible because unethical
- Observational study:
  - Dataset: all traffic accidents in a given time span
  - Two treatment conditions:
    - Treated: seat-belt wearers
    - Control: non-seat-belt wearers
  - Compare fraction dead in treated vs. control
- What problems do you see?

- Two treatment conditions:
  - Treated: seat-belt wearers
  - Control: non-seat-belt wearers
- Compare fraction dead in treated vs. control

- What problems do you see?

THINK FOR A MINUTE!

(Feel free to discuss with your neighbor.)

As a causal diagram

A matched observational study

- Consider only a particular subset of accident cars:
  - 2 people in car: driver + passenger
  - Exactly one of them died in accident
  - Exactly one of them wore seat belt at time of accident (i.e., 1 treated + 1 control per car)
- As before: compare fraction dead in treated vs. control
- New: many potential confounders are controlled for, incl. type of car, speed, severity of accident
- Fundamental concept: **matching**

As a causal diagram

Settling the seat-belt question

	Driver Passenger	Not Belted Belted	Belted Not Belted
Driver Died	Passenger Survived	189	153
Driver Survived	Passenger Died	111	363

Natural experiments

- Not researcher, but nature, “flips a coin” to decide treatment assignment
- Rosenbaum: “When investigators are especially proud, having found unusual circumstances in which treatment assignment, though not random, seems unusually haphazard, they may speak of a ‘natural experiment.’”
- Examples: twin studies, Vietnam draft, cholera in London
- Is matched seat-belt study a natural experiment?

NEWS | 13 October 2021

Nobel-winning 'natural experiments' approach made economics more robust

Joshua Angrist, Guido Imbens and David Card share the prize for finding a way to identify cause and effect in social science.

Philip Ball

Nature didn’t flip a coin for me – should I just go home and weep?

Commercial break

Nature didn’t flip a coin for me – should I go home and weep?

- No! You can still get good mileage if you’re smart about it
- Fundamental concept: **matching**
- Ideally: Pair up 2 identical people:
  - 1 treated, 1 control
  - Ex-post (vs. natural experiment: ex-ante)
- Compare outcome of treated vs. control
  - e.g., mean difference treated-minus-control
  - or regression analysis (see last lecture)

Matching

- Ideally: Pair up 2 identical people:
  - 1 treated, 1 control
- Such ideal matching usually not feasible

  - **Problem 1: Unobserved covariates:**
    You usually can’t even know if two people are identical

  - **Problem 2: Combinatorial explosion:**
    (Nearly) no two people are identical

Problem 1: Unobserved covariates

- You usually can't even know if two people are identical
- e.g., (hypothetical) gene that causes both desire to smoke and lung cancer

Addressing Problem 1 by ignoring it: A naive model

"People who look comparable are comparable"

or equivalently:

"Only observed covariates determine treatment assignment"

Naive model as a causal diagram

If the naive model were true…

- … you could "simulate" a randomized experiment:
  - Simply match subjects with identical observed covariates (1 treated, 1 control)
  - Subjects in a pair have the same probability to treat
  - So who gets treated is up to chance, as in experiment
  - Analysis: compare outcome for treated to outcome of control (e.g., mean difference treated-minus-control)

Problem 2!

Problem 2: Combinatorial explosion

- (Nearly) no two people are identical
- So finding two people to match is often impossible
  - Even when considering only observed covariates (as in the naive model)
- Do we really need to match people with identical **covariates?**
  - Recall "holy grail": randomized controlled trials
  - Coin makes sure everyone has identical **probability to be treated**
  - ⟶ Let's mimic what the coin does!

Addressing Problem 2: Propensity score

- Compress the (potentially many) observed covariates into a single number: the probability to receive the treatment (a.k.a. **propensity score**):

  Pr(subject is treated | observed covariates)

- Can be estimated from the data
  - E.g., via logistic regression (see next lecture)
  - Input: observed covariates
  - Output: treatment indicator (1 if treated, 0 if control)

Propensity score as a causal diagram

Balancing property of propensity score

- Fact: all subjects (treated and control) with equal propensity score (PS) *p* have equal distribution of observed covariates **x**:

$$\Pr(\mathbf{x} \mid \text{treated} = 1, \text{PS} = p) = \Pr(\mathbf{x} \mid \text{treated} = 0, \text{PS} = p)$$

- Subjects in a matched pair might not have equal **x**, but treated and control groups will have similar distributions of **x**

Balancing property is propensity score's *raison d'être*

- There are many other methods for achieving balance
  - e.g., exact matching, Mahalanobis distance matching, coarsened exact matching, …
- You can mix and match methods (e.g., match exactly on gender, use propensity scores for other covariates)
- What eventually matters is whether you achieve balance
  - Regardless of *how you try to achieve* balance, you need to verify *that you managed to achieve* balance (p.t.o.)

Assessing covariate balance

	Treated	Control	Unmatched
Female %	67.3	67.3	46.3
Age (mean)	14.9	14.7	13.5
Black %	14.3	15.1	28.1
Hispanic %	14.3	14.9	39.9
BMI (mean)	22.3	22.4	22.5
BMI missing %	4.1	2.2	0.0
Family income/poverty (mean)	2.4	2.4	2.1
Family income/poverty missing %	0.0	0.6	7.0
$\log_{10}$ (1 + cotinine)	0.4	0.4	0.3
Cotinine missing	2.0	2.7	11.1
Propensity score (mean, as a %)	1.7	1.7	0.7

Matching algorithms

- Goal: Match subjects into pairs (1 treated, 1 control), with identical propensity scores within each pair
- Unlikely that 2 subjects have identical propensity scores
- → Use approximate matching (remember your algo class!)
- Bipartite graph: each subject connected to all other subjects
- Edge weights: absolute (or squared) difference of propensity scores (or other matching criterion)
- Find minimum matching, e.g., via [Hungarian algorithm]

Ok, so are we done?

Let's ignore Problem 1!

Problem 1: Unobserved covariates:

You usually can't even know if two people are identical

We've been assuming the naive model:

"Only observed variables determine treatment assignment"

If the naive model isn't true…

- … propensity score may differ from true probability to treat:

$$\underbrace{\Pr(\text{treated} \mid \text{observed covariates})}_{\textit{have this}} \neq \underbrace{\Pr(\text{treated} \mid \text{all covariates})}_{\textit{need this}}$$

Violated naive model as a causal diagram

If the naive model isn’t true…

… you may end up matching

Person 1	Person 2
▪ Born 1948; grew up in England ▪ Married the second time ▪ 2 children ▪ successful in business ▪ Wealthy ▪ Spend their winter holidays in the Alps ▪ Likes dogs	▪ Born 1948; grew up in England ▪ Married the second time ▪ 2 children ▪ successful in business ▪ Wealthy ▪ Spend their winter holidays in the Alps ▪ Likes dogs

The naive model is easily attacked

Rosenbaum:

> It is common for a critic to argue that, in a particular study, the naïve model may be false. Indeed, it may be false. Typically, the critic accepts that the investigators matched for the observed covariates, **x**, so treated and control subjects are seen to be comparable in terms of **x**, but the critic points out that the investigators did not measure a specific covariate u, did not match for u, and so are in no position to assert that treated and control groups are comparable in terms of u. This criticism could be dismissed in a randomized experiment — randomization does tend to balance unobserved covariates — but the criticism cannot be dismissed in an observational study. This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: "This would never happen in my laboratory."

The sensitivity analysis model

- Idea: Quantify the degree to which the naive model may be wrong without you having to change your (causal) conclusions
- Assume that treatment odds of identical-looking subjects (i.e., identical observed covariates **x**) may differ by up to a factor Γ
- Then reason in spirit of proof by contradiction: "To change the conclusions of my study, two identical-looking people (1 treated, 1 control) would have to have hugely different treatment odds (i.e., huge Γ). Common sense (or domain knowledge) suggests that this is not the case, so my conclusions stand."

The sensitivity analysis model

$$\frac{1}{\Gamma} \leq \frac{\pi_k / (1 - \pi_k)}{\pi_\ell / (1 - \pi_\ell)} \leq \Gamma \quad \text{whenever} \quad \mathbf{x}_k = \mathbf{x}_\ell \,. \qquad \Gamma \geq 1.$$

subject ℓ's (true) probability to treat

- Bounded odds ratio (OR)
  - Reason for using OR:
    OR = Pr(k treated | **either k or ℓ treated**) / Pr(ℓ treated | **either k or ℓ treated**)
- Sensitivity $\Gamma = 1 \longrightarrow$ naive model is true
- Sensitivity $\Gamma = 2 \longrightarrow$ subject with same observed covariates **x** up to twice as likely to be the one to receive treatment
- Sensitivity $\Gamma = \infty \longrightarrow$ void statement (a.k.a. tautology)

Example: smoking and lung cancer

- Under naive model: matching on observed covariates gives a very small *p*-value for the null hypothesis that smoking does not increase lung cancer risk (using an appropriate hypothesis test), i.e., data hard to explain w/o a causal effect
- Tobacco lobby: “The naive model isn’t true! There may be hidden (e.g., genetic) correlates that increase both the probability to enjoy smoking and the probability of lung cancer. They, not smoking, cause cancer!”

Example: smoking and lung cancer

- Under sensitivity analysis model, increasing sensitivity Γ increases the p-value for null hypothesis
- Anti-tobacco lobby: "But making $p > 0.05$ would require $\Gamma > 6$; i.e., the odds of being a smoker would need to be six times higher for one of two people with the exact same observed features (age, gender, education, income, ...). It's unlikely that any unobserved covariate would have such a large effect on smoking habits. So smoking causes cancer!"

Test the West!
Super Geschmack. Super Preis.
West
West
West
Der Bundesgesundheitsminister: Rauchen gefährdet Ihre Gesundheit. Der Rauch einer Zigarette dieser Marke enthält: West 1,0 mg Nikotin und 15 mg Kondensat (Teer).
West «100» 1,2 mg N und 15 mg K. (Durchschnittswerte nach DIN).

„... und dann fand ich ein Angebot nach meinem Geschmack.“
9,90 €
34 STK
West
ORIGINAL
West
SILVER
Rauchen ist tödlich
Made for good times
LIDL

Two parts: mechanical vs. scientific

- Mechanical part:
  - Create pairs (1 treated + 1 control) with similar observed covariates (using exact or propensity-score matching)
- Scientific (i.e., fun) part:
  - Mitigate concerns that your findings might be caused by unobserved covariates, rather than treatment (e.g., using sensitivity analysis, ad-hoc arguments, natural experiments)

Summary

- Holy grail: randomized experiment
- When experiment not possible: observational study
- Crucial problem: treatment assignment not random (biases!)
- Semi-holy grail: natural experiment
- Matched studies: pair up treated/control based on observed covariates
- Problem: still, treatment assignment not random (biases via unobserved covariates)
- Solution: sensitivity analysis
- Keep this lecture (more here) in mind for your projects!

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec6-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Credits

- Much of the material is based on Paul Rosenbaum's amazing book "Design of Observational Studies", available for free here

Applied Data Analysis (CS401)

Lecture 7
Learning from data: Supervised learning
1 Nov 2023

Robert West

Annowwwwwncements

- Happy (belated) Halloween!
- Homework H1 is being graded. Feedback to be released next week.
- Project milestone P2 due on Fri 17 Nov
- Friday's lab session: two parallel tracks:
  - Quiz 6
  - Track 1: exercise on supervised learning (in BCH 2201)
  - Track 2: project office hours (on Zoom)
    - Logistics: see Ed post
    - Do come and ask for feedback – everyone will win!

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec7-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Pumpkin Pete?
- …

Why ML as part of data science?

- ML can facilitate most steps of the data analysis cycle

- But stay critical: “Can I trust my ML model?”

Machine learning

- **Supervised:** We are given input/output pairs (X, y) (a.k.a. "samples") that are related via a function $y = f(X)$. We would like to "learn" f, and evaluate it on new data. Types:
  - Discrete y (class labels): **"classification"**
  - Continuous y: **"regression"** (e.g., linear regression)
- **Unsupervised:** Given only samples X of the data, we compute a function f such that $y = f(X)$ is a "simpler" representation.
  - Discrete y (cluster labels): **"clustering"**
  - Continuous y: **"dimensionality reduction"**

Machine learning: examples

- **Supervised (*lecture 7, i.e., today*):**
  - Is this image a cat, dog, or cow?
  - How would this user rate that restaurant?
  - Is this email spam?
  - Is this blob on a telescope image a supernova?
- **Unsupervised (*lecture 9*):**
  - Cluster handwritten digit data into 10 classes
  - What are the top 20 topics in Twitter right now?
  - Find the best 2D visualization of 1000-dimensional data

Machine learning: techniques

- **Supervised learning:**
  - k-NN (k nearest neighbors)
  - Tree-based models: decision trees, random forests
  - Linear + logistic regression
  - Naïve Bayes
  - Support vector machines
  - Supervised neural networks
  - etc.
- **Unsupervised learning:**
  - Clustering
  - Dimensionality reduction: topic modeling, matrix factorization (PCA, SVD, word2vec)
  - Hidden Markov models (HMM)
  - etc.

Today

(particularly in light of bias/variance tradeoff)

Lectures 9, 10, 11

Intro to supervised learning: k nearest neighbors (k-NN)

k nearest neighbors (k-NN)

Given a query item:

Find k closest matches
in a labeled dataset ↓

k nearest neighbors (k-NN)

Given a query item:

Find k closest matches in a labeled dataset ↓

Return the most frequent label among the k

k nearest neighbors (k-NN)

k = 3 votes for “cat”

Properties of k-NN

The data is the model

- No training needed.
- Conceptually simple algorithm.
- Accuracy generally improves with more data.
- Usually need data in memory, but can also be run from disk.

Minimal configuration:

- Only one parameter: k (number of neighbors)
- But two other choices are also important:
  - Similarity metric
  - Weighting of neighbors in voting (e.g. by similarity)

k-NN flavors

Classification:

- Model is y = f(X), y is from a discrete set (labels).
- Given X, compute y = majority vote of the k nearest neighbors.
- Can also use a weighted vote* of the neighbors.

Regression:

- Model is y = f(X), y is a real value.
- Given X, compute y = average value of the k nearest neighbors.
- Can also use a weighted average* of the neighbors.

* Weighting function is usually the similarity.

k-NN distance (opposite of similarity) measures

- **Euclidean Distance:** Simplest, fast to compute

$$d(x, y) = \|x - y\|$$

- **Cosine Distance:** Good for documents, images, etc.

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

- **Jaccard Distance:** For set data:

$$d(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

- **Hamming Distance:** For string data:

$$d(x, y) = \sum_{i=1}^{n} (x_i \neq y_i)$$

k-NN distance (opposite of similarity) measures

- **Manhattan Distance:** Coordinate-wise distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Edit Distance:** for strings, especially genetic data.

```
stack.push(kNN)
```

Predicting from samples

- Most datasets are **samples** from a (maybe infinite) **population**.
- We are most interested in **models of the population**, but we only have access to a **sample** (blue points) from the population.

For a dataset consisting of pairs (X, y):

- features X, label y,

we aim to find the true model f:

- $y = f(X)$.

We train on a training sample D and denote the fitted model as f_D

Bias and variance

- Given a random training sample D, obtain model f_D
- For a new data point (X, y), prediction is $f_D(X)$
- (Squared) error = $\mathrm{E}[(f_D(X) - y)^2]$ (E is expectation over D!)
- Fact: error can be decomposed into two parts (derivation)
  - $\text{Error}^2 = \text{Bias}^2 + \text{Variance}$
  - $\text{Bias} = \mathrm{E}[f_D(X) - y]$
  - $\text{Variance} = \mathrm{E}[(f_D(X) - \mathrm{E}[f_D(X)])^2]$

Bias and variance

Our data-generated model $f_D(X)$ is a **statistical estimate** of the true function $f(X)$.

Because of this, its subject to bias and variance:

Bias: if we train models $f_D(X)$ on many training sets D, bias is the expected difference between their predictions and the true y's.

i.e.

$$Bias = \mathrm{E}[f_D(X) - y]$$

E[] is taken over points X and datasets D

Variance: if we train models $f_D(X)$ on many training sets D, variance is the variance of the estimates:

$$Variance = \mathrm{E}\left[\left(f_D(X) - \bar{f}(X)\right)^2\right]$$

Where $\bar{f}(X) = \mathrm{E}[f_D(X)]$ is the average prediction on X.

Consider a fixed testing point (X, y) not seen during training

"Full" bias/variance: average this picture over all testing points (X, y)

Bias/variance tradeoff

Since $\text{Error}^2 = \text{Bias}^2 + \text{Variance}$, there is a tradeoff, usually modulated via model complexity:

Complex models (many parameters) usually have lower bias, but higher variance.

Simple models (few parameters) have higher bias, but lower variance.

Bias/variance tradeoff

Example: A linear model can only fit a straight line. A high-degree polynomial can fit a complex curve. But the polynomial will fit the individual training sample, rather than the full population. Its shape can vary from sample to sample, so it has high variance.

Bias/variance tradeoff

The total expected error is

$$Bias^2 + Variance$$

Because of the bias-variance trade-off, we want to **balance** these two contributions.

If $Variance$ strongly dominates, it means there is too much variation between models. This is called **over-fitting**.

