RMACC TALKS, DAY 2

COLTON GRAINGER (SCRIBE)

1. Unsupervised clustering of telemetry data with deep neural networks

With Natalya Rapstine and Jeff Tracey. Here's the abstract.

> We use recurrent autoencoder neural network to encode the sequential California golden eagle telemetry data. The encoding is followed by K-means algorithm to cluster the data into behavior classes. The encoded data representation achieves better K-means clustering results as measured by the silhouette score compared to the K-means clustering of the original data.

1.1. **links.**

- "2.3. Clustering — scikit-learn 0.21.1 documentation"[1].
- "Building Autoencoders in Keras"[2]. "In this tutorial, we will answer some common questions about autoencoders, and we will cover code examples of the following models:; a simple autoencoder based on a fully-connected layer; a sparse autoencoder; a deep fully-connected autoencoder; a deep convolutional autoencoder; an image denoising model; a sequence-to-sequence autoencoder; a variational autoencoder"
- "von Mises distribution - Wikipedia"[3]. English Wikipedia.
- "Understanding LSTM Networks – colah's blog"[4]. "Humans don't start their thinking from scratch every second. As you read this essay, you understand each word based on your understanding of previous words. You don't throw everything away and start thinking from scratch again. Your thoughts have persistence.Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist."
- "Savitzky–Golay filter - Wikipedia"[5]. English Wikipedia. "A Savitzky–Golay filter is a digital filter that can be applied to a set of digital data points for the purpose of smoothing the data, that is, to increase the precision of the data without distorting the signal tendency. This is achieved, in a process known as convolution, by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares."

2. Tips and tricks for easy data management

With Mara Sedlins, Katie Mika, and Meg Eastwood. Here's the abstract.

Date: 2019-05-22.

url: `true`.

[1] `https://scikit-learn.org/stable/modules/clustering`

[2] `https://blog.keras.io/building-autoencoders-in-keras.html`

[3] `https://en.wikipedia.org/wiki/Von_Mises_distribution`

[4] `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

[5] `https://en.wikipedia.org/wiki/Savitzky%E2%80%93Golay_filter`

Data management is more than backing up files and cloud storage. Managing data, including sharing and access, security, active use, short and long term storage, metadata and description, and publishing practices are integral to the research process and require thoughtful planning. In addition to increasing research efficiency by improving organization, effective data management helps to ensure the quality of your research and supports published results. While data management often depends on project specifics like the type of data, how the data is collected, and how it's used throughout the life of the project, this panel of data management experts will provide some useful methods and common strategies for effectively managing raw and in-use data. Significant time will be afforded for audience questions about tricky data sets and anecdotes.

2.1. **Documentation.**

1. It's metadata: "metadata is a love note to the future"
   - relevant information for re-creation of re-use
   - READMEs
2. What's study level metadata? Includes *study context*.
   - source of data
   - data collection and methodology
   - measures or instrumentation used
   - about the data files *organizational schema, software*
   - programs used to process/manipulate data
3. What's data level metadata? Includes the *use of data*.
   - variable names and descriptions
   - units of measurement
   - derived variables
4. READMEs include what?
   - an abstract or an inventory?
   - it's *study-level*
   - describes the *content* of data files
5. Codebooks are?

variable name	description	SI units
`object_to_scale`	to be scaled	**R** in `meters`

6. Literate programming gives us (e.g., with Jupyter or Rmd)

   `human readable narrative + code = reproducible documents.`
7. Sharing data and code with electronic lab notebooks
   - Rspace
   - labarchives
8. Sharing software environments
   - docker
   - binder

2.2. **data sharing.** See SPARC `http://datasharing.sparcopen.org/` for data sharing requirements for funding from federal agencies. There's also a DMPTool[6] for writing a DMP quickly.

[6] `https://dmptool.org/`

We'll focus on sharing data with two groups.

1. With collaborators (incorporates "long-content").
   - aws, google drive, dropbox
   - slack, email threads
   - github, bitbucket
   - latex editing
   - do folks understand their responsibilities for data backup?
2. With strangers.
   - it's altruistic (avoids duplication)
   - published datasets should have `DOI`s (`https://datacite.org`)
   - healthy for the subject field (reduces fraud, improve integrity)
3. Excuses for not sharing data.
   - "it's too big" = it's unstructured.
     - go to the Center for Research Data & Digital Scholarship[7] to share "B"ig "D"ata.
   - "it's niche, exotic, eclectic, and only I understand it"
   - "it's highly sensitive"

2.3. **questions.**

- rubrics for data management plans? `https://dmptool.org/`
- version control for ~TB file sizes?
  - try functional programming, e.g., Haskell?

3. Open OnDemand Demo

These are notes from an informal introduction with Martin Cuma, `https://rmacc2019hpcsymposium.sched.com/event/Ja80/open-ondemand-demo`.

