Learning and Teaching Forum 2018

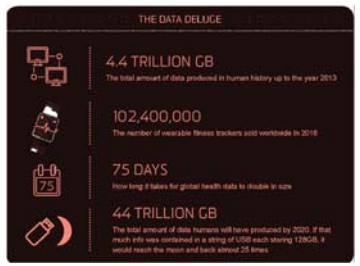
Learning and teaching health data science using real-life data science skills

Sanja Lujic[1], Oscar Perez Concha[1], Andrew Blance[1], Timothy Churches[2], James Farrow[1], Amy Gibson[1], Mark Hanly[1], Maarit Laaksonen[1]

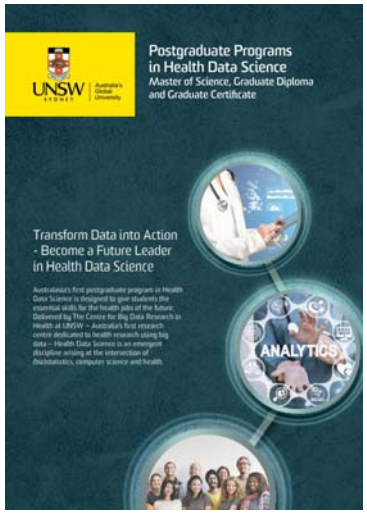
[1]Centre for Big Data Research in Health

[2] UNSW Medicine South Western Sydney Clinical School

Health Data Science @ UNSW

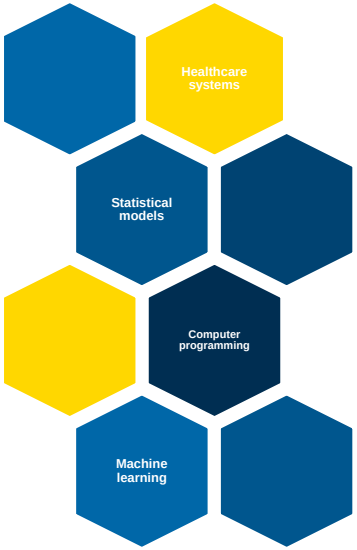


Health Data Science @ UNSW



Challenge

How to teach and learn HDS using industry-standard tools?



Solutions

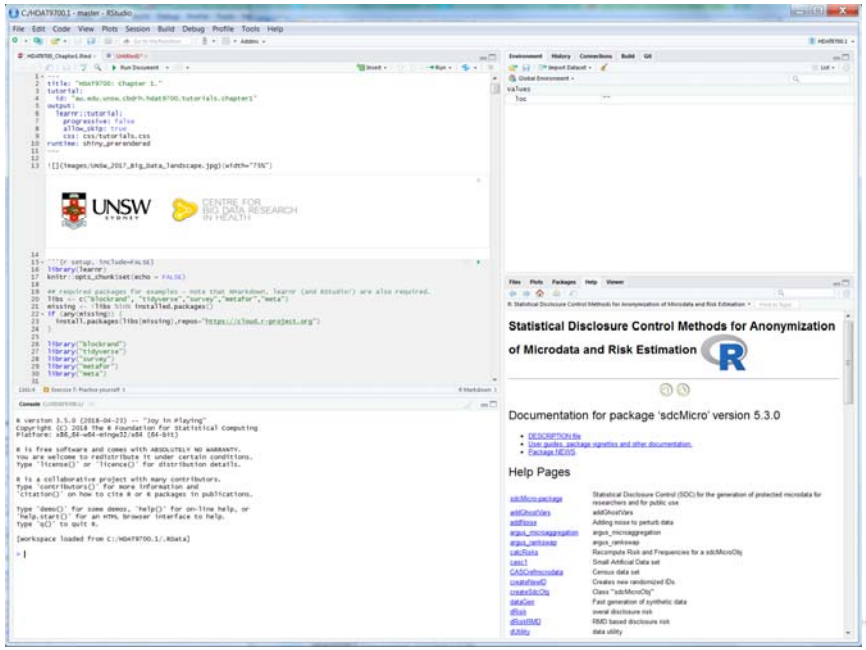
Learnr
(for RStudio)