If $Bias$ strongly dominates, then the models are not fitting the data well enough. This is called **under-fitting**.

```
kNN = stack.pop()
```

Choosing *k* for *k* nearest neighbors

We have a bias/variance tradeoff:

- Small $k \rightarrow$?
- Large $k \rightarrow$?

THINK FOR A MINUTE:

When *k* increases,
how do bias and variance change?

(Feel free to discuss with your neighbor.)

Choosing k

- Small $k \longrightarrow$ low bias, high variance
- Large $k \longrightarrow$ high bias, low variance
- Assume the real data follows the blue curve, with zero-mean additive noise. Red points are a data sample.

Choosing k

- **Small k** $\longrightarrow$ low bias, high variance
- Large k $\longrightarrow$ high bias, low variance

Choosing k

- **Small k** $\longrightarrow$ low bias, high variance
- Large k $\longrightarrow$ high bias, low variance

Choosing *k*

- Small *k* ⟶ low bias, high variance
- **Large *k*** ⟶ high bias, low variance

Choosing *k*

- Small $k \longrightarrow$ low bias, high variance
- **Large k** $\longrightarrow$ high bias, low variance

Choosing *k* in practice

Use leave-one-out (LOO) cross-validation:

- **Split:** Break data into train and test subsets, e.g. 80–20 % random split.
- **Predict:** For each point in the training set, predict using the *k* nearest neighbors from the set of all *other* points in training set. Measure the LOO error rate (classification) or squared error (regression).
- **Tune:** Try different values of *k*, and use the one that gives minimum leave-one-out error.
- **Evaluate:** Measure error on the test set to quantify performance.

Commercial break

American Diabetes Association®

Data is our candy.

Decision trees

Decision trees: example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Decision trees

Model: Flow-chart-like tree structure

- Nodes are tests on a single attribute
- Branches are attribute values of parent node
- Leaves are marked with class labels

Goal: Find decision tree that maximizes classification accuracy on given dataset

Optimization:

- NP-hard
- Heuristic: greedy top-down tree construction + pruning

Decision tree induction

Tree construction (top-down divide-and-conquer strategy)

- At the beginning, all training samples belong to the root
- Examples are partitioned recursively based on selected “most discriminative” attributes (partitioning data into most homogeneous subsets)
- Discriminative power based on information gain (in ID3 and C4.5 algorithms) or Gini impurity (in CART algorithm)

Partitioning stops if

- All samples belong to the same class → assign the class label to the leaf
- There are no attributes left → majority voting to assign the class label to the leaf

Decision tree induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Decision tree induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Decision tree induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute selection

For a given branch in the tree, the set of samples *S* to be classified has *P* positive and *N* negative samples

The amount of entropy in the set S is

$$H(P,\ N) = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}$$

Note that:

- If $P = 0$ or $N = 0$: $H(P, N) = 0 \longrightarrow$ no uncertainty
- If $P = N$: $H(P, N) = 1 \longrightarrow$ maximum uncertainty

Attribute selection

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

$H_S = H(9, 5) = 0.94$

Age [<=30] H(2, 3) = 0.97

Age [31...40] H(4, 0) = 0

Age [>40] H(3, 2) = 0.97

Income [high] H(2, 2) = 1

Income [med] H(4, 2) = 0.92

Income [low] H(3, 1) = 0.81

Student [yes] H(6, 1) = 0.59

Student [no] H(3, 4) = 0.98

Rating [fair] H(6, 2) = 0.81

Rating [exc] H(3, 3) = 1

Attribute selection

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

$$H_S = H(9, 5) = 0.94$$

$$H_{Age} = p([<=30]) \cdot H(2, 3) + p([31...40]) \cdot H(4, 0) + p([>40]) \cdot H(3, 2) =$$
$$= 5/14 \cdot 0.97 + 4/14 \cdot 0 + 5/14 \cdot 0.97 = 0.69$$

$$H_{Income} = p([high]) \cdot H(2, 2) + p([med]) \cdot H(4, 2) + p([low]) \cdot H(3, 1) =$$
$$= 4/14 \cdot 1 + 6/14 \cdot 0.92 + 4/14 \cdot 0.81 = 0.91$$

$$H_{Student} = p([yes]) \cdot H(6, 1) + p([no]) \cdot H(3, 4) = 7/14 \cdot 0.59 + 7/14 \cdot 0.98 = 0.78$$

$$H_{Rating} = p([fair]) \cdot H(6, 2) + p([exc]) \cdot H(3, 3) = 8/14 \cdot 0.81 + 6/14 \cdot 1 = 0.89$$

Attribute selection

Attribute *A* partitions *S* into $S_1, S_2, \ldots S_v$

Entropy of attribute *A* is

$$H(A) = \sum_{i=1}^{v} \frac{P_i + N_i}{P + N} H(P_i, N_i)$$

The *information gain* obtained by splitting *S* using *A* is

$$Gain(A) = H(P, N) - H(A)$$

Gain(Age) = 0.94 – 0.69 = 0.25 ← split on age

Gain(Income) = 0.94 – 0.91 = 0.03

Gain(Student) = 0.94 – 0.78 = 0.16

Gain(Rating) = 0.94 – 0.89 = 0.05

Pruning

The construction phase does not filter out noise → **overfitting**

Many possible pruning strategies

- Stop partitioning a node when the corresponding number of samples assigned to a leaf goes below a threshold
- Bottom-up cross validation: Build the full tree and replace nodes with leaves labeled with the majority class if classification accuracy on a **validation set (not seen during training!)** does not get worse this way

Comments

Decision trees are an example of a classification algorithm

- Many other out there (k-NN, naive Bayes, SVM, neural networks, logistic regression, random forests ...)

Maybe not the best one ...

- Sensitive to small perturbation in the data (high variance)
- Tend to overfit
- Non-incremental: Need to be re-trained from scratch if new training data becomes available

Decision tree models

- As tree depth increases, how do bias and variance change? (Hint: think about k-NN)

POLLING TIME

- Scan QR code or go to https://web.speakup.info/room/join/66626

Decision tree models

- As tree depth increases, bias decreases and variance generally increases. Why? (Hint: think about k-NN)

Bias decreases
with tree depth

Variance increases
with tree depth

Ensemble methods

Are, metaphorically, like **“democratic” machine learning algorithms**:

- Take a collection of simple or *weak* learners
- Combine their results to make a single, better learner

Types:

- **Bagging:** train learners in parallel on different samples of the data, then combine by voting (for discrete output) or by averaging (for continuous output).
- **Boosting:** train learner again, but after filtering/weighting samples based on output of previous train/test runs.
- **Stacking:** combine outputs from various models using a second-stage learner (e.g., linear regression).

Random forests

Grow K trees on datasets sampled from the original dataset (size N) with replacement (bootstrap samples), p = number of features.

- Draw K bootstrap samples of size N
- Grow each decision tree by selecting a random set of m out of p features at each node and choosing the best feature to split on.
- At testing time, aggregate the predictions of the trees (most popular vote, or average) to produce the final class (example of bagging).

Typically m might be e.g. sqrt(p), but can be smaller.

Random forests

Principles: we want to take a vote between different learners so we don't want the models to be too similar. The following two criteria ensure diversity in the individual trees:

- Draw K bootstrap samples of size N:
  - Each tree is trained on different data.
- Grow a decision tree by selecting a random set of m out of p features at each node, and choosing the best feature to split on.
  - Corresponding nodes in different trees (usually) can't use the same feature to split on.

Random forests

- **Very popular in practice,** probably the most popular classifier for dense data (up to a few thousand features)
- **Easy to implement** (simply train many normal decision trees)
- **Easy to parallelize**

- **Needs many passes over the data** – at least the max depth of the trees (<< boosted trees though, cf. next slide)

Boosted decision trees

- A more recent alternative to random forests (RF) [good intro here]
- In contrast to RF, whose trees are trained **independently** by bagging, BDT trees are trained **sequentially** by **boosting**: Each tree is trained to predict ("correct") residual errors of previous trees ($\longrightarrow$ bias reduction).
- Final prediction: sum of predictions made by individual trees.
- Both RF and boosted trees can produce very high-quality models. Superiority of one method or the other is dataset-dependent.
- Resource requirements are very different as well, so it's actually non-trivial to compare the methods.

Random forests vs. boosted trees

- The “geometry” of the methods is very different:
- Random forests use 10’s of deep, large trees:

Random forests vs. boosted trees

- The “geometry” of the methods is very different:
- Boosted decision trees use 1000’s of shallow, small trees:

Random forests vs. boosted trees

- RF training embarrassingly parallel, can be very fast
- Evaluation of trees (runtime) also much faster for RFs

For your personal perusal:

“A visual introduction to machine learning”

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Linear and logistic regression

Linear regression

- Your good friend from lecture 5 on regression analysis
- Goal: find the “best” line (linear function $y = f(X)$) to explain the data

How to model binary events?

- E.g., X: student features; y: did student pass ADA?
- Desired output: $f(X)$ = probability of passing ADA, given feats X
- Problem with linear regression:
  $f(X)$ can be below 0 or above 1

Logistic regression

- Trick: don't deal with probabilities, which range from 0 to 1, but with log odds, which range from $-\infty$ to $+\infty$
- Probability y ⇔ odds $y/(1-y)$ ⇔ log odds $\log[y/(1-y)]$
- Model log odds as a linear function of X

Logistic regression

- Model log odds as a linear function of X
- $\beta^T X = \log[y/(1-y)]$
- Solve for y: $y = 1 / (1 + \exp(-\beta^T X))$
- Finding best model β via maximum likelihood:
  - Don't use square loss as in linear regression (where y is assumed to be generated from Normal distribution)
  - Use cross-entropy loss instead (y assumed to be generated from Bernoulli distribution, i.e., biased coin)

Overfitting

- The more features the better?
  - **No!**
  - More features mean less bias, but more variance
  - Overfitting
- Carefully selected features can improve model accuracy
  - E.g., keep features that correlate with the label *y*
  - Forward/backward feature selection
  - Regularization (e.g., penalize norm of weight vector)
- More on such practical aspects: next lecture (“applied ML”)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec7-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Pumpkin Pete?
- …

Criteria

Predictive performance (accuracy, AUC/ROC, precision, recall, F1-score, etc.)

Speed and scalability

- Time to build the model
- Time to use the model
- In memory vs. on disk processing
- Communication cost

Robustness

- Handling noise, outliers, missing values

Interpretability

- Understanding the model and its decisions (black box vs. white box)

Compactness of the model

- Mobile and embedded devices

k-NN and the curse of dimensionality

The curse of dimensionality refers to “weird” phenomena that occur in high dimensions (100s to millions) that do not occur in low-dimensional (e.g. 3-dimensional) space.

In particular data in high dimensions are much sparser (less dense) than data in low dimensions.

For k-NN, this means there are fewer points that are very close in feature space (very similar) to the point X whose y we want to predict.

k-NN and the curse of dimensionality

From this perspective, it's surprising that kNN works at all in high dimensions.

Luckily real data are not like random points in a high-dimensional cube. Instead they live in **dense clusters** and near **much lower-dimensional surfaces**.

Also, points can be very "similar" even if their Euclidean distance is large. E.g. documents with the same few dominant words are likely to be on the same topic (--> use different distance)

k-NN and the curse of dimensionality

Example: Consider a collection of uniformly random points in the unit cube. In one dimension, the average squared Euclidean distance between any two points is:

$$\int_0^1 \int_0^1 (x-y)^2 dx\, dy = \frac{1}{6}$$

In N dimensions, we add up the squared differences for all N coordinates (because the coordinates are independent in a uniform random cube), giving:

$$d^2 = \mathrm{E}[\|x-y\|^2] = \frac{N}{6}$$

So the euclidean distance scales as $\sqrt{N}$

Applied Data Analysis (CS401)

ada

Lecture 8
Learning from data: Applied machine learning
8 Nov 2023

EPFL

Robert West

dlab

Announcements

- Homework H1
  - Feedback has been released
  - Postmortem: recorded video to be released early next week
- Project milestone P2 due on Fri 17 Nov 23:59
- Friday's lab session: two parallel tracks:
  - Quiz 7
  - Track 1: exercise on applied machine learning (BCH 2201)
  - Track 2: project office hours (Zoom)
    - Logistics: see Ed post
    - Do come and ask for feedback – everyone will win!

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec8-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Why an extra class on applied ML?

Machine Learning that Matters

Kiri L. Wagstaff KIRI.L.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

link

Classic ML class

ADA

It is easy to sit in your office and run a Weka (Hall et al., 2009) algorithm on a data set you downloaded from the web. It is very hard to identify a problem for which machine learning may offer a solution, determine what data should be collected, select or extract relevant features, choose an appropriate learning method, select an evaluation method, interpret the results, involve domain experts, publicize the results to the relevant scientific community, persuade users to adopt the technique, and (only then) to truly have made a difference (see Figure 1). An ML researcher might well feel fatigued or daunted just contemplating this list of activities. However, each one is a necessary component of any research program that seeks to have a real impact on the world outside of machine learning.

Classification pipeline

Data collection

The first step is collecting data related to the classification task.

- Definition of the attributes (or features) that describe a data item and the class label.

Domain knowledge is needed.

What if assigning the class label would be too time-consuming or even impossible?

→ Unsupervised methods (e.g., clustering); cf. next lecture!

Data collection

Features

Different types of features [more]

- Continuous (e.g., height, temperature ...)
- Ordinal (e.g., “agree”, “don’t care”, “disagree” ...)
- Categorical (e.g., country, gender ...)

New features can be generated from **simple stats**

- *Feature engineering* is considered a form of art, therefore it is often useful to find what other people did for similar problems

Some classifiers require categorical features => **discretization**

ML before 2012*

(but still very common today)

Much of the “heavy lifting” in here.
Final performance only as good as the feature set.

* Before publication of Krizhevsky et al.’s ImageNet CNN [paper]

A typical ML approach after 2012

Features and model learned together, mutually reinforcing

Data collection

Labels

Collecting a lot of data (features) is often easy.

Labeling data is time consuming, difficult, and sometimes even impossible.

Label: “Is page credible?”
Human dietary expert is needed

Potential labelers

- You
- Older days:
  - Undergraduate students
  - Domain experts ($$$)
- Now: crowdsourcing
  - Can get both amateurs (~ undergrad students)
  - and experts

"Is this webpage credible?"

1. Submit task

Requester

4. Collect answers

Credible
Credible
Not credible
Credible
Not credible

Crowdsourcing platform

Prolific

clickworker.com

amazon mechanical turk™
Artificial Artificial Intelligence

2. Accept task

Crowd workers

3. Return answers

Data collection

Discretization

Why?

- Some classifiers want discrete features (e.g., simplest kinds of decision trees)
- Discrete features let a linear classifier learn non-linear decision boundaries
- Certain feature selection methods require discrete (or even binary) features

Discretization

Unsupervised

- Equal width
  - Divide the range into a predefined number of bins (bad for skewed data, e.g., from a power law)
- Equal frequency
  - Divide the range into a predefined number of bins so that every interval contains the same number of values
- Clustering

Discretization

Supervised

- Start with very fine-grained discretization
- Test the hypothesis that membership in two adjacent intervals of a feature is independent of the class
- If they are independent, they should be merged
- Otherwise they should remain separate
- Independence test: χ^2 test ("chi-squared test") [example]
- Continue recursively

Data collection

Removing irrelevant features: Feature selection

- **Goal:** reduce set of N features to a subset of the best size $M < N$
- **Why?**
  - More interpretability
  - Less danger of overfitting
  - More efficient training
- **Problem:** There are 2^N possible subsets
- **Solutions:**
  - Filtering as preprocessing (“offline”)
  - Iterative feature selection (“online”)

Offline feature selection

Rank features according to their individual predictive power; then select the best ones

- Pros
  - Independent of the classifier (performed only once)
- Cons
  - Independent of the classifier (ignore interaction with the classifier)
  - Assumes features are independent

Ranking of features

Continuous features (and ideally labels):

- Pearson's correlation coefficient (capturing strength of linear [!] dependence)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Categorical features and labels:

- Mutual information (goes beyond linear dependence)

$$I(F;C) = H(C) - H(C \mid F) = H(F) + H(C) - H(F,C)$$

$$H(F) = -\sum_i P(f_i)\log_2 P(f_i)$$

$$H(F,C) = -\sum_i \sum_j P(f_i, c_j)\log_2 P(f_i, c_j)$$

Ranking of features

Categorical features and labels (cont'd):

- **χ^2 method** ("chi-squared")
  - Similar to χ^2 method for feature discretization
  - Test whether feature is independent of label
  - Difference w.r.t. mutual information: the χ^2 test checks the independence of the class and the feature, without indicating the strength or direction of any existing relationship (you just get a significance, a.k.a. *p*-value)

Ranking of features

Beware: collectively relevant features may look individually irrelevant!

Online feature selection

Forward feature selection: greedily **add** features; evaluate on validation dataset; stop when no improvement

- Pros
  - Interact with the classifier
  - No feature-independence assumption
- Cons
  - Computationally intensive

Online feature selection

Backward selection (a.k.a. ablation): greedily **remove** features; evaluate on validation dataset; stop when no improvement

- Pros
  - Interact with the classifier
  - No feature-independence assumption
- Cons
  - Computationally intensive

Commercial break

NETFLIX

ate questions on Tuesday's

District 2 seat are Harry
ayne Biggs, Harold Haines,
lliott, Bill R. Vandegriff.
es.

