COGS138: Neural Data Science

Lecture 15

C. Alex Simpkins, PhD

UCSD Dept. of Cognitive Science, Spring 2023

RDPRobotics, LLC

http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23

rdprobotics@gmail.com | csimpkinsjr@ucsd.edu

Plan for today

- Announcements

- Review of course status

- In class paper

- If time, Intro to parameterizing heterogeneous datasets in neural data science, issues, techniques, readings

Announcements

- ***Deadlines upcoming this week:***

- **Tuesday** (today) :

  - Previous project review 11:59pm

  - Mid-quarter checkin 11:59pm

  - Lecture Quiz 3 (EXTENDED) 11:59pm

- **Wednesday** (tomorrow) Assignment 3 (EXTENDED) 11:59pm

- **Saturday**:

  - Reading Quiz 3 11:59pm

- **Friday**:

  - Project proposal 11:59pm

Announcements

- **github** repos

  - created,

  - invites sent,

  - please accept (time limited)

  - login and be sure you can and files are there, rename

- if you don't have an invite, there's an issue with your group record in the main list

  - **Procedure** : Contact Siddhant, cc me, if no response in a day, reach out to me again, I'll help

Announcements

- **Project meetings**

  - You will need to sign up for a 10 min project meeting to introduce your project idea and get feedback ahead of your proposal so we can provide input early

  - Llook for the signup google form tonight

Project schedule

Task due	Date due	Description
Previous project review	5/23/2023 at 11:59pm (Tuesday)	Select 2 of the 3 available, review as individuals and then come together as a group to submit your responses to the questions after a discussion. This will orient you to the class project
Project proposal	5/26/2023 at 11:59pm (Friday wk8)	Generate your question, hypothesis, initial data sets you'll be working with, etc., describe your plan, schedule, who is doing what, potential issues, suggested analysis and how it will answer your question
Data checkpoint	6/2/2023 at 11:59pm (Friday wk9)	Builds on the proposal by taking the feedback from PP above and actually getting, loading, describing your data,
EDA checkpoint	6/9/2023 at 11:59pm (Friday wk10)	Builds on the previous checkpoint, essentially most of your analysis should be done by this point
Final report	6/15/2023 at 11:59pm (Thursday Fin wk)	Due Thursday of finals week so we can grade before the Tuesday deadline, otherwise your grade may be delayed
Group evaluations	6/15/2023 at 11:59pm (Thursday Fin wk)	You will evaluate each other based on participation and performance, this will contribute to your overall final project grade 5%)

Some notes on datasets

- 2 or more discussion and motivation

- What each portion represents as an idea (prev. project review, proposal, data, eda, final checkpoints)

Remaining assignments schedule

- A4 wk8-9, A5 wk9-10

- R4 wk8-9

- LQquiz wk 8, 9, 10

- Paper this week, mostly in class or via appointment

In class report development (~30m)

- Define this course's intent

- Draw comparisons between this course and requirements

- How does this course build upon what came before?

- How can you use your starting point in this course to expand your understanding?

- resampling neural data by example

- PCA for neural data by example

- parameterizing heterogeneous datasets

Parameterizing heterogeneous datasets

- Definition, review

  - What do we mean by **parameterization**?

  - Reminder of what data is and stepping back to the big picture - ***representation***

  - What are **heterogeneous** datasets?

  - What are the **challenges** and solutions?

- **Tools and practice** in neural data science

  - https://nwb-overview.readthedocs.io/en/latest/tools/tools_home.html

- Examples

Parameterization vs. Hyperparameterization

- **Parameterization** - the set of parameters that define the model unknowns to be fit, typically from data

  - For example, for y = ax + b, what are the parameters?

  - ANN - network weights

  - Calculated/learned from data

- **Hyperparameterization** - the set of parameters for machine learning in particular that define and control the learning process and are external to the model

  - Bisection algorithm for optimization - bisection parameter

  - ANN - parameters of the learning algorithm itself

  - Heuristic, can be set by practitioner, tunable for a given problem

Parameterization vs. Hyperparameterization

- **Parameters**

  - Calculated/learned from data ("the fit")

  - Internal to model

  - Chosen as part of model structure either manually or algorithmically

- **Hyperparameters**

  - Heuristic, can be set by practitioner, tunable for a given problem

  - External from model

  - Optimal parameters are not known, and are different for different problems

  - Other ex.: ANN learning rate, gradient descent step size, the k in k-nearest neighbors

Stepping back: What is data?

Stepping back: What is data?

