



Learning and Teaching Forum 2018

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

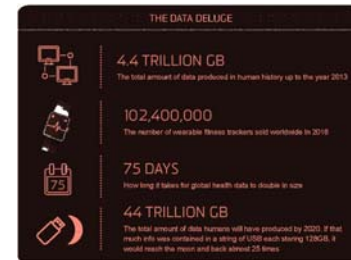
Sanja Lujic¹, Oscar Perez Concha¹, Andrew Blance¹, Timothy Churches², James Farrow¹, Amy Gibson¹, Mark Hanly¹, Maarit Laaksonen¹

¹Centre for Big Data Research in Health

²UNSW Medicine South Western Sydney Clinical School



Health Data Science @ UNSW



Health Data Science @ UNSW



Postgraduate Programs in Health Data Science
Master of Science, Graduate Diploma and Graduate Certificate

Transform Data into Action - Become a Future Leader in Health Data Science

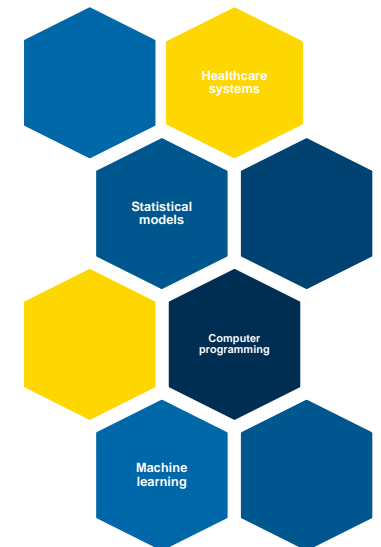
Australia's first postgraduate program in Health Data Science is designed by your students to ensure the essential skills for the health jobs of the future. Delivered by The Centre for Big Data Research in Health at UNSW - Australia's first research centre dedicated to health research using big data in Health Data Science as an emerging discipline arising at the intersection of biostatistics, computer science and health.



Challenge



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



Solutions

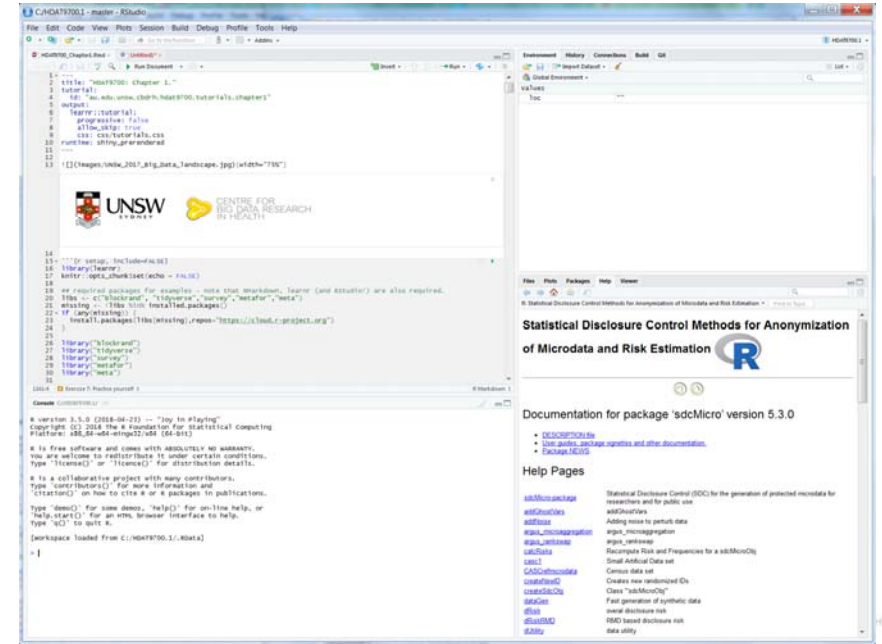
Learnr
(for RStudio)



Jupyter notebook
(for Python)



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.