Comedy

wall High School drama
sent a three act comedy,
e Girls," Monday at 7:30
hool auditorium.
evolves around three main
ying to catch a killer and
o discover the true culprit.
ncludes Korina Bushnell,
well, Lore Reagh, Sylidia
rri Teel, Tonja Kelley,
ano, and Johnna Winton.
kup is club sponsor and

is $1 adults and 50 cents for

ready to be primed and painted.

SUNDAY APRIL 28 1985

Unsolved, violent crimes haunt Ada

By DOROTHY HOGUE

The Norman man entered the convenience store about 9 a.m. that Saturday and stood at the checkout counter waiting for someone to come wait on him. He noticed school books were open behind the counter, a cigarette was burning in an ashtray, the cash drawer was open and empty. He looked around for the clerk but found no one. After a few minutes the police were remind local citizens of the murder of another young local woman but police say they seriously doubt there could be any connection between the Haraway case and that of Debbie Carter.

It has been almost 2½ years since Carter was brutally murdered in her apartment, located only a few blocks from the East Central University campus. Det. Dennis Smith said police have narrowed their focus to one

Precinct 25 — Arm
Arlington and Countr
Precinct 31 (2) —
Center, Fourth and O
Precinct 33 (2) — O
Church, 523 N. Oak.
Precinct 34 (2) —
825 W. 10th.
Precinct 42 — Ac
Room 101.
Precinct 43 (2) —
600 W. 17th.
Precinct 44 (2) —
Church, 15th and Ash
Precinct 45 — A
auditorium.

Willard
continu

44:14

The Innocent Man S1:E1 Debbie and Denice

Data collection

Feature normalization

- Some classifiers can’t easily handle features with very different scales, e.g.,
  - Revenue in CHF: 10,000,000
  - # of employees: 300
- Features with large values dominate the others, and the classifier tends to over-optimize for them
- Even single feature may span many orders of magnitude
  - e.g., city size (most cities small, some huge)
  - Too little resolution where data is dense, too much resolution where data is sparse

Logarithmic scaling

$$x_i' = \log(x_i)$$

- Consider order of magnitude, rather than direct value
- Good for heavy-tailed features (e.g., from power laws)

Min-max scaling

$$x_i' = (x_i - m_i)/(M_i - m_i)$$

where M_i and m_i are the max and min values of feature x_i respectively

The new feature x_i' lies in the interval [0,1]

Standardization

$$x_i' = (x_i - \mu_i)/\sigma_i$$

where μ_i is the mean value of feature x_i, and σ_i is the standard deviation

The new feature x_i' has mean 0 and standard deviation 1

Dangers of standardization and scaling

Standardization:

- Assume that the data has been generated by a Gaussian
- Uses mean and std → not meaningful for heavy-tailed data (may be mitigated by log-scaling)

Min-max scaling:

- If the data has outliers, they scale the typical values to a very small interval

Classification pipeline

Model selection: high level

Need to choose type of model

- k-NN?
- Decision trees?
- Random forest?
- Boosted decision trees?
- Logistic regression?
- Deep learning?
- …

Model selection: low level

Usually a classifier has some “hyperparameters” to be tuned

- Set of features to include
- Distance function (e.g., k-NN)
- Number of neighbors (e.g., k-NN)
- Number of trees (e.g., random forest)
- Decision threshold (e.g., logistic regression)
- Regularization parameter
- Learning rate (for gradient descent algorithms)

Loss function (more of them later!)

Categorical output

- e.g., 0-1 loss function, risk (= 1 minus accuracy):

$$J = \frac{1}{n}\sum_{i=1}^{n} \#(y \neq f(x_i))$$

Real-valued output

- e.g., squared error: $J = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))^2$

- e.g., absolute error: $J = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - f(x_i) \right|$

Model selection: on what data to evaluate?

What if you can’t afford a 3-way split because you have too little data?

→ Cross-validation! (p.t.o.)

Cross-validation

- Last lecture: leave-one-out cross-validation == *N*-fold cross-validation (where *N* is #data points)
- More efficient: *m*-fold cross-validation (in above picture: *m* = 5)
- Average performance over the *m* red portions → validation error
- Repeat procedure for all candidate models and pick the one with the lowest validation error

Model selection

Performance metrics for binary classification

For categorical binary classification, the usual metrics are based on the **confusion matrix,** which has 4 values:

- True Positives (positive examples classified as positive)
- True Negatives (negative examples classified as negative)
- False Positives (negative examples classified as positive)
- False Negatives (positive examples classified as negative)

		Class	
		Pos	Neg
Classified	Pos	TP	FP
	Neg	FN	TN

Accuracy

		Class	
		Pos	Neg
Classified	Pos	TP	FP
	Neg	FN	TN

$$A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$

Appropriate metric when

- classes are not skewed
- errors (FP, FN) have the same importance

Accuracy (skewed example)

Classifier 1		Class	
		Fraud	¬Fraud
Classified	Fraud	5	10
	¬Fraud	5	80

Accuracy = 85/100 = 85%

Always ¬Fraud		Class	
		Fraud	¬Fraud
Classified	Fraud	0	0
	¬Fraud	10	90

Accuracy = 90/100 = 90%

Question time

Which classifier is better?

- Classifier 1
- Classifier 2
- Both are equally good

Classifier 1		Class A	Class B
Classified	A	45	20
	B	5	30

100 data points

Classifier 2		Class A	Class B
Classified	A	40	10
	B	10	40

100 data points

POLLING TIME

- Scan QR code or go to https://web.speakup.info/room/join/66626

Question time

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

Which classifier is better?

- Classifier 1
- Classifier 2
- Both are equally good

Precision and recall

Precision: what fraction of positive predictions are actually positive?

$$P = \frac{TP}{TP + FP}$$

[source]

Recall: what fraction of actually positive examples did I recognize as such?

$$R = \frac{TP}{TP + FN}$$

Precision and recall

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

$P_1=45/65=0.69$

$R_1=45/50=0.9$

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

$P_2=40/50=0.8$

$R_2=40/50=0.8$

Everybody has cancer		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$P = 50/100 = 0.5$

$R = 50/50 = 1$

F-score

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Sometimes it's necessary to have a single metric to compare classifiers

F-score (or F1-score): harmonic mean of precision and recall

$$\text{F1} = 1 / (0.5 * (1/P + 1/R)) = 2 \cdot \frac{P \cdot R}{P + R}$$

Precision and recall can be **differently weighted,** if one is more important than the other

Precision and recall

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

$F_1=2*(0.69*0.9)/(0.69+0.9)$

$= 0.78$

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

$F_2=2*(0.8*0.8)/(0.8+0.8)$

$=0.8$

Everybody has cancer		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$F=2*(0.5*1)/(0.5+1)=0.66$

Precision/recall curve

Decreasing classification threshold

ROC curve

ROC = Receiver-Operating Characteristic (WTF?!)

Y-axis: true-positive rate = TP/(TP + FN), a.k.a. recall

X-axis: false-positive rate = FP/(FP + TN)

ROC AUC

ROC AUC is the “area under the curve” – a single number that captures the overall quality of the classifier. It should be between 0.5 (random classifier) and 1.0 (perfect).

Bias and variance

[lecture 7]

How to know where on the x-axis you are (without fiddling with model complexity)?

Bias and variance

Bias and variance can be assessed by comparing the error metric on the training set and the validation set => always **plot learning curves** (training set size vs. training/validation errors)

When more data helps

Fixed data set size
Varying model complexity

Fixed model complexity
Varying data set size

High bias

More data doesn't help

High variance

More data helps

For curious ADAventurers:
"Reconciling modern machine-learning practice and the classical bias–variance trade-off"
https://www.pnas.org/doi/abs/10.1073/pnas.1903070116

Classification pipeline

Model assessment

- Model assessment is the goal of estimating the performance of a fixed model (i.e., the best model found during model selection)
- Ideally under real-world conditions
- Use held-out test set that you've never seen during training

Useful reads

Machine Learning that Matters

Kiri L. Wagstaff KIRI.L.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

A Few Useful Things to Know about Machine Learning

Pedro Domingos
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350, U.S.A.
pedrod@cs.washington.edu

slides

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec8-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Crowdsourcing

Different types of workers

- Truthful
  - Expert
  - Normal
- Untruthful
  - Uniform spammer
  - Random spammer
  - Malicious spammer (a.k.a. a$s#*le)

Catching malicious spammers

- Insert obvious examples for which you already know the labels (“honeypot”)
  - Tell workers they won’t be paid if they don’t get those right
  - Filter out workers who don’t get them right
- Aggregate multiple answers
  - p.t.o.

Crowdsourcing

Answer aggregation problem

- Have each example labeled by several workers, aggregate:
  - e.g., majority vote (works if only a minority is malicious)
  - e.g., “peer prediction”, “prediction markets” (game theory: workers are paid more if they agree with others)

Crowd (workers)

Worker	Webpage	Credible
W_1	www.diet.com	C
W_2	www.diet.com	¬C
W_3	www.diet.com	C
…	…	…

Aggregation

www.diet.com	C

Recap

- Model selection:
  - Use training data in cross-validation
  - Need evaluation metric
    - Typically based on confusion matrix
    - e.g., accuracy, precision, recall

		Class	
		A	B
Classified	A	TP	FP
	B	FN	TN

Model selection

Need evaluation metric!

Fwd-selected features vs. performance

Training and testing with heaps of data

Data-efficient training and testing: Leave-one-out cross-validation

Data-efficient training and testing: k-fold cross validation

More data often beats better algorithms

EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Large Language Models in Machine Translation

Thorsten Brants **Ashok C. Popat** **Peng Xu** **Franz J. Och** **Jeffrey Dean**

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94303, USA
{brants,popat,xp,och,jeff}@google.com

Abstract

This paper reports on the benefits of large-scale statistical language modeling in machine translation. A distributed infrastructure is proposed which we use to train on up to 2 trillion tokens, resulting in language models having up to 300 billion n-grams. It is capable of providing smoothed probabilities for fast, single-pass decoding. We introduce a new smoothing method, dubbed *Stupid Backoff*, that is inexpensive to train on large data sets and approaches the quality of Kneser-Ney Smoothing as the amount of training data increases.

How might one build a language model that allows scaling to very large amounts of training data? (2) How much does translation performance improve as the size of the language model increases? (3) Is there a point of diminishing returns in performance as a function of language model size?

This paper proposes one possible answer to the first question, explores the second by providing learning curves in the context of a particular statistical machine translation system, and hints that the third may yet be some time in answering. In particular, it proposes a *distributed* language model training and deployment infrastructure, which allows direct and efficient integration into the hypothesis-search algorithm rather than a follow-on re-scoring phase.

Applied Data Analysis (CS401)

Lecture 9
Learning from data: Unsupervised learning
15 Nov 2023

Robert West

Announcements

- Project milestone P2 due on Fri 17 Nov 23:59 (see Ed)
  - Reminder: we won't answer questions asked in the final 24 hours before the deadline
- Homework H2 to be released on Fri 17 Nov
  - Due two weeks later, on Fri 1 Dec
- Friday's lab session:
  - Quiz 8
  - Exercises on unsupervised learning

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec9-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Machine learning

- **Supervised:** We are given input/output pairs (X, y) (a.k.a. "samples") that are related via a function $y = f(X)$. We would like to "learn" f, and evaluate it on new data.
  - Discrete y (class labels): **"classification"**
  - Continuous y: **"regression"** (e.g., linear regression)
- **Unsupervised:** Given only samples X of the data, we compute a function f such that $y = f(X)$ is a "simpler" representation.
  - Discrete y (cluster labels): **"clustering"**
  - Continuous y: **"dimensionality reduction"** (e.g., matrix factorization, unsupervised neural networks)

The clustering problem

- Given a **set of points**, with a notion of **distance** between points, **group the points** into some number of ***clusters***, such that
  - members of a cluster are close (i.e., similar) to each other
  - members of different clusters are far apart from each other
- **Usually:**
  - Points live in a high-dimensional space
  - Similarity is defined via a distance measure
    - Euclidean, cosine, Jaccard, edit distance, … (cf. lecture 7)

Characteristics of clustering methods

Quantitative: scalability (many samples), dimensionality (many features)

Qualitative: types of features (numerical, categorical, etc.), type of shapes (polyhedra, hyperplanes, manifolds, etc.)

Characteristics of clustering methods

Robustness: sensitivity to noise and outliers, sensitivity to the processing order

User interaction: incorporation of user constraints (e.g., number of clusters, max size of clusters), interpretability and usability

Example: clusters & outliers

A typical clustering example

Note: Above is 2D; real scenarios often much more high-dimensional, e.g., 10,000-dimensional for 100x100 images.

Some use cases for clustering

- Data exploration (especially for high-dimensional data, where visualization fails)
- Partitioning of data for more fine-grained subsequent analysis
- Marketing: building personas
- Supporting data labeling for supervised learning
- Supporting feature discretization for supervised learning (cf. lecture 8)
- Data compression (next slide)
- …

Clustering for condensation/compression

Here we don’t require that clusters extract meaningful structure, but that they give a coarse-grained version of the data.

Beware of “cluster bias”!

- Human beings conceptualize the world through categories represented as exemplars or prototypes

- We tend to see cluster structure whether it is there or not.
- Works well for dogs, but…

Cluster bias

Cluster bias

- **Clustering is used more than it should be**, because people assume an underlying domain has discrete classes in it
  - Especially true for characteristics of people, e.g., Myers-Briggs personality types like "ENTP".
- In reality the underlying data is often **continuous**.
- In such cases, continuous models (e.g., matrix factorization, "soft" clustering) tend to do better (cf. next slide)

Netflix

- Central portion: more of a continuum than discrete clusters
- Other methods (e.g., dimensionality reduction) may be more appropriate than discrete clustering models

Terminology

- **Hierarchical clustering:** clusters form a tree-shaped hierarchy. Can be computed bottom-up or top-down.
- **Flat clustering:** no inter-cluster structure

- **Hard clustering:** items assigned to a unique cluster
- **Soft clustering:** cluster membership is a probability distribution over all clusters

Clustering is a hard problem!

Why is it hard?

- Clustering in 2 dimensions looks easy
- Clustering small amounts of data looks easy
- And in these special cases, it actually is often easy, but...

- … many applications involve not 2, but hundreds or thousands of dimensions (and large amounts of data)
- High-dimensional spaces are different ("curse of dimensionality"): volume grows exponentially w/ #dims; space is sparsely populated; clusters become less tight

Clustering problem: galaxies

- A catalog of 2 billion “sky objects” represents objects by their radiation in 7 dimensions (frequency bands)
- Problem: Cluster into similar objects, e.g., galaxies, nearby stars, quasars, etc.
- Sloan Digital Sky Survey [link]

Clustering problem: movies

- **Intuitively: Movies divide into categories/genres, and customers prefer a few categories**
  - But what are categories really?
  - $\longrightarrow$ take a data-driven approach!

- Represent a movie by a set of customers who watched it ("collaborative filtering")
- **Idea:** Similar movies have similar sets of customers, and vice-versa

Clustering problem: movies

Space of all movies:

- Think of a space with one dimension for each customer
  - Values in a dimension may be 0 or 1 only
  - A movie is a point in this space $(x_1, x_2, \ldots, x_n)$, where $x_i = 1$ iff the i-th customer watched the movie
- For Amazon/Netflix, the dimensionality is in the millions
- **Task:** Find clusters of similar movies

Clustering problem: documents

Finding topics:

- Represent a document by a vector $(x_1, x_2, \ldots, x_n)$, where $x_i = 1$ iff the i-th word appears in the document (in any position)

- **Idea:** Documents with similar sets of words are about same topic

Cosine, Jaccard, Euclidean distances

- In both examples (movies, documents) we have a choice when we thinking of data points as sets of features (users, words):
  - **Sets as vectors:**
    - Measure similarity via **Euclidean distance**
    - Measure similarity via **cosine distance**
  - **Sets as sets:**
    - Measure similarity via Jaccard index

Overview: Methods of clustering

- **Hierarchical methods:**
  - **Agglomerative** (bottom-up):
    - Initially, each point is a cluster
    - Repeatedly combine the two "nearest" clusters into one
  - **Divisive** (top-down):
    - Start with one cluster and recursively split it

- **Flat methods (a.k.a. point-assignment methods):**
  - Maintain a set of clusters
  - Assign points to "nearest" cluster; recompute clusters; repeat

Agglomerative hierarchical clustering

- **Key operation:**
  Repeatedly combine two nearest clusters

- **Three important questions:**
  - **(1)** How to represent a cluster of more than one point?
  - **(2)** How to determine the “nearness” of clusters?
  - **(3)** When to stop combining clusters?

Agglomerative hierarchical clustering

- **Key operation: Repeatedly combine two nearest clusters**
- **(1) How to represent a cluster of many points?**
  - Euclidean case: represent a cluster via its **centroid** = average of points in cluster
  - What about non-Euclidean case?
- **(2) How to determine "nearness" of clusters?**
  - Euclidean case: distance between clusters = distance between centroids
  - What about non-Euclidean case?