What's Open OnDemand? From the README[8]:

> The Open OnDemand Project is an open-source software project, based on the Ohio Supercomputer Center's proven "OSC OnDemand" platform, that enables HPC centers to install and deploy advanced web and graphical interfaces for their users. More information can be found in the paper `http://dx.doi.org/10.1145/2949550.2949644`.

There are 2 OnDemand roadmaps for future development (hosted by OSC on Discourse).

- `https://discourse.osc.edu/t/ondemand-2-x-roadmap-engagement-goals/334`
- `https://discourse.osc.edu/t/ondemand-2-x-roadmap-accessibility-goals/295`
- Authentication is difficult to implement in one's institution.
- Message of the Day pulled from the cluster.
- If one authenticates as a cluster, then one ... TODO.

3.1. **interactive apps.** See also `https://www.chpc.utah.edu/documentation/software/ondemand.php#desktop`.

Cumas would like information for the nodes which are free, yet instead, Utah decided to make "Notchpeak shared partition" with limited queue time and resources, so that users can access an interactive desktop, with a mind for debugging.

[7] `https://www.colorado.edu/crdds/`

[8] `https://github.com/OSC/Open-OnDemand/blob/master/README.md`

Instead of starting an interactive desktop job, Cumas could also just start MATLAB (it's a slurm job, 1 node, 1 task, 2 cores) as an interactive job. Protects users from learning slurm commands.

- RStudio
- Jupyter Notebooks
- ANSYS
- VMD[9] is excluded, since the interactive applications are running on nodes with no GPUs.

3.2. **usage.** About 2 users each day.

3.3. **installation.** See Martin's instructions here: `https://www.ks.uiuc.edu/Research/vmd/`. Any tips?

1. Run two instances of the server, one for production, one for testing. > CHPC runs OOD on a VM which is mounting cluster file systems (needed to see users files, and SLURM commands). We have two VMs, one called ondemand.chpc.utah.edu is a production machine which we update only occasionally, the other is a testing VM called ondemand-test.chpc.utah.edu, where we experiment.
2. Setup with LDAP, with Keycloak.
3. Then setup `yaml` files for the clusters:

```
---
v2:
  metadata:
    title: "Ember"
  login:
    host: "ember.chpc.utah.edu"
  job:
    adapter: "slurm"
    cluster: "ember"
    bin: "/uufs/ember.arches/sys/pkg/slurm/std/bin"
  batch_connect:
    basic:
      set_host: "host=$(hostname -A | awk '{print $2}')"
    vnc:
      script_wrapper: |
        export PATH="/uufs/chpc.utah.edu/sys/installdir/turbovnc/std/opt/TurboVNC/bin:$PATH"
        export WEBSOCKIFY_CMD="/uufs/chpc.utah.edu/sys/installdir/websockify/0.8.0/bin/websockify"
        %s
      set_host: "host=$(hostname -A | awk '{print $2}')"
```

Here's logic for the group space:

```
User.new.groups.each do |group|
paths.concat Pathname.glob("/uufs/chpc.utah.edu/common/home/#{group.name}-group*")
end
```

3.4. **jupyter.** Fork from OSC, then modify, then push. See `https://github.com/CHPC-UofU/bc_osc_jupyter`. The `*.yml.erb` config files are read and filled by Ruby.

```
batch_connect:
  template: "basic"
script:
  native:
    - "-N"
    - "1" # specifies only 1 core for Jupyter Notebook
    - "-n"
    - "<%= num_cores %>"
```

[9] `https://www.ks.uiuc.edu/Research/vmd/`

4. How do I know you know what you say you know?

Software and data citations as part of the academic conversation, with Katie Mika.

4.1. **Joint declaration of data citation principles.** From `https://doi.org/10.25490/a97f-egyk`, *Force 11.*

1. Importance; Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. Credit and Attribution; Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. Evidence; In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. Unique Identification; A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. Access; Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. Persistence; Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe.
7. Specificity and Verifiability; Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. Interoperability and Flexibility; Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Six principles of software citation? Consider also

- *accessibility* includes READMEs for humans and machines, sufficient meta-data to determine whether or not to reuse software, data, etc.
- *specificity* include identification of and access to the specific version of the software, also the revision number, one's OS, other variants.

4.2. **helpful links.**

- Holdren's *Exit Memo*, 2017. `https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/OSTP%20Exit%20Memo.pdf`
  - See also the OSTP press release: "Scientific Integrity"[10]. The White House.
  - "Scientific Integrity: Fueling Innovation, Building Public Trust"[11]. whitehouse.gov. January 12, 2011.
- "Reproducible Research in Computational Science"[12]. Science.
- "Data Citation WG"[13]. RDA. April 18, 2019.

4.3. **questions.**

- What, if any, standards have been set for reproducibility of TB-scale data sets?

[10] `https://obamawhitehouse.archives.gov/administration/eop/ostp/library/scientificintegrity`
[11] `https://obamawhitehouse.archives.gov/blog/2010/12/17/scientific-integrity-fueling-innovation-building-public-trust`
[12] `https://science.sciencemag.org/content/334/6060/1226`
[13] `https://rd-alliance.org/groups/data-citation-wg.html`