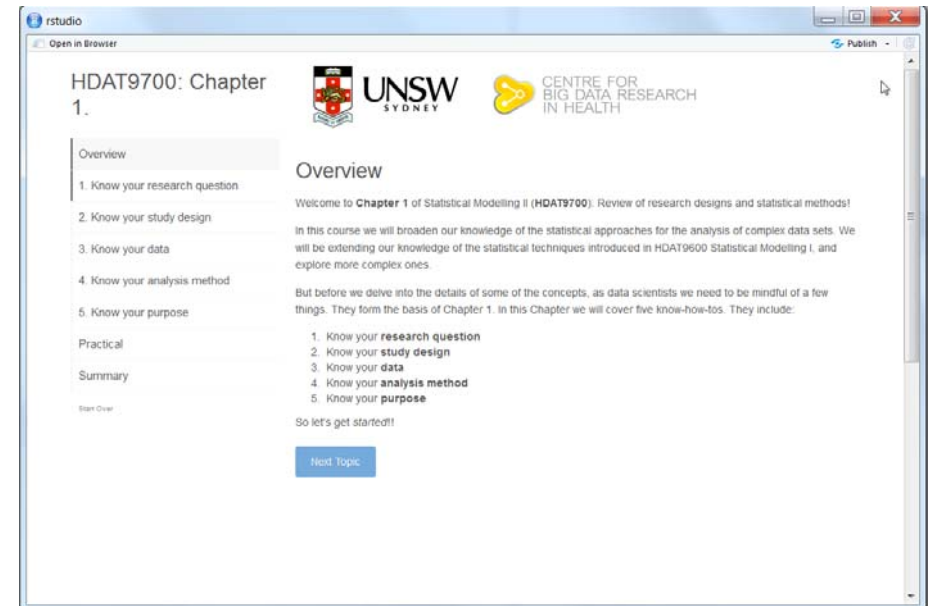
Chapter 1

To install the package enter:

```
devtools::install_github("cbrdh/HDAT9700.1", auth_token="f3583ae8f8d64391eac0851ad49585b48b34e7b")
```

To launch the tutorial enter:

```
cbrdh_hdat9700.1::tute()
```



Benefits

```


Example: Meta analysis
1 # Load RCG dataset
2 data(dat_bcg)
3 # Create total N for treatment and control groups in the RCG dataset
4 dat_bcg$tn <- dat_bcg$tpos + dat_bcg$tneg
5 dat_bcg$cn <- dat_bcg$cpos + dat_bcg$cneg
6 # Perform meta-analysis
7 meta_rr <- metabin(event.c = tpos, n.e = tn, event.c = cpos, n.c = cn,
8                   data = dat_bcg,
9                   sm = "RR",
10                  method = "Inverse",
11                  method.tau = "DL")
12
13 # Summary of the meta analysis results
14 summary(meta_rr)


Number of studies combined: k = 13

          RR      95%-CI      z p-value
Fixed effect model  0.6503 [0.6007; 0.7040] -10.42 < 0.0001
Random effects model 0.4896 [0.3449; 0.6950]  -4.00 < 0.0001

Quantifying heterogeneity:
tau^2 = 0.3089; I^2 = 3.56 [2.99; 4.34]; I^2 = 92.1% [88.34; 94.74]

Exercise: Stratified sample
1 # Compare average school size (~enroll) using SBS and stratified sample
2
3
    
```

Run and modify code within the document 

Embed exercises with hints and solutions 

Quiz (instant feedback)

Videos

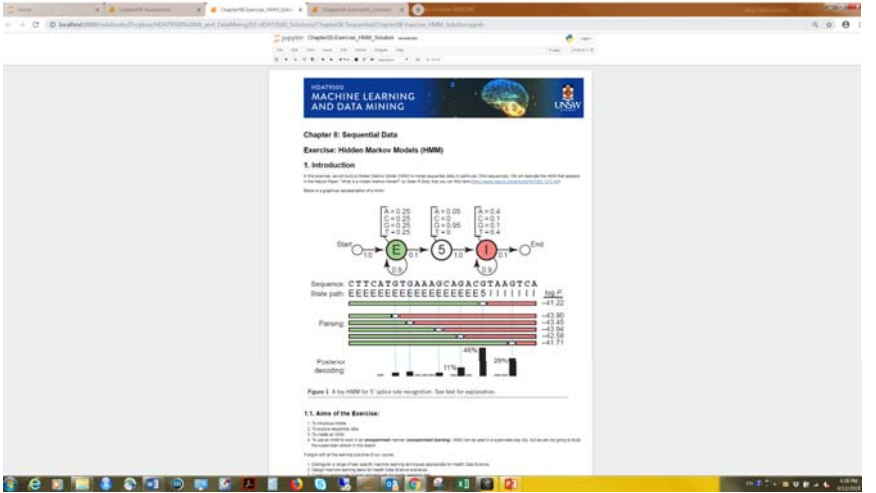
Graphics

Images

The end product: In students words



Jupyter Notebooks



jupyter Chapter08-Exercise_HMM_Solution [unsaved changes]