Example: Hierarchical clustering

Data:
o … data point
x … centroid

Dendrogram

Commercial break

A WORLD-CLASS UNIVERSITY IN AZERBAIJAN

UNİVERSİTET ADA UNIVERSITY

ada.edu.az

Data science from A to Z

Non-Euclidean case: clustroids

- **(1) How to represent a cluster of many points?**
  clustroid = actual data point that is **"closest"** to the other points
- **Possible meanings of "closest":**
  - Smallest average distance to other points (a.k.a. **medoid**)
  - Smallest sum of squares of distances to other points
  - Smallest maximum distance to other points

Centroid is the avg. of all data points in the cluster. This means centroid is an "artificial" point.
Clustroid is an **existing** data point that is "closest" to all other points in the cluster.

Non-Euclidean case: cluster “nearness”

- **(2) How do you determine the “nearness” of clusters?**
  - **Approach 1:**

    Intercluster distance = minimum of the distances between any two points, one from each cluster; or average of distances; or distance between clustroids; etc.

  - **Approach 2:**

    Pick a notion of “**cohesion**” (“tightness”) of a cluster

    Then: nearness of clusters = cohesion of their *union*

Cohesion

- **Approach 2.1:** Use the **diameter** of the merged cluster = maximum distance between points in the merged cluster
- **Approach 2.2:** Use the **average distance** between points in the merged cluster

How many branching points are there in a dendrogram for a dataset with *N* data points?

POLLING TIME

- Scan QR code or go to https://web.speakup.info/room/join/66626

Implementation

- **Naïve implementation of hierarchical clustering:**
  - At each step, compute pairwise distances between all pairs of clusters, then merge
  - $O(N^3)$, where N is the number of data points
- **Careful implementation using priority queue** can reduce time to $O(N^2 \log N)$
  - Still too expensive for really big datasets

Overview: Methods of clustering

- **Hierarchical methods:**
  - **Agglomerative** (bot
    - Initially, each cluster
    - epeatedly e the two
    - est" rs into one
  - **Div** own):
    - Start one cluster and recursively split it

- **Flat methods (a.k.a. point-assignment me ods):**
  - Maintain a set of clusters
  - Assign points to "nearest" cluster; recompute clusters; repeat

NEXT

K-means

The gorilla among the point-assignment clustering algorithms

K-means clustering

- Goal: assign each data point to one of *k* clusters such that the total distance of points to their centroids is minimized

- Solved by a simple greedy algorithm (Lloyd's algorithm): Locally minimize the "distance" (usually squared Euclidean distance) from data points to their respective centroids:
  - **Find the closest cluster centroid** for each data point, and assign the point to that cluster.
  - **Recompute the cluster centroid** (the mean of data points in the cluster) for each cluster.

K-means clustering

K-means clustering

How long to iterate?

- For fixed number of iterations
- or until no change in assignments (guaranteed to happen)
- or until only small change in cluster “tightness” (sum of [squared] distances from points to centroids)

K-means initialization

We need to pick some points for the first round of the algorithm:

- **Random sample:** Pick a random subset of *k* points from the dataset.
- **K-Means++:** Iteratively construct a random sample with good spacing across the dataset.

Note: Finding an optimal k-means clustering is NP-hard. The above help avoid bad configurations.

K-means++ [link]

Start: Choose first cluster center at random from the data points

Iterate:

- For every remaining data point x, compute the distance $D(x)$ from x to the closest previously selected cluster center.
- Choose a remaining point x randomly with probability proportional to $D(x)^2$, and make it a new cluster center.

Intuitively, this finds a sample of widely-spaced points from dense regions of the data space, avoiding "collapsing" of the clustering into a few internal centers.

K-means properties

- Greedy algorithm with random initialization – **solution may be suboptimal** & vary significantly with different initial points
- Very simple convergence proofs
- **Performance is O(*nk*) per iteration** — not bad, and can be heuristically improved
  n = number of points in the dataset, *k* = number clusters
- Many variants, e.g.
  - Fixed-size clusters
  - Soft clustering
- Works well for data condensation/compression

K-means drawbacks

Often terminates at a **local but non-global optimum** (mitigated by smart initialization such k-means++, or by re-running multiple times with different initializations)

Requires the notion of a **mean**

Requires specifying ***k*** (number of clusters) in advance

Doesn't handle **noisy data and outliers** well

Clusters **only** have **convex shapes**

How to choose *k*?

For k = 1, 2, 3, …

Run k-means with k clusters

For each data point i, compute "silhouette width" $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

S = average of $s(i)$ over all i

Plot S against k

Pick k for which S is greatest

$b(i)$: avg. distance to points in closest other cluster

$a(i)$: avg. distance to points in own cluster

DBSCAN

- "**D**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise"
- Motivation: centroid-based clustering methods like k-means favor clusters that are spherical, and have great difficulty with anything else

- But with real data we often have:

DBSCAN

- DBSCAN performs density-based clustering, and follows the shape of dense neighborhoods of points.

- **Def.: core points** have at least *minPts* neighbors in a sphere of diameter ε around them.
- The red points here are core points with at least *minPts* = 3 neighbors in an ε-sphere around them.

DBSCAN

- More **definitions (!):**
  - Core points can **directly reach** neighbors in their ε-sphere
  - From non-core points, no other points can be reached
  - Point q is **density-reachable** from p if there is a series of points $p = p_1, \ldots, p_n = q$ such that p_{i+1} is directly reachable from p_i
  - All points not density-reachable from any other points are **outliers/noise**

DBSCAN clusters

- Even more **definitions:**
  - Points *p*, *q* are **density-connected** if there is a point *o* such that both *p* and *q* are density-reachable from *o*.
  - A **cluster** is a set of points that are **mutually density-connected**.
  - That is, if a point is density-reachable from a cluster point, it is part of the cluster as well.
  - In the above figure, red points are mutually density-reachable; B and C are density-connected; N is an outlier.

DBSCAN algorithm

```
DBSCAN(DB, dist, eps, minPts) {
   C = 0                                                  /* Cluster counter */
   for each point P in database DB {
      if label(P) ≠ undefined then continue               /* Previously processed in inner loop */
      Neighbors N = RangeQuery(DB, dist, P, eps)          /* Find neighbors */
      if |N| < minPts then {                              /* Density check */
         label(P) = Noise                                 /* Label as Noise */
         continue
      }
      C = C + 1                                           /* next cluster label */
      label(P) = C                                        /* Label initial point */
      Seed set S = N \ {P}                                /* Neighbors to expand */
      for each point Q in S {                             /* Process every seed point */
         if label(Q) = Noise then label(Q) = C            /* Change Noise to border point */
         if label(Q) ≠ undefined then continue            /* Previously processed */
         label(Q) = C                                     /* Label neighbor */
         Neighbors N = RangeQuery(DB, dist, Q, eps)       /* Find neighbors */
         if |N| ≥ minPts then {                           /* Density check */
            S = S ∪ N                                     /* Add new neighbors to seed set */
         }
      }
   }
}
```

DBSCAN performance

- DBSCAN uses all-pairs point distances, but using an efficient indexing structure, each RangeQuery (for finding neighbors within ε-sphere) takes only O(log *n*) time
- The algorithm overall can be made to run in **O(*n* log *n*)**
- Fast neighbor search becomes progressively harder (higher constants) in higher dimensions, due to the

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec9-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

What are the most important statistical ideas of the past 50 years?*

Andrew Gelman† and Aki Vehtari‡

3 June 2021

Abstract

We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

https://arxiv.org/abs/2012.00174

Applied Data Analysis (CS401)

ada

Lectures 10+11
Handling text data
22 Nov 2023
+ 29 Nov 2023

EPFL

Robert West

dlab

Looking back at ADA so far…

What are the most important statistical ideas of the past 50 years?*

Andrew Gelman† and Aki Vehtari‡

3 June 2021

Abstract

We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

https://arxiv.org/abs/2012.00174

Announcements

- To all ADAmericans: Happy Thanksgiving!
- Milestone P2 being graded; feedback to be released next week
- Homework H2 due on Fri 1 Dec
- Friday's lab session:
  - Quiz 9
  - Homework H2 office hours (in person in BCH 2201, **not** on Zoom!)
  - Alumnus Niccolò Stefanini (ML scientist @ Expedia) will give "report from the trenches" (a.k.a. the real world of data science)

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec10-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Textual data

- Much modern data is unstructured text
  - Web
  - Social media
  - News
  - Several ADA project datasets…
- Frequently, “clean” datasets can be derived from “dirty” textual data
  - e.g., search queries are short texts; Google Trends time series for concepts (e.g., Q36611 *Gorilla*) are obtained by aggregating all search queries referring to the concept (e.g., “gorilla”, “big black Rwandan apes”, “are gorillas humans?”)

Nov 2022: the dawn of a new era

ChatGPT 3.5

EP **You**

Which of the following Google search queries relates to the concept "gorilla" (Wikidata knowledge base ID Q36611)?
- Arctic Monkeys
- gorilla
- big black Rwandan apes
- pistachio ice-cream
- are gorillas humans?
- what kind of animal did Jane Goodall work with?

Your answer must consist of a list of gorilla-related search queries, one per line. Add no other text.

< 3 / 3 >

ChatGPT

gorilla

are gorillas humans?

what kind of animal did Jane Goodall work with?

Outline

- 4 typical tasks on text data:
  - Document retrieval
  - Document classification
  - Sentiment analysis
  - Topic detection
- How to phrase these tasks as machine learning problems
- How to preprocess text so it can be fed to ML algorithms
- Next lecture: pointers to miscellaneous more advanced topics
- ADA spirit: show you what’s there; give you basic feel
  - For more, take classes on NLP and information retrieval

Typical task 1: document retrieval

- Given:
  - Document collection (a.k.a. corpus)
  - Query document (can be short query string)
- Task:
  - Rank all docs in collection by similarity to query
- An old problem (e.g., libraries)
- Document retrieval is the core task solved by Web search engines (“10 blue links”)

Document retrieval

- Straightforward approach: neighbor search (as in kNN)
- Define a distance function between documents
- Given query *q*, find the *k* docs with smallest distance to *q*
- *k* = 10, docs sorted by distance, blue links, ads → G
- The hard part: craft/learn a distance function (and scale it to the Web…)

Typical task 2: document classification

- Given:
  - Document *d*
  - Set of classes (e.g., topics: news, sports, tech, music, romance)
- Task:
  - Decide to which one of the classes document *d* belongs
- Example scenario:
  - Find gangster movies in CMU Movie Summary Corpus

Document classification

- Supervised learning
- Obtain a large collection of documents
- Label each doc with the class it belongs to
- Represent docs as feature vectors
- Train a supervised classifier based on the labeled docs:
  - e.g., kNN, logistic regression, decision tree, random forest, boosted decision trees, neural network, …

Typical task 3: sentiment analysis

- Given:
  - Document *d* (e.g., product review)
- Task:
  - “Sentiment” score capturing how positive/negative *d* is
- Example scenarios:
  - Infer what people think about a product from text only (i.e., without explicitly given ratings)
  - Historical opinion analysis; e.g., how has people’s attitude toward certain politicians changed over time?

Sentiment analysis

- Supervised learning
  - Regression
  - Classification
- Same setup as for document classification:
  - Label a training set with ground-truth sentiment scores
  - Represent documents as feature vectors
  - Train supervised model: kNN, linear/logistic regression, …

Typical task 4: topic detection

- Given:
  - Unlabeled document collection
- Task:
  - Determine a set of prevalent topics in the docs
  - Determine for each document to which topics it belongs
- Example scenario:
  - Detection of trending topics in social media (e.g., Twitter)
  - Detection of distinct viewpoints on a political subject
  - Exploratory analysis of a large doc collection

Topic detection

- Clustering
- Represent documents as feature vectors
- Run hierarchical or point-assignment clustering algorithm
  - Hierarchical: agglomerative or divisive
  - Point-assignment: e.g., k-means, DBSCAN
- Alternative: matrix factorization (cf. next lecture)

Feature vectors

- Nearly all ML methods work with feature vectors
  - E.g., previous slides: document retrieval; document classification; sentiment analysis; topic detection
- Text is not immediately a feature vector
  - Variable length
  - Even for fixed length (e.g., tweet…): Positions don’t correspond to meaningful features

Feature vectors

- Need to transform arbitrarily long string to fixed-length vector
  - Traditional and vetted: bag of words
  - More recent: *learn* a mapping from strings to vectors (buzzword: “text embedding”)

Bag of words

Tom Mitchell (CMU)

- Bag == multiset
  - “multi-”: keep multiplicity of words
  - “-set”: don’t keep order of words
  - E.g., document “what you see is what you get”
    → bag of words {get:1, is:1, see:1, what:2, you:2}
- To have fixed-length representation for all documents:
  - Vector with one entry for each unique word in vocabulary
  - Bag-of-word vectors are very high-dimensional (typically 1e5 or 1e6) and very sparse
  - E.g., above: [0…0 1 0…0 1 0…0 1 0…0 2 0…0 2 0…0]

An extra reason for sparsity: Zipf's law

A famous power law

The probability of observing a word scales inversely with its frequency rank:

$p(w_i) \propto 1/i$ (where w_i is the i-th most frequent word)

POLLING TIME

- Scan QR code or go to https://web.speakup.info/room/join/66626

Bag-of-words matrix

- Combine document vectors as rows in a matrix
  - One row per doc
  - One column per word in vocabulary
- This matrix is huge!
  - E.g., Wikipedia: 6M docs, 2M words → 12 trillion entries
- Solution: use a sparse matrix format
  - Triples: (doc_idx, word_idx, count)
  - E.g., Wikipedia, assuming 2000 words per article on avg.: 12 billion non-zero entries (fits in memory)
- With matrix representation, you're ready to use any ML model

… or are you really?

- In theory, yes
- In practice: “garbage in, garbage out”
- Be careful when mapping raw text to bag-of-words matrix!
  - Character encoding
  - Language identification
  - Tokenization
  - Stopword removal
  - Word normalization
- Tweaking the matrix a bit can lead to much better performance
  - Reweight/normalize rows and/or columns of matrix

Bag of tricks for bags of words

Character encoding

- Mapping from (abstract) characters to bytes
- Old school: ASCII, Latin-1
- New school: Unicode (e.g., UTF-8, UTF-16, UTF-32)
- E.g., W → 0x57 (one byte)
- Reading text from file:
  - Need to read with encoding that was used to write file
  - Especially important for non-English text: à, ê, ü, ß, …
- Writing to file: Always use UTF-8 or UTF-16; hard-code the output format!

```
file = codecs.open("temp", "w", "utf-8")
file.write(codecs.BOM_UTF8)
file.close()
```

- Otherwise, your future self will be very angry at you (example in speaker notes 👇)

Language identification

- Typically, you're interested in text from a single language
- Increasingly, content is multilingual (e.g., Twitter, Wikipedia)
- Ideally, language code is specified (e.g., headers in HTML; JSON field in Twitter API results)
- But not always…
- There are good libraries (e.g., Python, Java)
  - Most commonly based on letter trigrams (e.g., "eau", "ghi", "ijs", "sch", "eiß", "ção")
  - Much harder if you messed up character encoding…

Commercial break

WIKIPEDIA
The Free Encyclopedia

Create account Log in

Amigos dos Amigos

4 languages

Article Talk dit View history Tools

Looking for friends?

From Wikipedia, the free encyclopedia

Amigos dos Amigos (**ADA**, *Friends of Friends*) is a criminal organization that operates in the Brazilian city of Rio de Janeiro. It was started up in 1998[1] when a member of Comando Vermelho was expelled from the organization for ordering the murder of another member. The gang's main rivals are Comando Vermelho and Terceiro Comando Puro. ADA controls many drug selling points in the North and West zones.

Between 2004 and 2017, ADA controlled Rocinha, the largest favela in Rio de Janeiro,[2] along with many other smaller favelas. With the assassination of the gang leader Bem-Te-Vi in 2005 by police, there was a renewed wave of violence as gangs fought for control over favelas previously controlled by ADA.

ADA are thought to wield significant social power in the communities they control, winning support through handouts, throwing parties, and providing some services, while their rivals, the Red Command, imposes itself more through violence.

Amigos dos Amigos

Founding location	Rio de Janeiro, Brazil
Years active	1998-present
Territory	Rocinha, various neighborhoods of Rio
Ethnicity	Brazilians
Membership	300[1]
Activities	Murder, drug trafficking, extortion, prostitution, illegal gambling, human trafficking, kidnapping
Rivals	Comando Vermelho, Terceiro Comando Puro,

Tokenization

- Maps character string into sequence of tokens (≈ words)
- E.g., “Hello! How are you?” → Hello_!_How_are_you_?
- Tempting to do this yourself by splitting at whitespaces and punctuation
- But many corner cases:
  - “Hello, Mr. President! How are you?! :-)”
    → Hello_,_Mr._President_!_How_are_you_?!_:-)
- Don’t do it yourself, use libraries instead; e.g.,
  - Python: spaCy, nltk; Java: Stanford CoreNLP
  - Rule-based, deterministic, fast

Tokenization

- Optimal tokenizer different for different languages (e.g., Swedish “Sankt Peter” → “S:t Peter”), but English tokenizer often good enough
- Tokenization relatively straightforward in English
- Hard in, e.g., Chinese: no whitespace between words
- Compound words, e.g. in German:
  - Advanced models can split “Donaudampfschifffahrtskapitän” into “Donau dampf schiff fahrts kapitän”
  - But what to keep together…? “Schiff fahrt” or “Schifffahrt”?