- **Data** can be of many forms

- **Data** - any representation of information that has been recorded in a fixed or dynamic state [Simpkins, 2023]

- *"(1) Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation" [Webster's, 2023]*

- *(2) "Information in digital form that can be transmitted or processed" [Webster's, 2023]*

What do we mean by 'representation'

What do we mean by 'representation'

- Some sort of arbitrary symbolic link between the reality and the symbols we use to model reality, such as words, numbers, pictures, graphs, sounds, videos, smells, textures, vibrations, gestures
- Further information: See Simpkins COGS100 lecture 10 on representation, and read Norman ch3: "The power of representation"

Representation defined

- Cognitive age, Norman argues started when we started using sounds, gestures and symbols to refer to objects, things and concepts - when we started generating data!

- **Representation** : The sound, gesture, symbol is not the thing itself, it stands for, refers to it

- On representation not the reality

Powers of cognition come from abstraction and representation

- Ability to represent perceptions, experiences, thoughts in some medium other than what they occurred in

- Abstracted away from irrelevant details

- "The essence of intelligence" as he states - if representation is just right, new experiences, insights, creations emerge

- **We can make symbols then use them to do our reasoning**

Representing the dimensions requires different types of data entirely

- Ultimately in neural data science we are reasoning about the brain and behavior, how it's all interconnected and the dynamics of it

- Data makes it possible to reach beyond our immediate cognitive limitations to operate on information

  - We cannot see a neuron firing when we look at each other, we measure, but then must do something with that data, related it and connect it meaningfully to other things

  - As we have been reasoning, we need massive amounts of connections to understand the patterns of it all

  - Recording it all the same way often is impossible

    - EEG vs. Behavior, text, other dimensions

Data Structures Review

Structured data

- Can be stored in database SQL

- Tables with rows and columns

- Requires a relational key

- 5-10% of all data

Semi-structured data

- Doesn't reside in a relational database

- Has organizational properties (easier to analyze)

- CSV, XML, JSON

Unstructured

- Non-tabular data

- 80% of the world's data

- Images, text, audio, videos

(Semi-)Structured Data

Data that is stored in such a way that it is easy to search and work with. These data are stored in a particular format that adheres to organization principles imposed by the file format. These are the data structures data scientists work with most often.
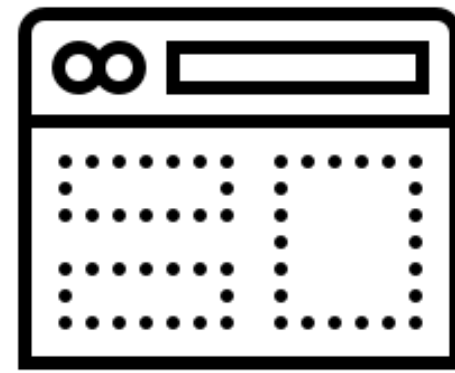
Unstructured Data

Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.

Unstructured Data Types

Text files
and
documents

Websites
and
applications

Sensor
data

Image
files

Audio
files

Video
files

Email
data

Social
media
data

What are heterogeneous datasets?

- Given that **data can represent anything that can be represented**, we can have many forms of sampling and recording systems

- What have we covered thus far for data types and forms?

- MOCAP

- EEG/MEG

- fMRI

- Eye tracking

- Text

- Single unit recording

Why is it a challenge to integrate them?

- Sampling rate mismatch

- Time/frequency/spatial domains

- Sample rate variability (why does this matter?)

- Sample time mismatch

- Format, software

- Missing data, data mixture/non-tabular etc

- Memory usage

- (Not an exhaustive list)

Integration strategies - *Sample rate mismatch*

- Resampling

  - Sub-sampling - ("down-sampling") - every Mth sample, LowPass first (aliasing)

  - Super-sampling - ("up-sampling") - padding with 0's, then LowPass to interpolate

- Interpolation/extrapolation (what are the differences?)

  - <u>Linear (LERP, BERP, TERP, SLERP)</u>

  - Piecewise continuous

    - <u>Splines, Bezier</u>

  - Polynomial

    - <u>Lagrange</u>, etc

Integration strategies - Time/frequency/spatial domains

Integration strategies - Sample time mismatch

- Super/sub sampling with filtering

- Time/sample shift to align data

Integration strategies - Software and format

- What if you have image/video data, EEG, text, audio?

- Each is in its own format, with different sample timings, not keyed events

- Traditional way?

- Newer way?

Integration strategies - Sample rate variability

- Do you have an accurate time measure and know the variability?

  - Yes - then you can simply interpolate and resample to create a new equally spaced set

- Inaccurate time measure, some information is lost

  - Computer timers for example do not provide accurate time measures unless they are specialized hardware

  - Can assume it's accurate if sampling much much faster than dynamics

  - Reduce sample rate (sub-sample) below estimated variability

  - Cannot use for time-critical associations

Integration strategies- missing data, mixture, non-tabular

- Addressed in earlier lectures (NANs)

- Wrangling

- Manual labor

- May need to use portions of the data

- Large sets need automated detection means

Integration strategies - Memory and processor usage