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 famous Paper "What is a Hidden Markov Model?" by Sean R. Eddy that you can find here <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1151140/>

Below is a graphical representation of a HMM:

Sequence: CTTTCATGTGAAAAGCAGACGTAAGTCA
State path: EEEEEEEEEEEEEEEEEEE5IIIIIIII

log P: -41.22, -43.90, -43.45, -43.94, -42.58, -41.71

Parsing: -43.90, -43.45, -43.94, -42.58, -41.71

Posterior decoding: 11%, 46%, 28%

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

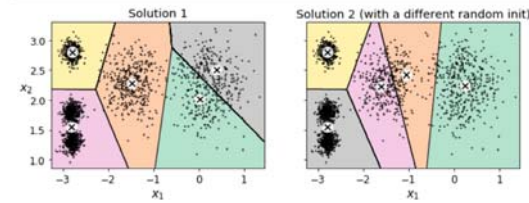
1.1. Aims of the Exercise:

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

CENTRE FOR BIO DATA RESEARCH IN HEALTH

```
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()
```

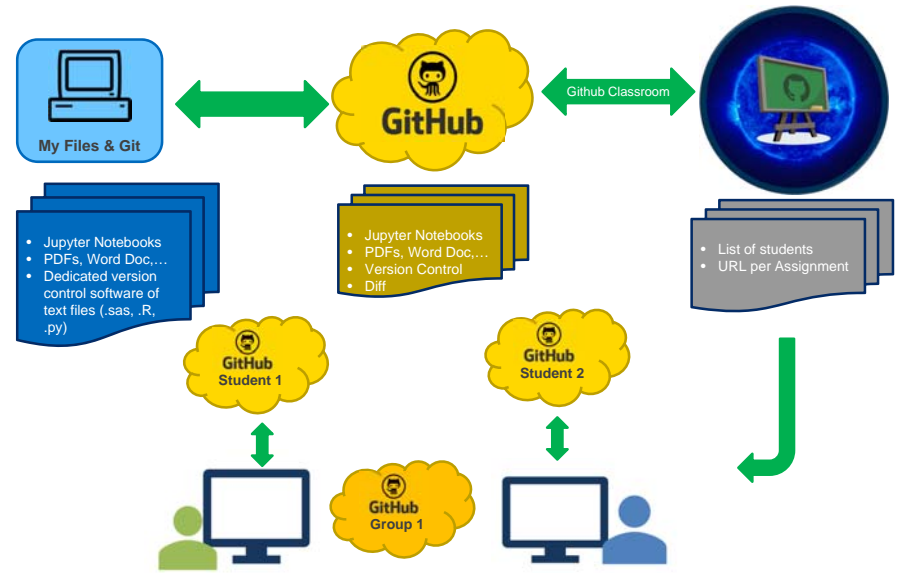
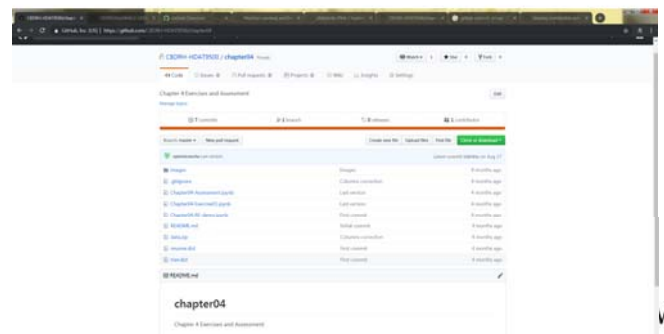


