Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

All content is publicly available under the Creative Commons Attribution License: https://creativecommons.org/licenses/by/4.0/

Name / affiliation / email / twitter
Tim Dennis / ucla / timdennis@ucla.edu/ @jt14den
Martino Sorbaro / Edinburgh / martino.sorbaro@ed.ac.uk
Silvia Di Giorgio / ZB MED - Cologne / digiorgio@zbmed.de / @digiorgiosilvia
Benjamin Mummery / Hartree Centre, STFC Daresbury / benjamin.mummery@stfc.ac.uk / @BenjaminMummery
Ghada Alfattni / Manchester / ghada.alfattni@manchester.ac.uk / @Ghada_a_f
Marc Galland / University of Amsterdam / m.galland@uva.nl
Alexander Konovalov / University of St Andrews / alexander.konovalov@st-andrews.ac.uk (2nd half only)
Niall Beard / University of Manchester / niall.beard@manchester.ac.uk / twit: @niall_beard git: @njall
Akshay Khare / Indian Institute of Technology, Madras / akshaykhare.1997@gmail.com / linkedin:

Colin's repo: https://github.com/SCW-Aberystwyth/machine-learning-novice/blob/ccmcr19/Carpentry%20Connect%20Notes.md

Seattle Cycling habits https://jakevdp.github.io/blog/2015/07/23/learning-seattles-work-habits-from-bicycle-counts/
similar data for Edinbrugh

Titanic - lots of preprocessing

Kaggle competition datasets

built in sklearn data   (but they should learn how to load data from files too)

load_breast_cancer from sklearn

wine https://archive.ics.uci.edu/ml/datasets/Wine

Main areas:

- Supervised
    - Regression (and quality measures)

    - Classification (and quality measures)
- Unsupervised
    - Clustering (k-means, hierarchical and DBscan clustering)
    - Dimensionality reduction (PCA, t-SNE and UMAP)
- General
    - testing
    - cross-validation
    - ethics

Peter Steinbach, https://gcoe-dresden.de/author/steinba/Contact (person for the supervised machine learning material, especially for Machine Learning for Medical Imaging material---> we could ask if they are interested in joining us)

Personas:
Name
Discipline
Type(s) of data
What kind of analysis
Existing knowledge

Adam Smith is a PhD student studying in Geography and Remote Sensing. He is analysing satellite images looking at the expansion of urban areas in sub-Sharan Africa. He wants to measure how many buildings are visible in an image. He has a moderate amount of programming experience with some experience of using Excel, SPSS and simple Matlab code. His supervisor has suggested he look at neural networks or random forest as a method of doing this.

A postdoc in psychology, with a moderate amount (~1000) of data consisting of questionnaire responses.
Knows statistics relatively well and has used SPSS; only recently, he started learning R. He is interested in experimenting with simple machine learning models in order to link answers about family and social relations to self-reported behaviour and psychiatric conditions. Is eager to learn and very committed, but is very busy with research most of the time.
Is interested in both classification and regression, but would particularly appreciate interpretable models, so that they can be used in diagnosis.
The class should have a hands-on approach that will empower him to directly use his own data and judge by himself whether a specific ML model can make reliable predictions on them.

A PhD student in Plant Biology (Max) who has generated one transcriptomic dataset (12 samples x 25,000 genes). He need to extract candidate genes that are related to a phenotype of interest (let's say resistance to a pathogenic fungi). He might have access to other transcriptomic datasets of a similar type and size. He wants to combine them and train a ML model to classify sample phenotypes based on their gene expression.

Name: Jennifer
Discipline: A recent joinee of an RSE team who completed their PhD in Physics, but will now be applying software engineering solutions to a variety of domains over their career.
Types of Data: Unknown
What kind of Analysis: She wants to learn about the lanscape of machine learning solutions that she can apply it in her future work, but she does not yet have a specific use-case in mind.
Existing Knowledge: She has some knowledge of bayesian mathematics and probability from her PhD.

A PhD student in Biology who has generated some "omics" data sets. After preprocessing the data he/she needs to do a more advance analysis, for example using some unsupervised machine learning approaches to see if there are meaningful information coming from the data he/she acquired.

Concepts:

Pre processsing

- Cleaning Data
- Unsupervised Learning
-
    - Clustering
        - Kmeans
            - Methods to choose K
        - Dimesionality Reduction
            - PCA
            - TSNE
- Supervised Learning
    - *All models, objectives:*
        - *What it is;*
        - *when to use it and on what type of data;*
        - *how to evaluate the fit, over/underfitting;*
        - *computational complexity*
-
    - Regression
        - Linear
        - Polynomial
            - Overfitting/underfitting
            - Test sets (how and why)
-
    - Classification
        - Logistic regression
            - Over/underfitting can happen in regression too
            - Accuracy
                - Confusion Matrix
                - Precision
                - Recall
    -

        - Random Forest
        - Neural Networks
            - Evaluation
            - Cross Validation
-
    -
        -
            -
- Ethics

datasets:
 Edinburgh cycling: https://edinburghcyclehire.com/open-data
 temperature vs number of bikes on the road
 Daily climate data https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/data-formats

Unsupervised Learning
 - Understand what unsupervised learning is
 - Discriminate when unsupervised learning is appropriate to use

Clustering
 - Apply the k-means clustering to a dataset
 - Visualize the clusters
 - Assess the success of the resulting clusters
 - Be able to decide what the optimal value of k is

Dimensionality Reduction
 - Understand what the motivation for dimensionality reduction is
 - Use different methods to perform dimensionality reduction
 - Visualize the data
 - Transforming the dataset in order to make it fit for modelling

Further steps

Consider submitting the lesson to the CarpentriesLab: https://github.com/carpentrieslab

You can read further details about the purpose of the CarpentriesLab and submission instructions here:
 https://github.com/carpentrieslab/proposals