Jupyter notebook
(for Python)

jupyter

GitHub

GitHub Classroom
Your course assignments on GitHub



Instructions to access interactive learnr tutorials

The content of this course will be delivered as learnr tutorials with supporting face-to-face sessions (2 X 3 hours per week). The learnr tutorials run as packages in RStudio. Please make sure you have the latest versions of R and RStudio installed. You will also need to have the packages `learnr` and `devtools` installed:

```
install.packages("learnr")
install.packages("devtools")
```

Chapter tutorials will be released just prior to the assigned face-to-face session. Please check this page to obtain the RStudio run commands for the chapter tutorials.

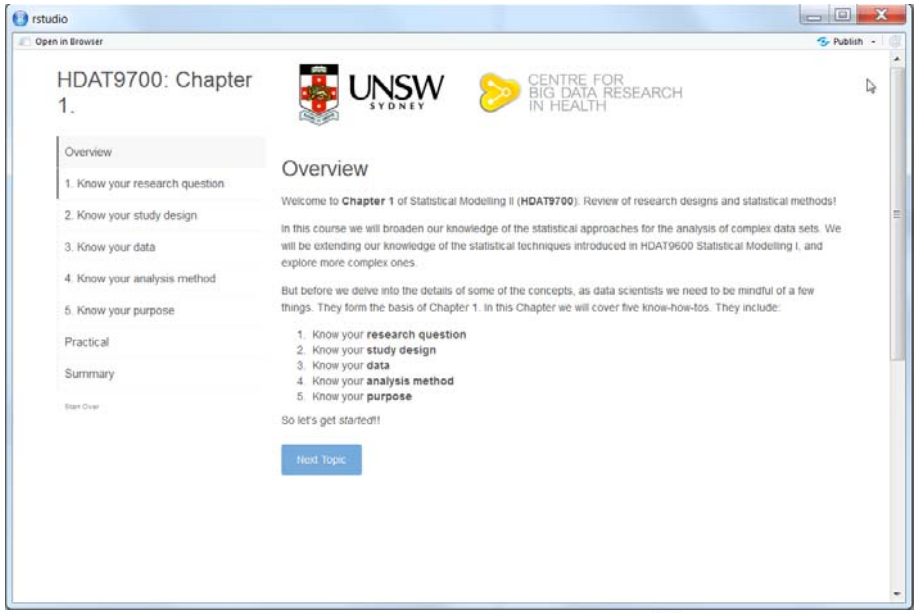
Chapter 1

To install the package enter:

```
devtools::install_github("cbdrh/HDAT9700.1", auth_token="f3583ae88f8d64391eac0851ad49505b48b34e7b")
```

To launch the tutorial enter:

```
cbdrh.hdat9700.1::tute()
```



Benefits

Run and modify code within the document

Embed exercises with hints and solutions

Quiz
(instant feedback)