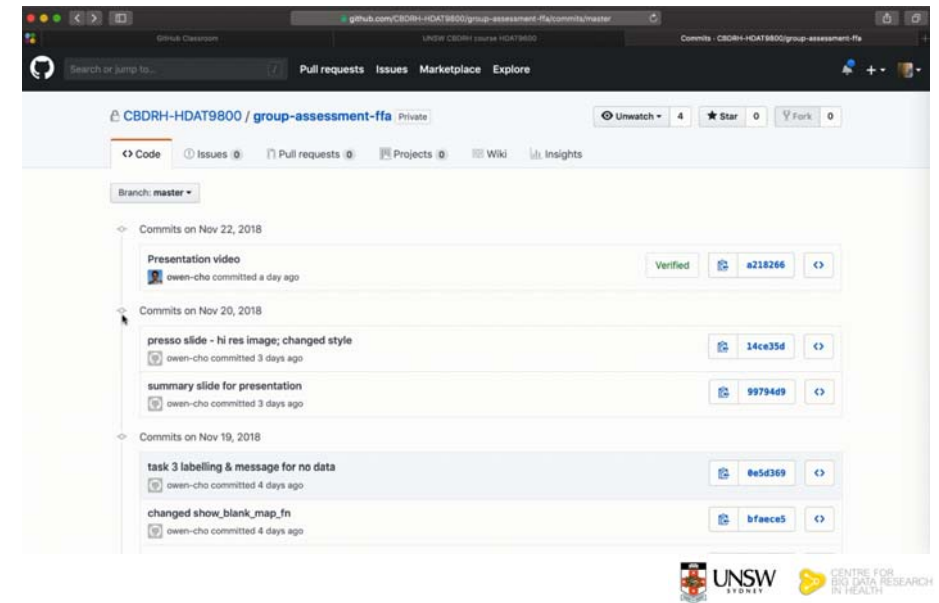
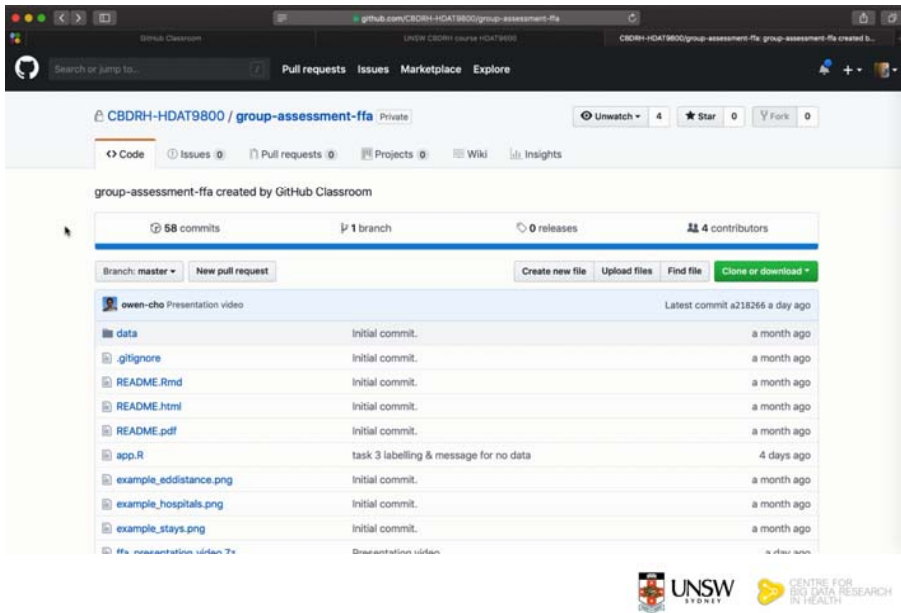
Inertia

To select the best model, we will need a way to evaluate a K-Means 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





```

345 - geon_violin(trim = F, show.legend = F, alpha = 0.2, fill = "grey", linetype = 0) +
346 - geon_jitter(colour = "#CC79A7", alpha = 1/3, width = .2) +
347 - geon_segment(
348 -   aes(x = as.numeric(Time.period) - 0.3, xend = as.numeric(Time.period) + 0.3,
349 -     y = Peer.group.average, yend = Peer.group.average),
350 -   size = 1
351 - ) +
352 - geon_text(aes(label = paste("Peer", Peer.group.average),
353 -   y = Peer.group.average + .05),
354 -   show.legend = F, size = 3) +
355 - ylim(1, 6) +
356 - scale_x_discrete(drop = F) +
357 - labs(
358 -   title = "Average length of stay in hospital",
359 -   x = "Time period", y = "Average length of stay (days)",
360 -   size = "Number of\overnight stays"
361 + if (nrow(task3.plot.data()) == 0) {
362 +   ggplot() +
363 +   geon_text(aes(x = 0, y = 0, label = "No data to display for your selection"), size = 5) +
364 +   theme_void()
365 + } else {
366 +   ggplot(
367 +     task3.plot.data(),
368 +     aes(x = Time.period, y = Average.length.of.stay, size = Number.of.overnight.stays)
369 +   ) +
370 - theme(
371 -   panel.background = element_blank(), axis.ticks = element_blank(),
372 -   axis.line = element_line(
373 -     colour = "black", size = 1,
374 -     arrow.length = unit(0.2, "cm"), type = "open")
375 - ),

```

UNSW SYDNEY | CENTRE FOR BIO DATA RESEARCH IN HEALTH

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

* UNSW Medicine South Western Sydney Clinical School