Stopword removal

- Very frequent, “small” words carry little information for most tasks and can “drown out” information contained in real content words
- E.g., “a”, “the”, “is”, “you”, “I”, punctuation marks
- Many stopword lists online, but be careful!
  - Different tasks require removing different stopwords
  - Good heuristic: remove words appearing in at least *p*% of all documents (but what should *p* be…?)
  - Sometimes stopword removal hurts!
    - Author identification, psychological modeling; punctuation can be useful as well: e.g., “!!!”, “:-)”

Don’t throw out the baby with the bathwater!

Word normalization: casefolding

- E.g., “I love yams. Yams are yummy.”
- Should “yams” and “Yams” really be different features?
- Simple solution: make everything lower-case (“casefolding”)
- But then: “I’d rather have an apple than an Apple.”
- Hand-code exceptions?
- In practice (especially when dataset is large), typically best to **not** do casefolding
- But when dataset is small, might help because less sparsity

Word normalization: Stemming

- Map different forms of same word to same, normalized form, by stripping affixes
- E.g., “walking”, “walks”, “walked” → “walk”
  “business”, “busy” → “busi”
- Typically done in hacky, heuristic way (e.g., Porter stemmer)
- Pro: decreases sparsity in bag-of-words matrix
- Con: discards information
  - E.g., “business” vs. “busy”; “operating” (as in “operating system”)
- In English (esp. with big data) typically not done anymore
- Still very useful in morphologically richer languages (e.g., German, Finnish, Bantu languages)

Word normalization: Lemmatization

- Lemmatization == stemming++
- Map tokens to lexicon entries
- E.g., “U.S.A.”, “US” → “United States”
  “Grüße”, “Gruesse” → “Grüße”
  “You **lie** in the grass” vs. “You **lie** to me”
- Frequently omitted, as it requires complete lexicon and complex mapping rules
- Especially hard for non-English

Social media

A real tweet:
“ikr smh he asked fir yo last name so he can add u on fb lololol”

- Translation:
  - *“ikr”* means “I know, right?”
  - *“smh"* means “shake my head”
  - *“fb"* means “Facebook”
  - *“yo”* is being used as equivalent to *“your”*
  - *“fir”* is a misspelling or spelling variant of the preposition *for* (But who knows?!)
- Also common: repeating letters/syllables (“yeahhh”, “hahahaha”, “haha”)
- Good luck with traditional NLP tools…
- Need dedicated toolkits such as TweetNLP

Tokens vs. n-grams

- So far: bag-of-words matrix
  - Rows: documents
  - Columns: tokens (a.k.a. unigrams, or 1-grams)
- Frequently, longer sequences belong together
  - E.g., “United States”, “operating system”
- Brute-force approach: use $n > 1$
  - E.g., all bigrams ($n = 2$), all trigrams ($n = 3$)
  - Using all 5-grams can beat neural networks (Table 1 here)
  - Problem: combinatorial explosion

Tokens vs. n-grams

- Smarter:
  - Feature selection (“multi-word expressions”, “phrase extraction”)
  - Simple approach for bigrams: keep bigram if *mutual information* between constituent tokens is large
  - How to generalize to $n > 2$?
    - Frequent itemset/sequence mining
    - Wikipedia anchor texts
    - Compressive feature learning (link)

Postprocessing the BOW matrix

Inverse document frequency

- Not all words equally informative
- This is the reason for removing stopwords (“a”, “the”, “is”, …)
- Beyond discarding stopwords, want to give less weight to more common words
  - E.g., “per” vs. “perceptron”
- Standard way: **IDF = inverse document frequency**
  - docfreq(w): number of documents that contain word w
  - N: overall number of documents
  - $\text{idf}(w) = -\log(\text{docfreq}(w) / N) = \log(N) - \log(\text{docfreq}(w))$

Inverse document frequency

- idf(w) = –log(docfreq(w) / N)
- Interpretation: information content (in terms of #bits) of event "randomly drawing a document that contains w"
- Beyond this theoretical justification, IDF weighting has been shown to work well in practice

TF-IDF matrix

- tf(d, w): term frequency of word w in doc d
  - This is what the bag of words captures
  - E.g., document "what you see is what you get"
    $\longrightarrow$ bag of words {get:1, is:1, see:1, what:2, you:2}
- idf(w): inverse doc freq of w (computed on entire corpus)
- TF-IDF matrix:
  - Entry in row d and column w has value
    tf(d, w) * idf(w)
  - Amounts to multiplying column w with constant idf(w)

Row normalization of TF-IDF matrix

- Longer docs have more non-zero entries
- Interpreted as vectors, longer docs have longer vectors
- This may throw off ML algorithms
  - Long vectors far away from short vectors
  - Dot product: random vector has higher dot product with longer vector
- Fix: normalize doc vectors, i.e., rows of TF-IDF matrix
  - L2-normalization: all rows have Euclidean distance 1 from origin (all data points lie on a unit sphere)
  - L1-normalization: all rows sum to 1, i.e., can be interpreted as distribution
- How to know which one is better?

Column normalization

- IDF-scaling may be seen as column normalization
- Additionally, it may help to apply any of the normalization techniques we discussed in lecture 8 (“Applied ML”)
  - Min-max scale
  - Standardize: subtract mean; divide by standard deviation
- How to know which one (if any) to use?

Bag of tricks for bags of words

Stay tuned!

Next week: Part 2

Applied Data Analysis (CS401)

ada

Lectures 10+11
Handling text data
22 Nov 2023
+ 29 Nov 2023

EPFL

Robert West

dlab

Announcements

- Homework H2 due this Fri 1 Dec 23:59
  - Reminder: We won’t answer questions asked during final 24h
- Projects:
  - Milestone P2 grades have been released
  - Start working on Milestone P3 after Friday!
- Friday’s lab session:
  - Exercises on handling text

Recap

Reminder:
bag-of-words matrix

Revisiting the 4 typical tasks

- Document retrieval
- Document classification
- Sentiment analysis
- Topic detection

- TF-IDF matrix to the rescue
  - Entry for doc *d*, word *w*:
    tf(*d*, *w*) * idf(*w*)

Typical task 1: document retrieval

- Nearest-neighbor method in spirit of kNN
- Compare query doc *q* to all documents in the collection (i.e., rows of the TF-IDF matrix)
- Rank docs from collection in increasing order of distance
- Distance metrics
  - Typically cosine distance (= 1 – cosine similarity)
  - Recall: cosine similarity of *q* and *v* = $<q/|q|, v/|v|>$
  - If rows are L2-normalized, may simply take dot products $<q, v>$

Typical task 1: document retrieval

- This is just the most basic approach
- For efficiency
  - Start by filtering documents by presence of query terms (use efficient full-text index)
  - Hugely narrows down set of documents to be ranked
- Google et al. do much more…
  - Query-independent relevance: PageRank
  - Boost recent results
  - Personalization, contextualization
  - …

Typical task 2: document classification

- Use TF-IDF matrix as feature matrix for supervised methods (cf. lecture 7)
- Often more features (words) than documents
- What's the danger with this?
- High model capacity can lead to overfitting (high variance)
- Potential solutions:
  - Use more data (i.e., more labeled training docs)
  - Decrease model capacity:
    - Feature selection
    - Regularization (two slides from now)
    - Dimensionality reduction (a few slides from now)
  - Use ensemble methods such as random forests

Typical task 3: sentiment analysis

- When treated as classification:
  Ctrl-C Ctrl-V previous slide

- When treated as regression:
  Pretty much the same (most supervised methods work for both classification and regression)

Regularization

- E.g., linear regression:
  Find weight vector $\boldsymbol{\beta}$ that minimizes $\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$

  ($\mathbf{x}_i$: feature vector of i-th data point; y_i: label of i-th data point, e.g., sentiment [1–5 stars] expressed in document i)
- If one word j appears only in docs with sentiment 5, we can obtain very small training error on these docs by making β_j large enough
- But doesn't generalize to unseen test data!
- Remedy: penalize very large positive and very large negative weights:

$$\text{minimize} \quad \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$$

Regularization

Which curve resulted from adding a regularization term to the loss function?

POLLING TIME

- Scan QR code or go to https://web.speakup.info/room/join/66626

Typical task 4: topic detection

- Cluster rows of TF-IDF matrix (each row a data point)
- Manually inspect clusters and label them with descriptive names (e.g., “news”, “sports”, “romance”, “tech”, “politics”)
- In principle, may use k-means, k-medoids, etc.
- But can be difficult if dimensionality is large (#words $\gg$ #docs)
  - “Curse of dimensionality”
  - Many outliers

Typical task 4: topic detection

- Alternative approach: **matrix factorization**

- Assume docs and words have representation in (latent) "topic space"
- (IDF-weighted) word frequency modeled as dot product of doc's vectors and word's vectors in topic space
- #topics ≪ #words (→ "dimensionality reduction"): D*W → (D+W)*T
- Topics interpretable in doc space (*A*'s cols) and word space (*B*'s rows)

Typical task 4: topic detection

- Optimization problem:
  - Find *A*, *B* such that *AB* is as close to TF-IDF matrix as possible
  - That is, minimize $$\sum_{d=1}^{N} \sum_{w=1}^{M} (T_{dw} - A_d \; B_w)^2$$

    where *T* is TF-IDF matrix, A_d is *d*-th row of *A*, and B_w is *w*-th column of *B*
- This is called **latent semantic analysis (LSA)**

Typical task 4: topic detection

You already know how to efficiently compute this, from your linear algebra class: **singular-value decomposition** (SVD)

- $T = USV^T$
- Freebie: columns of U and V are orthonormal bases (yay!)
- S is diagonal (with values in decreasing order) and captures "importance" of topic (amount of variation in corpus w.r.t. topic)
- If you want k topics, keep only the first k columns of U and V, and the first k rows and columns of S
  $\longrightarrow U', S', V'$
- E.g., $A = U'$, $B = S'V'^T$ or $A = U'S'$, $B = V'^T$

Typical task 4: topic detection

- Recall potential problem with clustering and classification and regression: "curse of dimensionality"
- Matrix factorization via LSA solves these problems for you:
  - Use *A* instead of original TF-IDF matrix
  - That is, cluster (or learn to classify or regress) in topic space, rather than word space

- Topic representation from LSA is simply a vector, not a probability distribution over topics
- Probabilistic: LDA = Latent Dirichlet Allocation (p.t.o.)

Commercial break

ada

ça roule pour moi

LDA: probabilistic topic modeling

- **L**atent **D**irichlet **A**llocation (*not* Latent Discriminant Analysis!)
- Document := bag of words
- Topic := probability distribution over words
- Each document has a (latent) distribution over topics
- "Generative story" for generating a doc of length n:
  d := sample a topic distribution for the doc ($\leftarrow$ "Dirichlet")
  for $i = 1, \ldots, n$
    t := sample a topic from topic distribution d
    w := sample a word from topic t
    Add w to the bag of words of the doc to be generated

Topics
gene 0.04
dna 0.02
genetic 0.01
. . .
life 0.02
evolve 0.01
organism 0.01
. . .
brain 0.04
neuron 0.02
nerve 0.01
. . .
data 0.02
number 0.02
computer 0.01
. . .
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Haemophilus genome 1703 genes
Genes in common 333 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
255 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
SCIENCE • VOL. 272 • 24 MAY 1996

Topic inference in LDA

- LDA is unsupervised (topics come out “magically”)
- Input:
  - Docs represented as bags of words
  - Number *K* of topics
- Output:
  - *K* topics (distributions over words)
  - For each doc: distribution over *K* topics
- How is this done?
  - Find distributions (i.e., topics, docs) that maximize the likelihood of the observed documents (maximum likelihood)

Question:

- "Which of these word pairs is more closely related?"
  - (car, bus)
  - (car, astronaut)
- How to quantify this?
- Detour:
  - How to quantify closeness of two docs?
  - E.g., cosine of **rows** of TF-IDF matrix
- Retour:
  - How to quantify closeness of two words?
  - E.g., cosine of **cols** of TF-IDF matrix

Sparsity in TF-IDF matrix

- Two docs (i.e., rows of TF-IDF matrix)
  - “Do you love men?”
  - “Adorest thou the likes of Adam?”
  - Cosine of row vectors of TF-IDF matrix == 0
- Same problem when comparing two words (i.e., cols of matrix)
- Solution:
  - Move from sparse to dense vectors
  - But how?
  - Latent semantic analysis (LSA)!
    - Use columns of *B* as dense vectors representing words

“Word vectors”

- Columns of TF-IDF matrix (sparse)
  or of word-by-topic matrix B (dense)

- Problem:
  - Entire doc treated as one bag of words
  - All information about word proximity, syntax, etc., is lost
- Solution:
  - Instead of full docs, consider local contexts: windows of *L* (e.g., 3) consecutive words to left and right of the target word
  - Rows of matrix: not docs, but contexts

"Word vectors"

- What to use as entries of word/context matrix?
- Straightforward: same as TF-IDF, but with contexts as "pseudo-docs": $M[c,w]$ = TF-IDF(c,w)
- May use any other measures of statistical association
- E.g., pointwise mutual information (PMI):
  $M[c,w] = \text{PMI}(c,w) = \log \frac{\Pr(c,w)}{\Pr(c)\ \Pr(w)}$
  "How much more likely are *c* and *w* to occur together than if they were independent?"
- word2vec: factor PMI matrix and use columns of *B* as word vectors

Beyond bags of words

From words to texts

- Word vectors represent, well, words
- How to represent larger units, such as sentences, paragraphs, docs?
- Typical approach: take sum/average of word vectors
- Note: this is roughly also what bags of words are (when using "one-hot" encoding for words, i.e., vector with exactly one 1, rest 0)
- More recently: **learn** vectors for longer units
  - Cr5, sent2vec
  - Convolutional neural networks
  - Recurrent neural networks, e.g., LSTM, ELMo
  - Transformer-based models, e.g., BERT (next slides), GPT-*

Contextualized word vectors

- Motivating example:
  - “My ass likes carrots”
  - vs. “He’s such an ass”

- Classic word vectors (e.g., word2vec) cannot distinguish these two cases; same vector used for both instances of “ass”

- Solution: contextualized word vectors
  - E.g., BERT

BERT in a nutshell

- Introduced in 2018 by Google Research

NLP pipeline

- Tokenization
- Sentence splitting
- Part-of-speech (POS) tagging
- Named-entity recognition (NER)
- Coreference resolution
- Parsing
  - Shallow parsing (a.k.a. chunking)
  - Constituency parsing
  - Dependency parsing

NLP pipeline

- Implemented by Stanford CoreNLP, nltk, spaCy, etc.
- Sequential model
  - Fixed order of steps
  - Early errors will propagate downstream
  - Fixed order not optimal for all cases (e.g., syntax usually done before semantics, but semantics might be useful for inferring syntax)
- Hence, current research: learn all tasks jointly (early example)
- To learn how all this magic is implemented
  - Take CS-431 (Intro to NLP), CS-552 (Modern NLP)

Today's trend: generative language models

- E.g., OpenAI GPT-*, ChatGPT
- Input: text
- Output: text
- Many NLP tasks can be formulated in this framework, by "prompting" the language model with the right input

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec10-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Applied Data Analysis (CS401)

ada

Lecture 12
Handling network data
6 Dec 2023

EPFL

Robert West

dlab

Announcements

- Homework H2 is being graded
  - Feedback to be released next week
- Final project milestone P3 due on Fri 22 Dec 2023
- Friday's lab session:
  - Project office hour (on Zoom)
    - Second-to-last project office hour before due date
    - To secure your team's personal time slot, follow protocol described in this Ed post
  - Exercises on handling networks (in BCH 2201)
    - In parallel to office hour

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec12-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Beyond flat tables

People

id	name	age
1	Bob	36
2	Willy	32
…	…	…

Marriages

husband_id	wife_id
1	34
2	5
2	87
…	…

Examples

Social networks

Economic networks

Communication graphs

Information networks:
Web & citations

Internet

Networks of neurons

Examples

Patient networks

Hierarchies of cell systems

Disease pathways

Genetic interaction networks

Gene co-expression networks

Cell-cell similarity networks

Networks as graphs

- **Network:** a real-world system of dependent variables, e.g.,
  - WWW is a network of hyperlinked documents
  - Society is a network of individuals linked by family, friendship, professional ties
  - Metabolic network is sum of all chemical reactions in a cell
- **Graph:** mathematical abstraction for describing networks
- In practice, “network” and “graph” are used interchangeably
- You can make a graph out of almost anything (e.g., connect all people whose name starts with the same letter), so must ask: Does this graph correspond to a meaningful network?

