Practical instructions Essentials Course Unsupervised learning day 1

General overview this week

Day 1
11:00-11:30: short introductory lecture
Until 14:00 : R refresher + data visualization practical
14:00-15:15 : lecture clustering
15:15-17:00: start practical

Day 2
11:00-11:45 lecture dimensionality reduction
11:45-15:15 continue practical (as far as you get, no problem if not finished!)
15:15-16:00 practical discussion
16:00-16:30 answering questions paper

Today

The goal of today's practical is to get your hands dirty doing some data visualization from 11:00-14:00 (excluding lunch). The final goal is to generate a plot (or a few, if you wish) on some dataset of your choice, and upload it and the code that generates it to Github complete with a readme.md that describes what was done and why (i.e. what insight you can gain from the plot). Finally, you should invite Adrien and Dieter to your repository. Below we offer: 1. Some resources to get up to speed with R, and specifically with data visualization using ggplot2. 2. Some sample datasets that you could use, and resources to find other ones. After this data visualization part, we move on to a lecture on clustering, followed by getting started on a practical (continued Thursday).

Data visualisation (until 14:00)
Make a plot (or more). Upload the plot, code that generates the plot, and a readme.md describing what the data represents and what you did to GitHub and invite us to your repo!

R refresher (optional as needed)

The first paragraph below gives you links to cloud-based R tutorials. Sometimes these can be slow. If that's the case, instead follow the tutorials in the second paragraph.

Here are some tutorials for working with R, how you define a function and what that is, how to use loops, how to manipulate data frames or tibbles (i.e. tables of data, like samples and their gene expression values), and how to plot things (using ggplot2). Be sure to go through at least the visualisation basics and programming basics (or that you are already at a level that you don't need to). Please follow the tutorials there that apply to you if you already have the basics down, or skip this entirely!

For the real basics, see here. For data visualization, we will turn to one of Hadley Wickham's great e-books. Hadley Wickham is the originator of the tidyverse suite of R packages and has written many free e-books on R. For data visualization, check out this part of the freely available R for Data Science e-book. Also look at exploratory data analysis (EDA) in the same book, getting to know your data by plotting (parts) of it and its distributions, etc. If you are the video type: here is an introduction to the tidyverse, here is a video of what you can do with Rmarkdown.

Tips of datasets to visualise

The Human Cell Atlas is an attempt to (single-cell) sequence all the human tissue types that exist and construct an atlas of it. There are many datasets, of healthy cells but also of various cancer cells. Of course the data can be *very* complex so be careful what you take on! Have a look at Seurat for single-cell data visualization (this will give you a sneak preview of clustering and dimensionality reduction already).

You can make an account on Zenodo and search for free open datasets. Another idea is to search on Kaggle, a Machine Learning platform that regularly hosts contests to find the best ML solutions to big problems. For instance, here is the Drosophila melanogaster genome along with some metadata on Kaggle.

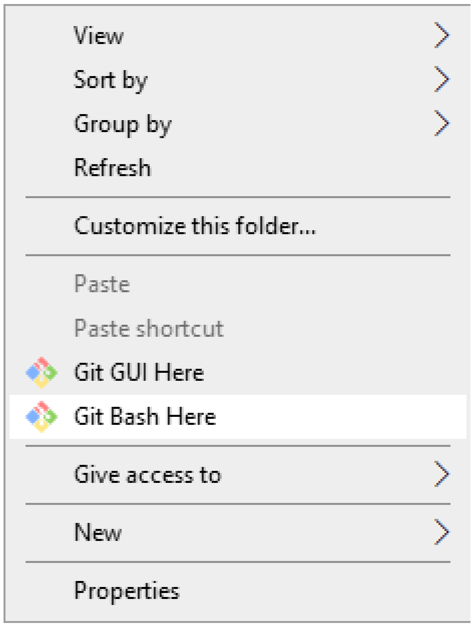
Finally, there's a dataset that's quite different and also very simple: it details different chicken breeds and their optimal characteristics. With this, the data is simple so be sure to pimp your ggplot2 plot to the max! Get it here.

When you are done making plots, upload your work to GitHub and invite Adrien and Dieter to your repository

Clustering practical start (from ~15:30, continued Thursday)

Getting materials

You first need to download the practical materials to your computer. They are in a repository on Github. A simple way to download this data is to install Git Bash (if on Windows), go to the folder where you want to download the materials, and open Git Bash there. You can open it in any folder by right clicking:



Then, simply type git clone "https://github.com/DieStok/PracticalEssentialsCourseUUSeptember212021.git". This should work also on Mac or Linux (you don't need Git Bash then, git is just in your terminal).

If that's not for you, you can also go here and download all the files in a zip.

Working on the practical

The practical is self-paced and should be self-explanatory. It starts with k-means clustering and showing how it comes about, while showcasing the usage of the ggplot2 and gganimate libraries. Then comes hierarchical clustering. That's it for the first part. The second part, which we deal with on Thursday, discusses k-means and PCA in the context of dimensionality and noise. What follows is a part about Eigenfaces. This is the most non-essential, though quite interesting, part of the practical. Skip it if you are running low on time. We end with a comparison of PCA, t-SNE and UMAP on some data (where you are also required to watch videos about t-SNE and UMAP) and **optionally** trying to recreate a paper figure subpanel from actual published RNAseq data.

Please go through the practical individually, but be sure to discuss, compare, and contrast with your neighbours. Please help others who might have had less exposure to R if you think you can. If there are any questions, please raise your hand and we will diligently come over to exterminate any and all questions.

Good luck!