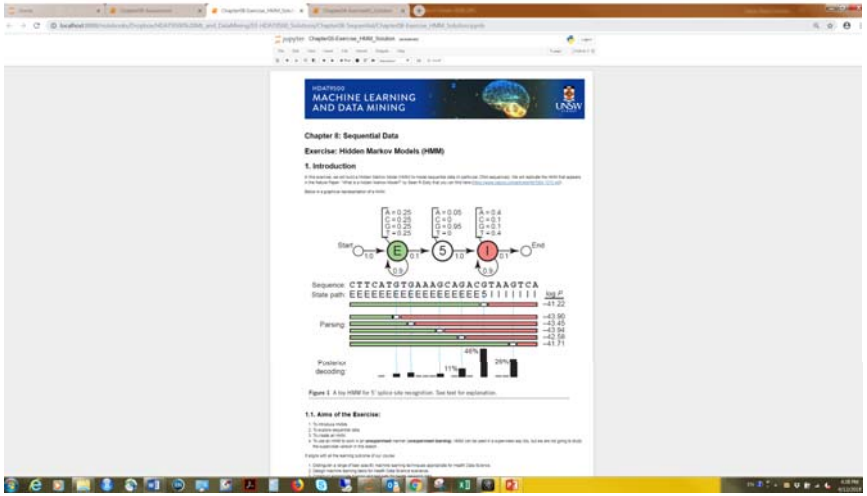
Videos

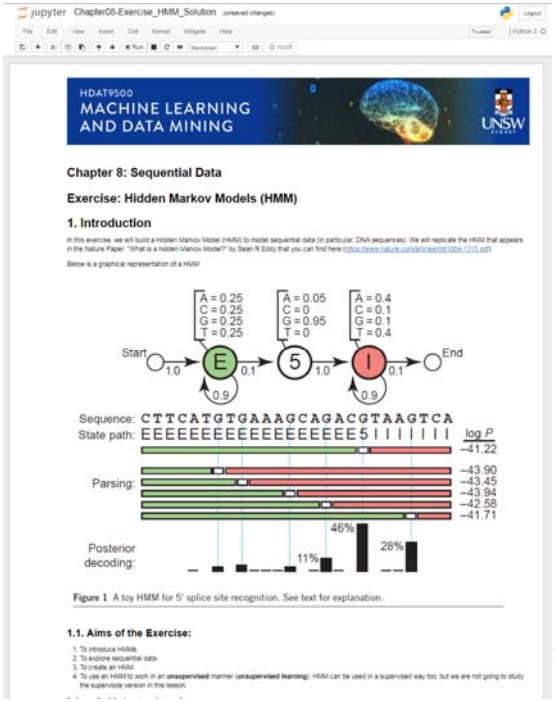
Graphics

Images

The end product: In students words

Jupyter Notebooks





HDAT9500
MACHINE LEARNING AND DATA MINING

Chapter 8: Sequential Data

Exercise: Hidden Markov Models (HMM)

1. Introduction

In this exercise, we will build a Hidden Markov Model (HMM) to model sequential data (in particular, DNA sequences). We will replicate the HMM that appears in the Nature Paper: "What is a hidden Markov Model?" by Sean R Eddy that you can find here (https://www.nature.com/articles/nbt1004-1315.pdf)

Below is a graphical representation of a HMM:

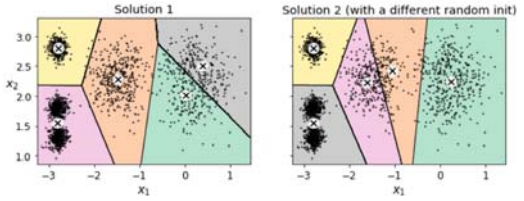
Figure 1 A toy HMM for 5' splice site recognition. See text for explanation.

1.1. Aims of the Exercise:

1. To introduce HMMs
2. To explore sequential data
3. To create an HMM
4. To use an HMM to work in an **unsupervised** manner (**unsupervised learning**). HMM can be used in a supervised way too, but we are not going to study the supervised version in this lesson.


```
In [24]: kmeans_rnd_init1 = KMeans(n_clusters=5, init="random", n_init=1,
                                   algorithm="full", random_state=11)
         kmeans_rnd_init2 = KMeans(n_clusters=5, init="random", n_init=1,
                                   algorithm="full", random_state=19)

         plot_clusterer_comparison(kmeans_rnd_init1, kmeans_rnd_init2, X,
                                   "Solution 1", "Solution 2 (with a different random init)")
         plt.show()
```