Today’s lecture

- **Part 1:** Types of graphs
- **Part 2:** Representing graphs on computers
- **Part 3:** Properties of real-world networks
- **Part 4:** Measuring node importance

Part 1:
Types of graphs

Most basic type: undirected graphs

- Entities:
  nodes/vertices V
- Relationships/interactions:
  edges/links E
- Entire system:
  graph $G = (V, E)$

Node degree

Node degree, k_i: the number of edges adjacent to node $\boldsymbol{i}$

$$k_A = 4$$

Types of graphs: undirected vs. directed

Undirected

- **Links:** undirected (symmetrical, reciprocal)

- **Examples:**
  - Collaborations
  - Friendship on Facebook

Directed

- **Links:** directed (arcs)

- **Examples:**
  - Phone calls
  - Following on Twitter

Average node degree

POLLING TIME

Scan QR code or go to
https://web.speakup.info/room/join/66626

Undirected

Node degree, k_i: the number of edges adjacent to node i

$k_A = 4$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i =$

Average node degree

Undirected

Directed

Source: Node with $k^{in} = 0$
Sink: Node with $k^{out} = 0$

Node degree, k_i: the number of edges adjacent to node $\boldsymbol{i}$

$$k_A = 4$$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N}$

In directed networks we define an **in-degree** and **out-degree.**

The (total) degree of a node is the sum of in- and out-degrees.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\bar{k} =$$

Types of graphs: weighted

- **Unweighted**
  (undirected)

Examples: Friendship, Hyperlink

- **Weighted**
  (undirected)

Examples: Collaboration, Internet, Roads

Types of graphs: bipartite

- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V

- **Examples:**
  - Authors-to-Papers (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
  - Recipes-to-Ingredients (they contain)
- **"Folded" networks (a.k.a. "projections"):**
  - Author collaboration networks
  - Movie co-rating networks

Projections of bipartite graph

Example: flavor networks

Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási
Flavor network and the principles of food pairing , Scientific Reports 196, (2011).

[paper]

Part 2:
Representing graphs on computers

Representing graphs on computers: adjacency matrix

$A_{ij} = 1$ if there is a link from node ***i*** to node ***j***

$A_{ij} = 0$ otherwise

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

Adjacency matrix usually sparse

Representing graphs on computers: edge list

- **Represent graph as a set of edges:**
  - (2, 3)
  - (2, 4)
  - (3, 2)
  - (3, 4)
  - (4, 5)
  - (5, 2)
  - (5, 1)

Representing graphs on computers: adjacency list

- **Adjacency list:**
  - Easier to work with if network is
    - Large
    - Sparse
  - Allows us to quickly retrieve all neighbors of a given node
    - 1:
    - 2: 3, 4
    - 3: 2, 4
    - 4: 5
    - 5: 1, 2

Graph processing libraries

- Good overview + benchmark available here
- NetworkX
  - Written in Python
  - Popular but slow
  - Ok when your graph is small
  - **This Friday's lab session: intro to NetworkX**
- NetworkKit, SNAP, iGraph, graph-tool
  - Written in C++
  - Much faster than NetworkX
  - Libraries also available in other languages (incl. Python)
  - Consider these if your graph is large

Part 3:
Properties of real-world networks

Properties of real-world networks

- Real networks are different from arbitrary graphs
- Real networks tend to share certain properties
- Remarkable, given the diversity of networks
  - Information networks (e.g., Web graph, knowledge graphs)
  - Social networks (e.g., Facebook, sexual networks)
  - Biological networks (e.g., protein–protein interaction)
  - …

Properties of real-world networks: sparsity

- Every node connected to only small fraction of all other nodes
- i.e., $k_i \ll N$
- Often bounded by a constant
  - e.g., social networks: Dunbar's number (cognitive limit to the number of people with whom one can maintain stable social relationships; allegedly 150)

Infomercial break

Today!
And happy ADAvent
to all of you!

Random wiki fact:

Gemiler Island

Article Talk

From Wikipedia, the free encyclopedia

Gemiler Island (Turkish: *Gemile Adası or Gemiler Adası*, Greek: Γκεμιλέρ) is an island located off the coast of Turkey near the city of Fethiye. On the island are the remains of several churches built between the fourth and sixth centuries AD, along with a variety of associated buildings. Archaeologists believe it was the location of the original tomb of Saint Nicholas. The original Turkish name is Gemile from the Greek word καμήλα (kamila) meaning camel, so called because of its geographical shape (see Fethiye Gemile Island Archaeological Site).

Properties of real-world networks: degree distribution

- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

  N_k = # nodes with degree k
- Normalized histogram:

  $P(k) = N_k / N$ ➔ **plot**

Properties of real-world networks: degree distribution

Properties of real-world networks: degree distribution

Linear axes

Logarithmic axes

Properties of real-world networks: triadic closure

- "A friend of my friend is my friend"
- Measured via clustering coefficient C_i of node i:
  C_i = (#edges among neighbors of i) / (#potential edges among neighbors of i)
- #potential edges among neighbors of i in undirected graph: $k_i (k_i - 1) / 2$
- $C_A = 1 / (4*3/2) = 1/6$
- In real networks, nodes have large clustering coefficients

Properties of real-world networks: community structure

- Triadic closure makes real networks cluster into locally dense "communities"
- Communities connected via "weak ties"
- "The strength of weak ties" (Granovetter 1973)
- Weak ties fill "structural holes"

Properties of real-world networks: community structure

- In real life, communities are often not disjoint, but overlapping:

[George Costanza's ideal world; unfortunately not realistic]

Extra homework: Start watching *Seinfeld*,

so you understand the [“Worlds Collide” theory]

Properties of real-world networks: average shortest-path length

- **What is the typical shortest path length between any two people?**
  - **Experiment on the global friendship network**
    - Can't measure, need to probe explicitly
- **Small-world experiment** [Milgram '67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter to a stock-broker in Boston by passing it through friends
- **How many steps did it take?**

[Movie]

Properties of real-world networks: average shortest-path length

- **64 chains completed:**
  (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus **"6 degrees of separation"**
- **Further observations:**
  - People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
  - People from the Boston area have even closer paths: 4.4

Milgram's small world experiment

Properties of real-world networks: navigability

- For decades, people focused on the fact that short paths exist in social networks
- But this is true even in random graphs (e.g., Erdős-Rényi model)
- The truly amazing fact is not that short paths exist, but that they are discoverable via greedy decentralized routing (as in Milgram's experiment)
- Intrigued? Read Jon Kleinberg: "The Small-World Problem: An Algorithmic Perspective"

Don’t believe me?

- Play a game and see for yourself how well you can navigate an a-priori unknown network:
  - Wikispeedia.net

Properties of real-world networks: homophily/heterophily

- "Birds of a feather flock together"
- Especially in social networks
- Big confound and cause of debate: Influence vs. homophily
- E.g., obese people's friends are more likely to also be obese
  - Influence: I copy eating behavior of those around me [argument for influence]
  - Homophily: people with similar eating behavior prone to become friends [argument for homophily]

Properties of real-world networks: summary

Real-world networks (across many types)

- are sparsely connected,
- but some nodes are much more connected than most others (i.e., skewed degree distribution);
- form locally dense clusters via triadic closure,
- which leads to community structure;
- have short paths between random node pairs (partly due to “hubs” [skewed degree distribution!]),
- and the short paths are easily discoverable.

Part 4:
Measuring node importance

How to measure “importance” of a node?

- Formalized via *centrality measures*
- Map each node i to a scalar value $C(i)$ capturing its importance in the overall network

Degree centrality

- Simplest centrality measure
- Many neighbors → important node
  - $C(i)$ = number of neighbors of i
- Very brittle, easy to “rig”
  - E.g., Twitter: scam account, followed by 100,000 other scam accounts

Closeness centrality

- Farness(x) = total distance to x from other nodes
- $C(x)$ = 1 / Farness(x)
  - Reciprocal to turn farness into closeness
  - Only defined for connected graphs (otherwise d infinite)
- Under closeness centrality, nodes that are easy to reach from anywhere in the network are considered important
- Variant: harmonic centrality: switch sum and reciprocal
  - $C(x)$ = total reciprocal distance of x to other nodes
  - Defined even for disconnected graphs (define $1/d$ of disconnected nodes as 0)

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

$$C(x) = \sum_{y \neq x} \frac{1}{d(y,x)}$$

Betweenness centrality

- $C(i)$ = average fraction of all shortest paths in the network that pass through node i
- Computation of betweenness centrality of node i:
  - For each pair of vertices (s, t):
    - Find all shortest paths from s to t
    - Compute the fraction of these shortest paths that pass through i
  - Average this fraction over all pairs of vertices (s, t)
- Expensive to compute

Katz centrality

- Generalization of degree centrality
- Degree centrality counts only number of direct neighbors (i.e., neighbors at distance 1)
- Katz centrality also counts neighbors at distances 2, 3, …
- More precisely, number of paths from other nodes to *i*, giving less weight to larger distances:

$$C(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ji}$$

k-th power of adjacency matrix *A* contains number of length-*k* paths for each node pair

- More robust than degree centrality

PageRank centrality

- Recursive definition: my centrality $C(i) =: x_i$ is high if I receive inlinks from many other central nodes:

$$x_i = \sum_j a_{ji} \frac{x_j}{L(j)}$$

$$L(j) = \sum_i a_{ji}$$

a_{ji}: entry (j, i) of adjacency matrix A (1 if j links to i, else 0)
$L(j)$: out-degree of j

- Some extra tweaks to make it work with any graph (e.g., disconnected)

PageRank centrality

$$x_i = \sum_j a_{ji} \frac{x_j}{L(j)}$$

- Matrix notation: $x = M\,x$
  (where M is computed from adjacency matrix A)
- Do you recognize this?
  - x is eigenvector of M with eigenvalue 1 $\rightarrow$ we're in linear-algebra land (home sweet home)
  - x is the steady state the Markov chain induced by the network: x_i is fraction of time a random walker will have spent in node i, after a very long ($\rightarrow\infty$) random walk

PageRank centrality

- The technology that made Google huge
  - “Page” in PageRank for Larry Page
  - MapReduce (next lecture!) was invented to compute PageRank on full Google Web crawl
  - “The $25,000,000,000 eigenvector” [link]
- Bottom line:
  Pay attention in your linear algebra class

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec12-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Credits

- Some slides borrowed from CS224W class (Jure Leskovec, Stanford)

Applied Data Analysis (CS401)

Lecture 13
Scaling to massive data
13 Dec 2023

Robert West

Announcements

- Homework H2
  - Feedback to be released later this week
  - Reminder (see Ed): Please participate in ML4Ed study by Fri 14:00
- Project milestone P3 due next week (Fri 22 Dec)
- Friday's lab session:
  - Last lab session! → Last quiz (on lecture 12)
  - Project office hour (same sign-up protocol as last week)
  - Exercises on Spark (useful for your future projects, your job, your love life)
- **Course eval is available on IS-Academia!**
  - Note: different from the eval from a few weeks ago!

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec13-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo?
- …

So far in this class...

- We made one big assumption:
  - All data fits on a single machine
  - Even more, all data fits into memory on a single machine (Pandas)
- Realistic assumption for **prototyping,** but frequently not for production code

The big-data problem

Data is growing faster than computation speed

+ Growing data sources
  (e.g, Web, mobile, sensors, …)

+ Cheap hard-disk storage

– Stalling CPU speeds

– RAM bottlenecks

Examples

These numbers (anno domini 2016) are outdated (too small)!

Facebook's daily logs: 60 TB

1000 Genomes project: 200 TB

Google Web index: 100+ PB

Cost of 1 TB of disk: $50

Time to read 1 TB from disk: 3 hours (100 MB/s)

The big-data problem

A single machine can no longer store, let alone process, all the data

The only solution is to **distribute** over a large cluster of machines

But how much data should you get?

Of course, “it depends”, but for many applications the answer is:

As much as you can get

Big data about people (text, Web, social media) tends to follow heavy-tailed distributions
(e.g., power laws)
Example: Web search

59% of all Web search queries are unique
17% of all queries are made only twice
8% are made three times

Hardware for big data

Budget (a.k.a. commodity) hardware

Not “gold-plated” (a.k.a. custom)

Many low-end servers

Easy to add capacity

Cheaper per CPU and per disk

Increased complexity in software:

- Fault tolerance
- Virtualization (e.g., distributed file systems)

Google Corkboard server: Steve Jurvetson/Flickr

Problems with cheap hardware

Failures, e.g. (Google numbers)

- 1–5% hard drives/year
- 0.2% DIMMs (dual in-line memory modules)/year

Commodity network (1–10 Gb/s) speeds vs. RAM

- Much more latency (100–100,000x)
- Lower throughput (100–1,000x)

Uneven performance

- Inconsistent hardware (e.g., old + new)
- Variable network latency
- External loads

These numbers are constantly changing thanks to new technology!

Google datacenter

How to program this thing?

What’s hard about cluster computing?

1. How to split work across machines?

2. How to deal with failures?

How do you count the number of occurrences of each word in a document?

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?"

I: 3
am: 3
Sam: 3
do: 1
you: 1
like: 1
...

A hashtable (a.k.a. dict)!

“I am Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?”

{}

A hashtable!

“I am Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?”

{I: 1}

A hashtable!

"I $\boxed{\text{am}}$ Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?"

{I: 1,
am: 1}

A hashtable!

"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?"

{I: 1,
am: 1,
Sam: 1}

A hashtable!

“I am Sam
I am Sam
Sam I am
Do you like
Green eggs and
ham?”

{I: 2,
am: 1,
Sam: 1}

What if the document is really big?

What if the document is really big?

“Divide and Conquer”

“Divide and Conquer”

“Divide and Conquer”

Recall: What's hard about cluster computing?

1. How to divide work across machines?
   - Moving data may be very expensive
   - Must consider network, data locality

2. How to deal with failures?
   - 1 server fails every 3 years ⇒ 10K servers see~10 faults/day
   - Even worse: stragglers (node not failed, but slow)

Solution: MapReduce

Jeff Dean [facts]

- Smart systems engineers have done all the work for you
  - Task scheduling
  - Virtualization of file system
  - Fault tolerance (incl. data replication)
  - Job monitoring
  - etc.
- “All” you need to do: implement Mapper and Reducer classes

Applied Machine Learning Days ’19 [link]

How to deal with failures?

Just launch another task!

How to deal with slow tasks?

Just launch another task!

Solution: MapReduce

Jeff Dean

- Smart systems engineers have done all the work for you
  - Task scheduling
  - Virtualization of file system
  - Fault tolerance (incl. data replication)
  - Job monitoring
  - etc.
- "All" you need to do: implement Mapper and Reducer classes

Need to break more complex jobs into sequence of MapReduce jobs

Example task

Suppose you have user info in one file, website logs in another, and you need to find the top 5 pages most visited by users aged 18–25

Example from http://wiki.apache.org/pig-data/attachments/PigTalksPapers/attachments/ApacheConEurope09.ppt

In MapReduce

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.Iterator;
import java.util.List;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.Writable;
import org.apache.hadoop.io.WritableComparable;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.KeyValueTextInputFormat;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.RecordReader;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
import org.apache.hadoop.mapred.SequenceFileInputFormat;
import org.apache.hadoop.mapred.SequenceFileOutputFormat;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.jobcontrol.Job;
import org.apache.hadoop.mapred.jobcontrol.JobControl;
import org.apache.hadoop.mapred.lib.IdentityMapper;

public class MRExample {
    public static class LoadPages extends MapReduceBase
        implements Mapper<LongWritable, Text, Text, Text> {

        public void map(LongWritable k, Text val,
                OutputCollector<Text, Text> oc,
                Reporter reporter) throws IOException {
            // Pull the key out
            String line = val.toString();
            int firstComma = line.indexOf(',');
            String key = line.substring(0, firstComma);
            String value = line.substring(firstComma + 1);
            Text outKey = new Text(key);
            // Prepend an index to the value so we know which file
            // it came from.
            Text outVal = new Text("1" + value);
            oc.collect(outKey, outVal);
        }
    }
    public static class LoadAndFilterUsers extends MapReduceBase
        implements Mapper<LongWritable, Text, Text, Text> {

        public void map(LongWritable k, Text val,
                OutputCollector<Text, Text> oc,
                Reporter reporter) throws IOException {
            // Pull the key out
            String line = val.toString();
            int firstComma = line.indexOf(',');
            String value = line.substring(firstComma + 1);
            int age = Integer.parseInt(value);
            if (age < 18 || age > 25) return;
            String key = line.substring(0, firstComma);
            Text outKey = new Text(key);
            // Prepend an index to the value so we know which file
            // it came from.
            Text outVal = new Text("2" + value);
            oc.collect(outKey, outVal);
        }
    }
    public static class Join extends MapReduceBase
        implements Reducer<Text, Text, Text, Text> {

        public void reduce(Text key,
                Iterator<Text> iter,
                OutputCollector<Text, Text> oc,
                Reporter reporter) throws IOException {
            // For each value, figure out which file it's from and
store it
            // accordingly.
            List<String> first = new ArrayList<String>();
            List<String> second = new ArrayList<String>();

            while (iter.hasNext()) {
                Text t = iter.next();
                String value = t.toString();
                if (value.charAt(0) == '1')
first.add(value.substring(1));
                else second.add(value.substring(1));
                reporter.setStatus("OK");
            }

            // Do the cross product and collect the values
            for (String s1 : first) {
                for (String s2 : second) {
                    String outval = key + "," + s1 + "," + s2;
                    oc.collect(null, new Text(outval));
                    reporter.setStatus("OK");
                }
            }
        }
    }
    public static class LoadJoined extends MapReduceBase
        implements Mapper<Text, Text, Text, LongWritable> {

        public void map(
                Text k,
                Text val,
                OutputCollector<Text, LongWritable> oc,
                Reporter reporter) throws IOException {
            // Find the url
            String line = val.toString();
            int firstComma = line.indexOf(',');
            int secondComma = line.indexOf(',', first Comma);
            String key = line.substring(firstComma, secondComma);
            // drop the rest of the record, I don't need it anymore,
            // just pass a 1 for the combiner/reducer to sum instead.
            Text outKey = new Text(key);
            oc.collect(outKey, new LongWritable(1L));
        }
    }
    public static class ReduceUrls extends MapReduceBase
        implements Reducer<Text, LongWritable, WritableComparable,
Writable> {

        public void reduce(
                Text key,
                Iterator<LongWritable> iter,
                OutputCollector<WritableComparable, Writable> oc,
                Reporter reporter) throws IOException {
            // Add up all the values we see

            long sum = 0;
            while (iter.hasNext()) {
                sum += iter.next().get();
                reporter.setStatus("OK");
            }

            oc.collect(key, new LongWritable(sum));
        }
    }
    public static class LoadClicks extends MapReduceBase
        implements Mapper<WritableComparable, Writable, LongWritable,
Text> {

        public void map(
                WritableComparable key,
                Writable val,
                OutputCollector<LongWritable, Text> oc,
                Reporter reporter) throws IOException {
            oc.collect((LongWritable)val, (Text)key);
        }
    }
    public static class LimitClicks extends MapReduceBase
        implements Reducer<LongWritable, Text, LongWritable, Text> {

        int count = 0;
        public void reduce(
                LongWritable key,
                Iterator<Text> iter,
                OutputCollector<LongWritable, Text> oc,
                Reporter reporter) throws IOException {

            // Only output the first 100 records
            while (count < 100 && iter.hasNext()) {
                oc.collect(key, iter.next());
                count++;
            }
        }
    }
    public static void main(String[] args) throws IOException {
        JobConf lp = new JobConf(MRExample.class);
        lp.setJobName("Load Pages");
        lp.setInputFormat(TextInputFormat.class);
        lp.setOutputKeyClass(Text.class);
        lp.setOutputValueClass(Text.class);
        lp.setMapperClass(LoadPages.class);
        FileInputFormat.addInputPath(lp, new
Path("/user/gates/pages"));
        FileOutputFormat.setOutputPath(lp,
            new Path("/user/gates/tmp/indexed_pages"));
        lp.setNumReduceTasks(0);
        Job loadPages = new Job(lp);

        JobConf lfu = new JobConf(MRExample.class);
        lfu.setJobName("Load and Filter Users");
        lfu.setInputFormat(TextInputFormat.class);
        lfu.setOutputKeyClass(Text.class);
        lfu.setOutputValueClass(Text.class);
        lfu.setMapperClass(LoadAndFilterUsers.class);
        FileInputFormat.addInputPath(lfu, new
Path("/user/gates/users"));
        FileOutputFormat.setOutputPath(lfu,
            new Path("/user/gates/tmp/filtered_users"));
        lfu.setNumReduceTasks(0);
        Job loadUsers = new Job(lfu);

        JobConf join = new JobConf(MRExample.class);
        join.setJobName("Join Users and Pages");
        join.setInputFormat(KeyValueTextInputFormat.class);
        join.setOutputKeyClass(Text.class);
        join.setOutputValueClass(Text.class);
        join.setMapperClass(IdentityMapper.class);
        join.setReducerClass(Join.class);
        FileInputFormat.addInputPath(join, new
Path("/user/gates/tmp/indexed_pages"));
        FileInputFormat.addInputPath(join, new
Path("/user/gates/tmp/filtered_users"));
        FileOutputFormat.setOutputPath(join, new
Path("/user/gates/tmp/joined"));
        join.setNumReduceTasks(50);
        Job joinJob = new Job(join);
        joinJob.addDependingJob(loadPages);
        joinJob.addDependingJob(loadUsers);

        JobConf group = new JobConf(MRE xample.class);
        group.setJobName("Group URLs");
        group.setInputFormat(KeyValueTextInputFormat.class);
        group.setOutputKeyClass(Text.class);
        group.setOutputValueClass(LongWritable.class);
        group.setOutputFormat(SequenceFileOutputFormat.class);
        group.setMapperClass(LoadJoined.class);
        group.setCombinerClass(ReduceUrls.class);
        group.setReducerClass(ReduceUrls.class);
        FileInputFormat.addInputPath(group, new
Path("/user/gates/tmp/joined"));
        FileOutputFormat.setOutputPath(group, new
Path("/user/gates/tmp/grouped"));
        group.setNumReduceTasks(50);
        Job groupJob = new Job(group);
        groupJob.addDependingJob(joinJob);

        JobConf top100 = new JobConf(MRExample.class);
        top100.setJobName("Top 100 sites");
        top100.setInputFormat(SequenceFileInputFormat.class);
        top100.setOutputKeyClass(LongWritable.class);
        top100.setOutputValueClass(Text.class);
        top100.setOutputFormat(SequenceFileOutpu F ormat.class);
        top100.setMapperClass(LoadClicks.class);
        top100.setCombinerClass(LimitClicks.class);
        top100.setReducerClass(LimitClicks.class);
        FileInputFormat.addInputPath(top100, new
Path("/user/gates/tmp/grouped"));
        FileOutputFormat.setOutputPath(top100, new
Path("/user/gates/top100sitesforusers18to25"));
        top100.setNumReduceTasks(1);
        Job limit = new Job(top100);
        limit.addDependingJob(groupJob);

        JobControl jc = new JobControl("Find top 100 sites for users
18 to 25");
        jc.addJob(loadPages);
        jc.addJob(loadUsers);
        jc.addJob(joinJob);
        jc.addJob(groupJob);
        jc.addJob(limit);
        jc.run();
    }
}
```

Enter: Spark

- A high-level API for programming MapReduce-like jobs

```
sc = SparkContext()
print "I am a regular Python program, using the pyspark lib"
users = sc.textFile('users.tsv')  # user <TAB> age
          .map(lambda s: tuple(s.split('\t')))
          .filter(lambda (user, age): age>=18 and age<=25)
pages = sc.textFile('pageviews.tsv')  # user <TAB> url
          .map(lambda s: tuple(s.split('\t')))
counts = users.join(pages)
              .map(lambda (user, (age, url)): (url, 1)
              .reduceByKey(add)
              .takeOrdered(5)
```

- Implemented in Scala (go EPFL!)
- Additional APIs in
  - Python
  - Java
  - R

Commercial break

RDD: resilient distributed dataset

- To programmer: looks like one single list (each element represents a “row” of a dataset)
- Under the hood: oh boy…
  - RDDs “live in the cloud”: split over several machines, replicated, etc.
  - Can be processed in parallel
  - Can be transformed to a single, real list (if small…)
  - Typically read from the distributed file system (HDFS)
  - Can be written to the distributed file system

Spark architecture

Your Python script runs in the driver ("ground")

RDD operations are run in executors ("cloud")

RDD operations

- **“Transformations”**
  - Input: RDD; output: another RDD
  - Everything remains “in the cloud”
  - Example: for every entry in the input RDD, count chars
    - RDD:[‘I’, ‘am’, ‘you’] → RDD:[1, 2, 3]
- **“Actions”**
  - Input: RDD; output: a value that is returned to the driver
  - Result is transferred “from cloud to ground”
  - Examples: take a sample of entries from RDD and print it on the driver’s shell; or store results to file (local or distributed)

Lazy execution [unrelated]

- **Transformations** (i.e., RDD→RDD operations) are not executed until it's really necessary (a.k.a. "lazy execution")
- Execution of transformations triggered by **actions**
- Why?
  - If you never look at the data, there's no point in manipulating it…
  - Smarter query processing possible:
    E.g., rdd2 = rdd1.map(f1)
    rdd3 = rdd2.filter(f2)
    Can be done in one go — no need to materialize rdd2

"I have good news and bad news"

RDD transformations [full list]

- **map**(*func*): Return a new distributed dataset formed by passing each element of the source through a function *func*
  - {1,2,3}.map(lambda x: x*2) → {2,4,6}
- **filter**(*func*): Return a new dataset formed by selecting those elements of the source on which *func* returns true
  - {1,2,3}.filter(lambda x: x <= 2) → {1,2}
- **flatMap**(*func*): Similar to map, but each input item can be mapped to 0 or more output items (so *func* should return a list rather than a single item)
  - {1,2,3}.flatMap(lambda x: [x,x*10]) → {1,10,2,20,3,30}

RDD transformations [full list]

- **sample**(*withReplacement?*, *fraction*, *seed*): Sample a fraction *fraction* of the data, with or without replacement, using a given random number generator *seed*
- **union**(*otherDataset*): Return a new dataset that contains the union of the elements in the source dataset and the argument.
- **intersection**(*otherDataset*): …
- **distinct**(): Return a new dataset that contains the distinct elements of the source dataset.

RDD transformations [full list]

- **sample**(*withReplacement?*, *fraction*, *seed*): Sample a fraction *fraction* of the data, with or without replacement, using a given random number generator *seed*

Why ***relative fraction,*** and not ***absolute number?***

POLLING TIME

Scan QR code or go to https://web.speakup.info/room/join/66626

RDD transformations [full list]

- **groupByKey**(): When called on a dataset of (K, V) pairs, returns a dataset of (K, Iterable<V>) pairs.
  - {(1,a), (2,b), (1,c)}.groupByKey() → {(1,[a,c]), (2,[b])}
- **reduceByKey**(*func*): When called on a dataset of (K, V) pairs, returns a dataset of (K, V) pairs where the values for each key are aggregated using the given reduce function *func*, which must be of type (V, V) => V.
  - {(1, 3.1), (2, 2.1), (1, 1.3)}.reduceByKey(lambda (x,y): x+y) → {(1, 4.4), (2, 2.1)}

RDD transformations [full list]

- **sortByKey**(): When called on a dataset of (K, V) pairs, returns a dataset of (K, V) pairs sorted by keys
- **join**(*otherDataset*): When called on datasets of type (K, V) and (K, W), returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key
  - {(1,a), (2,b)}.join({(1,A), (1,X)}) → {(1, (a,A)), (1, (a,X))}
- Analogous: **leftOuterJoin**, **rightOuterJoin**, **fullOuterJoin**
- (There are several other RDD transformations, and some of the above have additional arguments; cf. tutorial)

RDD actions [full list]

- **collect**(): Return all the elements of the dataset as an array at the driver program. This is usually useful after a filter or other operation that returns a sufficiently small subset of the data.
- **count**(): Return the number of elements in the dataset.
- **take**(*n*): Return an array with the "first" *n* elements of the dataset.
- **saveAsTextFile**(*path*): Write the elements of the dataset as a text file in a given directory in the local filesystem or HDFS.
- (There are several other RDD actions; cf. tutorial)

Broadcast variables

- my_set = set(range(1e80))
  rdd2 = rdd1.filter(lambda x: x in my_set)
  ^ This is a bad idea: my_set needs to be shipped with every task (one task per data partition, so if rdd1 is spread over *N* partitions, the above will require copying the same object *N* times)
- Better:
  my_set = sc.broadcast(set(range(1e80)))
  rdd2 = rdd1.filter(lambda x: x in my_set.value)
  ^ This way, my_set is copied to each executor only once and persists across all tasks (one per partition) on the same executor
- Broadcast variables are **read-only**

Accumulators

- def f(x): return x*2
  rdd2 = rdd1.map(f)
  ^ How can we easily know how many rows there are in rdd1 (without running a costly reduce operation)?
- Side effects via accumulators!
  counter = sc.accumulator(0)
  def f(x): counter.add(1); return x*2
  rdd2 = rdd1.map(f)
- Accumulators are **write-only** ("add-only") for executors
- Only driver can read the value: counter.value

RDD persistence

```
rdd2 = rdd1.map(f1)
list1 = rdd2.filter(f2).collect()
list2 = rdd2.filter(f3).collect()
```

} rdd1.map(f1) transformation is executed twice

```
rdd2 = rdd1.map(f1)
rdd2.persist()
list1 = rdd2.filter(f2).collect()
list2 = rdd2.filter(f3).collect()
```

} Result of rdd1.map(f1) transformation is cached and reused (can choose between memory and disk for caching)

Spark DataFrames

- Bridging the gap between your experience with Pandas and the need for distributed computing
  - RDD = list of rows
  - DataFrame = table with rows and typed columns
- Important to understand what RDDs are and what they offer, but today most of the tasks can be accomplished with DataFrames **(higher level of abstraction ⇒ less code)**
- https://www.databricks.com/spark/getting-started-with-apache-spark/dataframes

Spark SQL [link]

```
sc = SparkContext()

sqlContext = HiveContext(sc)

df = sqlContext.sql("SELECT * from table1 GROUP BY id")
```

Spark's Machine Learning Toolkit

MLlib: Algorithms [more details]

Classification

- Logistic regression, decision trees, random forests

Regression

- Linear (with L1 or L2 regularization)

Unsupervised:

- Alternating least squares
- K-means
- SVD
- Topic modeling (LDA)

Optimizers

- Optimization primitives (SGD, L-BGFS)

Example:
Logistic regression with MLLib