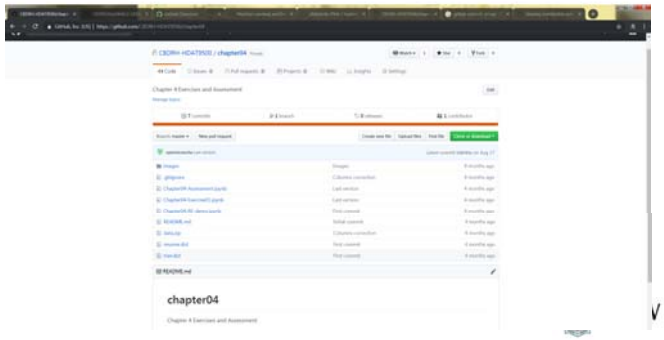


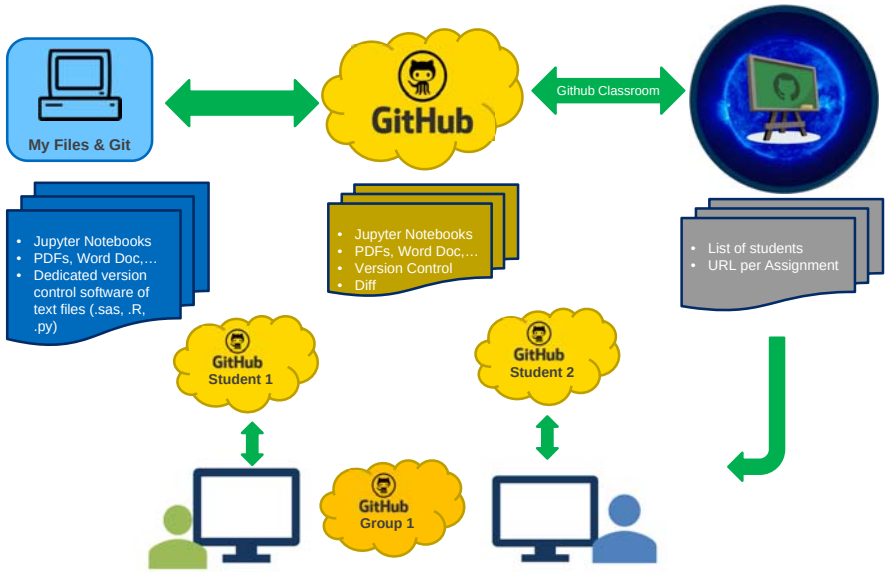
Inertia

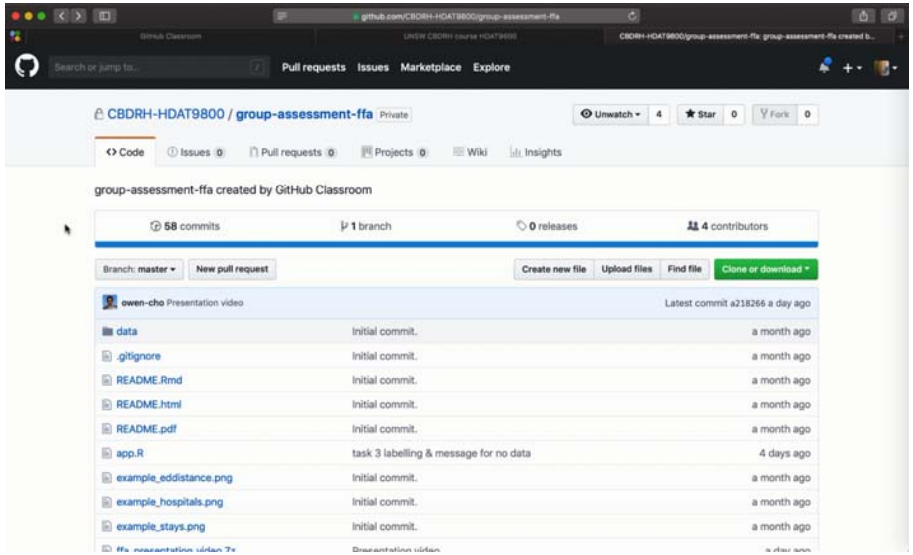
To select the best model, we will need a way to evaluate a K-Mean model's performance. Unfortunately, clustering is an unsupervised task, so we do not have the targets. But at least we can measure the distance between each instance and its centroid. This is the idea behind the _inertia_ metric:

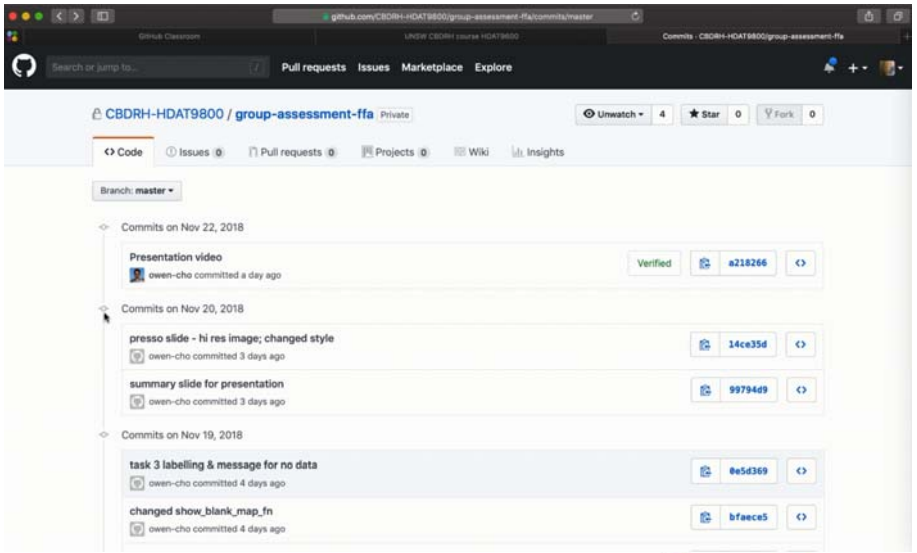
Git Hub

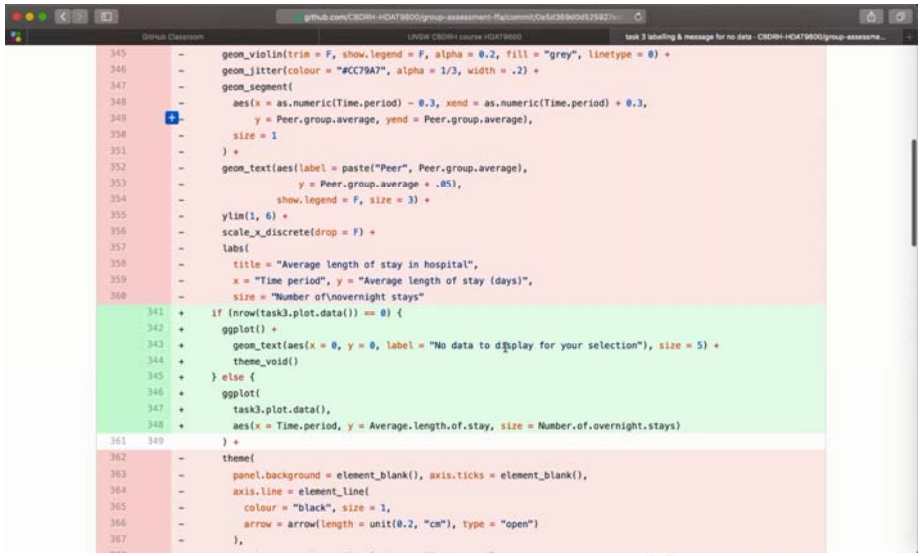
- Web-Based Hosting Service for Version Control
- Dedicated version control software
  - Text files (.sas, .R, .py)
  - Data, spreadsheets, images, PDFs, word processing docs











Thank you

Sanja Lujic (s.lujic@unsw.edu.au) and
Oscar Perez Concha (o.perezconcha@unsw.edu.au)

On behalf of the postgraduate programs in HDS teaching faculty:

Andrew Blance
Timothy Churches*
Oscar Perez Concha
James Farrow
Amy Gibson
Mark Hanly
Maarit Laaksonen
Sanja Lujic
Sebastiano Barbieri
Kylie-ann Mallitt
Andrea Schaffer

* *UNSW Medicine South Western Sydney Clinical School*