```
from pyspark.mllib.classification \
   import LogisticRegressionWithSGD

trainData = sc.textFile("...").map(...)
testData = sc.textFile("...").map(...)
model = LogisticRegressionWithSGD.train(trainData)
predictions = model.predict(testData)
```

Remarks

- This lecture is not enough to teach you Spark!
- To use it in practice, you'll need to delve into further online material
- Also: Friday's lab session
- You can't learn it without some frustration :(
- Important skill: assess whether you'd benefit from Spark
  - E.g., >1TB: yes, you'll need Spark
  - 20GB: it depends…

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec13-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo?
- …

Cluster etiquette

- Develop and debug locally
  - Install Spark locally on your personal computer
  - Use a small subset of the data
- When ready, launch your script on the cluster using spark-submit
- **Never (never!) use the Spark shell (a.k.a. pyspark) -- it's hereby officially forbidden**
- Useful trench report from a dlab member: "What I learned from processing big data with Spark"

Applied Data Analysis (CS401)

ada

Lecture 14
ADA in action
20 Dec 2023

EPFL

Robert West

dlab

Announcements

- Today: last lecture [click here for a famous last lecture]
- No lab session this Friday
- Homework H2 feedback slightly delayed 🤯 — ETA tomorrow
- Final project milestone P3 due on Fri 22 Dec 2023, 23:59
- Final exam: Tue 16 Jan 2024, 15:15–18:15
  - Announcement re: exam protocol, room assignment, etc., to be made in early January (on Ed with email notification)

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec14-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Today I will

- present a research paper of mine,
- highlight how everything you've learned in lectures 1–13 comes together in one project.

Paper available at https://doi.org/10.1073/pnas.2106152118

Postmortem memory of public figures in news and social media

Robert West[a,1], Jure Leskovec[b], and Christopher Potts[c]

[a]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [b]Department of Computer Science, Stanford University, Stanford, CA 94305; and [c]Department of Linguistics, Stanford University, Stanford, CA 94305

Edited by Henry L. Roediger III, Washington University in St. Louis, St. Louis, MO, and approved June 24, 2021 (received for review April 1, 2021)

Deceased public figures are often said to live on in collective memory. We quantify this phenomenon by tracking mentions of 2,362 public figures in English-language online news and social media (Twitter) 1 y before and after death. We measure the sharp spike and rapid decay of attention following death and model collective memory as a composition of communicative and cultural memory. Clustering reveals four patterns of postmortem memory, and regression analysis shows that boosts in media attention are largest for premortem popular anglophones who died a young, unnatural death; that long-term boosts are smallest for leaders and largest for artists; and that, while both the news and Twitter are triggered by young and unnatural deaths, the news additionally curates collective memory when old persons or leaders die. Overall, we illuminate the age-old question of who is remembered by society, and the distinct roles of news and social media in collective memory formation.

computational social science | collective memory | news and social media analysis | forgetting

Being remembered after death has been an important concern for humans throughout history (1), and conversely, many cultures have considered *damnatio memoriae*—being purposefully erased from the public's memory—one of the most severe punishments conceivable (2). To reason about the processes by which groups and societies remember and forget, the French philosopher and sociologist Maurice Halbwachs introduced the concept of collective memory in 1925 (3), which has since been a subject of study in numerous disciplines, including anthropology, ethnography, philosophy, history, psychology, and sociology, and which gave rise to the new discipline of memory studies (4). Over the decades, collective memory has moved from being a purely theoretical construct to becoming a practical phenomenon that can be studied empirically (5), e.g., in order to quantify to what extent US presidents are remembered across generations (6) or how World War II is remembered across countries (7).

Whereas oral tradition formed the basis for collective memory in early human history, today the media play a key role in determining what and who is remembered, and how (8–11). Researchers have studied the role of numerous media in constructing the postmortem memory of deceased public figures. A large body of work has investigated the journalistic format of the obituary (12–16), which captures how persons are remembered around the time of their death (14). Taking a more long-term perspective, other work has considered how deceased public figures are remembered in the media over the course of years and decades (17–21). As ever more aspects of life are shifting to the online sphere, the Web is also gaining importance as a global memory place (22), which has led researchers to study, e.g., how social media users (23–27) and Wikipedia editors (28) react to the death of public figures. In the context of social media, the detailed analysis of highly visible individual cases, such as Princess Diana (24), pop star Michael Jackson (25, 26), or race car driver Dale Earnhardt (27), has revealed how people experience and overcome the collective trauma that can ensue following the death of celebrities.

Although such studies of individuals have led to deep insights at a fine level of temporal granularity, they lack breadth by excluding all but some of the very most prominent public figures. What is largely absent from the literature is a general understanding of patterns of postmortem memory in the media that goes beyond single public figures.

To bridge this gap, we draw inspiration from a body of related work that has studied the temporal evolution of collective memory using large-scale datasets—although, unlike our work, not with a focus on the immediate postmortem period of public figures. For instance, van de Rijt et al. (20) tracked thousands of person names in news articles, finding that famous people tend to be covered by the news persistently over decades. In a similar analysis, Cook et al. (19) further showed that the duration of fame had not decreased over the course of the last century. Beyond news corpora, the online encyclopedia Wikipedia has become a prime resource for the data-driven study of collective memory. Researchers have leveraged the textual content of Wikipedia articles (29), as well as logs of both editing (30) and viewing (31, 32), as proxies for the collective memory of traumatic events such as terrorist attacks or airplane crashes. Jatowt et al. (33) characterized the coverage and popularity of historical figures in Wikipedia, observing vastly increased page-view counts for people from the 15th and 16th centuries, a fact that Jara-Figueroa et al. (34) later attributed to the invention of the printing press. In addition to news and

Significance

Who is remembered by society after they die? Although scholars as well as the broader public have speculated about this question since ancient times, we still lack a detailed understanding of the processes at work when a public figure dies and their media image solidifies and is committed to the collective memory. To close this gap, we leverage a comprehensive 5-y dataset of online news and social media posts with millions of documents per day. By tracking mentions of thousands of public figures during the year following their death, we reveal and model the prototypical patterns and biographic correlates of postmortem media attention, as well as systematic differences in how the news vs. social media remember deceased public figures.

Author contributions: R.W., J.L., and C.P. designed research; R.W. performed research; R.W. analyzed data; and R.W., J.L., and C.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

[1]To whom correspondence may be addressed. Email: robert.west@epfl.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2106152118/-/DCSupplemental.

Published September 15, 2021.

Downloaded from https://www.pnas.org by 157.143.15.69 on December 20, 2022 from IP address 157.143.15.69.

COMPUTER SCIENCES

SOCIAL SCIENCES

Philip Seymour Hoffman † 2014

Amy Winehouse † 2011

Questions

- Who is remembered by society after death?
  - “Postmortem collective memory”
- Are there prototypical patterns of postmortem collective memory?
- Are certain kinds of people remembered in certain ways?
- Are dead people remembered differently in news vs. social media?

ada #6 Observational studies

Why should we care about postmortem collective memory?

Fact: Humans care a lot about being remembered after death

Ara Pacis, Rome
[Wikipedia]

Damnatio memoriae
[Wikipedia]

An ADA approach

Stuffed Count von Count counting stuff

Let's use lots of data and count stuff!

- **Detect names** of dead people in big corpus of news and social media
- Build time series of name **counts**
- Analyze the **shape of time series**
- **Correlate** shapes with biographic info about dead people from knowledge base

Part 1:
Getting the data

The raw data: spinn3r

- "Spinn3r provides APIs for social media, weblogs, news, video, and live web content to our customers in any language and in large volumes." (Source: spinn3r.com)
- Firehose stored to dis
- Several billions of doc
- 40 terabytes

Postmortem memory of public figures in news and social media

Robert West[a,1], Jure Leskovec[b], and Christopher Potts[c]

[a]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [b]Department of Computer Science, Stanford University, Stanford, CA 94305; and [c]Department of Linguistics, Stanford University, Stanford, CA 94305

Edited by Henry L. Roediger III, Washington University in St. Louis, St. Louis, MO, and approved June 24, 2021 (received for review April 1, 2021)

ada #13 Scaling to massive data

Detecting news and social media in Spinn3r

- Found a list of all 151K online news articles about Osama bin Laden's killing (2 May 2011) indexed by Google News
- Assume that every relevant news outlet had reported on bin Laden's death, and that Google News is reasonably complete
- News defined as documents from the 6,608 Web domains appearing in the "bin Laden list"

- Detect via URL pattern: e.g., https://twitter.com/UserName/status/1605097142552403968

ada #2 Handling data

Data volume: Number of docs per day

ada #3 Visualizing data

Detecting mentions of people names

ada

#2 Handling data
#12 Handling network data
#10+11 Handling text data

- Easy: unambiguous names
  - e.g., “Amy Winehouse”
- Hard: ambiguous names
  - e.g., “Michael Jackson”

Wikipedia-based solution:

- Given: a name *X* (e.g., “Michael Jackson”)
- Consider all *N* links in English Wikipedia with *X* as anchor text (cf. examples on the right)
- Build distribution over the *N* link targets
- If there is a target *T* to which ≥90% of all links point: consider *X* “sufficiently unambiguous” and assume that all mentions of *X* in Spinn3r refer to *T*
- Else: consider *X* “too ambiguous” and ignore it in the analysis

The city of Gary is known as the birthplace of singer Michael Jackson.

Beer and whisky expert Michael Jackson was born in Leeds.

…

WIKIPEDIA

Michael Jackson (disambiguation)

Article Talk

Michael Jackson (1958–2009) was an American singer, songwriter and dancer known as the "King of Pop".

Michael Jackson, **Mike Jackson**, or **Mick Jackson** may also refer to:

People

Entertainment industry

- Michael Jackson (radio commentator) (1934–2022), American radio talk show host, KABC and KGIL, Los Angeles
- Michael Jackson (writer) (1942–2007), *Beer Hunter* show host, beer and whisky expert
- Mick Jackson (director) (born 1943), British film and TV director, known for *The Bodyguard*
- Michael J. Jackson (born 1948), English actor from Liverpool, best known for his role in *Brookside*
- Michael Jackson (television executive) (born 1958), British television executive
- Mick Jackson (author) (born 1960), British writer, known for *The Underground Man*
- Mike Jackson (photographer) (born 1966), British abstract and landscape photographer, known for *Poppit Sands* images
- Michael Jackson (actor) (born 1970), Canadian actor
- Mike Jackson (film producer) (born 1972), American film producer and talent manager
- Michael R. Jackson (born 1981), American playwright, composer, and lyricist

Musicians

- Mike Jackson (musician) (1888–1945), American jazz pianist and composer
- Mike Jackson (Australian entertainer) (born 1946), Australian multi-instrumentalist, songwriter and children's entertainer
- Mick Jackson (singer) (born 1947), English singer-songwriter
- Michael Gregory (jazz guitarist) (born 1953), American jazz guitarist, born Michael Gregory Jackson
- Mike and Michelle Jackson, Australian multi-instrumental duo
- Michael Jackson (English singer) (born 1964), British singer with the heavy metal band Satan/Pariah
- Oh No (musician), birth name Michael Woodrow Jackson (born 1978), American rapper
- Michael Lee Jackson, guitarist
- Mick Jackson, bassist with British band Love Affair (1950-)

Military and militants

- Michael Jackson (American soldier) (1734–1801), soldier from Massachusetts, wounded at Bunker Hill
- Mike Jackson (British Army officer) (born 1944), former head of the British

Recruiting the army of the dead: Freebase

	All	Included
Age		
N/A	10%	6%
1st quartile	68	64
Mean	76	74
Median	80	77
3rd quartile	88	87
Gender		
N/A	27%	7%
Female	16%	17%
Male	84%	83%
Manner of death		
N/A	76%	60%
Natural	85%	86%
Unnatural	15%	14%
Language		
N/A	45%	27%
Anglophone	60%	82%
Non-anglophone	40%	18%
Notability type		
N/A	1%	0%
Arts	40%	50%
Sports	14%	14%
Leadership	11%	14%
Known for death	26%	16%
General fame	7%	4%
Academia/engineering	2%	2%
Count	33 340	2 362

- Freebase, a once-popular knowledge graph
  - Now defunct, but still available
- Contains information about >3M people
  - e.g., Philip Seymour Hoffman = /m/02qgqt
  - More info about some, less about others
- Start from 33K people with death date 2009–2014
  - Further filtering → 2,362 people
- Extract biographic information from Freebase

ada #2 Handling data
#12 Handling network data

Tools used

ada #13 Scaling to massive data

- Counting names in Spinn3r dump: Hadoop (Java)
- Extracting info from Freebase: Python, Perl
- Once data was small enough: R (script, repo)
  - Statistical analyses
  - Plotting
  - Read data as CSV once (minutes), then serialized to binary format and deserialized in later runs (seconds)

ada #2 Handling data

Intrat: Mention time series (our protagonist)

- t: relative time t, counting days since death
  - $t = 0$: day of death
- $S_i(t)$: fraction of documents in which person i was mentioned, out of all documents published on day t
  - For mention time series, consider logarithms: $\log_{10} S_i(t)$

ada #1–13 ADA loves logs!

Mention time series: examples

News

Twitter

Smoothed via "Friedman's super smoother"

Part 2:
The shape of postmortem memory

Average mention time series

News

ada #4 Describing data

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \cdots a_n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln a_i\right)$$

Modeling the data

ada #3 Visualizing data

$y(t) = v(t)/(u(t) + v(t))$

Inflection point after about 1 month:
→ short vs. long term

Curve characteristics

Pre-mortem mean: arithmetic mean of days 360 through 30 before death

Short-term boost: maximum of days 0 through 29 after death, minus the premortem mean

Long-term boost: arithmetic mean of days 30 through 360 after death, minus the premortem mean

Halving time: number of days required to accumulate half of the total area between the postmortem curve (including the day of death) and the minimum postmortem value

Median over people:
1.98
95% CI [1.90, 2.03]

Median over people:
0.00055
95% CI [-0.00091, 0.0017]

ada #4 Describing data

The Surgeon General warns:

WIKIPEDIA
The Free Encyclopedia

Create account Log in

Adenosine deaminase

17 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Adenosine deaminase (also known as **adenosine aminohydrolase**, or **ADA**) is an enzyme (EC 3.5.4.4) involved in purine metabolism. It is needed for the breakdown of adenosine from food and for the turnover of nucleic acids in tissues.

Its primary function in humans is the development and maintenance of the immune system.[5] However, the full physiological role of ADA is not yet completely understood.[6]

ADA

Structure [edit]

ADA exists in both small form (as a monomer) and large form (as a dimer-complex).[6] In the monomer form, the enzyme is a polypeptide chain,[7] folded into eight strands of parallel α/β barrels, which surround a central deep pocket that is the active site.[5] In addition to the eight central β-barrels and eight peripheral α-helices, ADA also contains five additional helices: residues 19-76 fold into three helices, located between β1 and α1 folds; and two antiparallel

Are there prototypical curve shapes?

- Each curve one data point
- Represented via its 4 curve characteristics
  - ≈ manual dimensionality reduction
  - Values standardized via *z*-scores
- Cluster via *k*-means

ada

#8 Applied ML
#9 Unsupervised learning

Q: How to find the best number *k* of clusters?

Average silhouette width!

ada #9 Unsupervised learning

News

Twitter

Cluster analysis

ada #4 Describing data

Part 3:
Biographic correlates of postmortem memory

A first stab

- Measure correlation coefficient between outcome (e.g., short-term memory boost) and each biographic properties (e.g., gender)
- Higher correlation ⇒ higher outcome for the respective group of people (e.g., women)

Age
N/A
1st quartile
Mean
Median
3rd quartile
Gender
N/A
Female
Male
Manner of death
N/A
Natural
Unnatural
Language
N/A
Anglophone
Non-anglophone
Notability type
N/A
Arts
Sports
Leadership
Known for death
General fame
Academia/engineering
Count

Problem: Biographic properties are correlated

E.g., leaders (politicians, CEOs, etc.) are

- more likely to have died old,
- more likely to have died of a natural death,
- more likely to be men,

compared to artists

ada #5 Regression analysis

Regression analysis allows us to compare averages across subgroups of the data while accounting for correlations among averaged values!

Linear regression

Avg. outcome for "baseline persona": male anglophone artist of median premortem popularity who died a natural death at age 70–79

Rank-transformed, then linearly scaled/shifted to [–0.5, 0.5]; i.e., median has value 0

$$y_i = \beta_0 + \beta_1\,\text{premortem_mention_freq}_i + \beta_2\,\text{age_at_death}_i + \beta_3\,\text{manner_of_death}_i + \beta_4\,\text{notability_type}_i + \beta_5\,\text{language}_i + \beta_6\,\text{gender}_i$$

Outcome for person *i*:

- short-term boost or
- long-term boost

8 discrete levels (dummy-coded): 20–29, 30–39, …, 70–79, 80–89, 90–99

2 levels: natural, unnatural

6 levels: arts, sports, leadership, known for death, general fame, academia/engineering

3 levels: anglophone, non-anglophone, unknown

2 levels: male, female

Linear regression results

	Short-term boost (news)	Short-term boost (Twitter)	Long-term boost (news)	Long-term boost (Twitter)
(Intercept)	2.322 (0.063)***	2.670 (0.067)***	0.088 (0.014)***	0.095 (0.015)***
Premortem mean (relative rank)	0.804 (0.093)***	0.948 (0.100)***	0.031 (0.020)	0.086 (0.022)***
Manner of death: unnatural	0.618 (0.095)***	0.282 (0.100)**	0.097 (0.021)***	0.106 (0.022)***
Language: non-anglophone	−0.316 (0.074)***	−0.116 (0.078)	−0.061 (0.016)***	−0.037 (0.017)*
Language: unknown	−0.446 (0.086)***	−0.325 (0.091)***	−0.079 (0.019)***	−0.081 (0.020)***
Gender: female	0.083 (0.072)	−0.034 (0.076)	0.034 (0.016)*	0.006 (0.017)
Notability type: academia/engineering	0.181 (0.197)	0.340 (0.208)	−0.032 (0.043)	0.023 (0.046)
Notability type: general fame	0.070 (0.124)	0.132 (0.131)	−0.010 (0.027)	−0.008 (0.029)
Notability type: known for death	−0.107 (0.099)	−0.088 (0.106)	−0.021 (0.022)	0.008 (0.023)
Notability type: leadership	0.152 (0.083)	0.113 (0.087)	−0.058 (0.018)**	−0.040 (0.019)*
Notability type: sports	0.049 (0.083)	0.072 (0.088)	−0.034 (0.018)	−0.034 (0.020)
Age: 20–29	0.162 (0.170)	0.718 (0.180)***	0.060 (0.037)	0.192 (0.040)***
Age: 30–39	0.400 (0.167)*	0.649 (0.177)***	0.028 (0.037)	0.118 (0.039)**
Age: 40–49	−0.046 (0.126)	0.351 (0.133)**	−0.001 (0.028)	0.100 (0.030)***
Age: 50–59	−0.075 (0.099)	0.181 (0.104)	−0.058 (0.022)**	−0.024 (0.023)
Age: 60–69	−0.109 (0.082)	0.130 (0.086)	−0.050 (0.018)**	−0.025 (0.019)
Age: 80–89	0.022 (0.078)	0.021 (0.082)	−0.018 (0.017)	−0.013 (0.018)
Age: 90–99	0.174 (0.098)	0.034 (0.103)	−0.011 (0.021)	−0.024 (0.023)
R^2	0.213	0.192	0.123	0.178
Adj. R^2	0.197	0.176	0.106	0.161
No. obs.	870	870	870	870
RMSE	0.772	0.815	0.169	0.181

SEs of coefficients are in parentheses. ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$.

Age at death vs. postmortem memory

News plays two simultaneous roles (more so than Twitter):

- Catering to public curiosity stirred by a young or unnatural death
- But also when old person or accomplished leader dies

Part 4:
Discussion

Summary: The shape of postmortem memory

- Sharp **pulse** of media attention with death:
  - Median: +9,400% in news, +28,000% on Twitter
- Then sharp **drop** (around 1 month long) toward premortem level
- **Two components** of collective memory:
  - Baseline level of **cultural** memory built up during life (constant)
  - Added layer of **communicative** memory sparked by death (power law)
- **Cluster analysis** revealed a set of four prototypical memory patterns: “Blip”, “silence”, “rise”, “decline”
- Same patterns in news and Twitter; **same person** tends to fall into the **same cluster** across the two media

Summary: Biographic correlates

- Notability types: all regression coefficients for long-term boosts negative
  ⇒ All types have lower average long-term boost than default type (artists)
  ⇒ **Artists more present in collective memory**
- **Low R^2** (10–20%): human lives/legacies rich, hard to model
  - But all **model fits highly significant** (*F*-statistic, *p*-value)
  - **Effects** not only significant, but also **large:** e.g., short-term boost (on linear scale) for unnatural vs. natural death: 4x in news, 2x on Twitter
- **Largest boost:**
  - **Premortem popular anglophones who died a young, unnatural death**
  - Long-term boosts **largest for artists, smallest for leaders**

Merry Christmas ADA happy New Year!