Lab session 3

Machine Learning for Behavioral Data (CS-421)

March 10, 2021

Today

- **08:15 08:30 Part I Introduction**
  - Feedback from tutorial 2
  - Introduction to Scikit-learn & SpeakUp
- **08:30 09:10 Part II Tutorial on Regression**
- **09:10 09:15 BREAK**
- **09:15 09:50 Part III Tutorial on Classification**
- **09:50 10:00 Part IV Homeworks and Next Steps**
  - Debrief on homework 1
  - Introduction to homework 2
  - Closing and feedback

Feedback last tutorial (n=2)

1 The volume and complexity of the material were appropriate, from 1 (totally disagree) to 5 (totally agree)

	Average rank
1 2 3 4 5	5.0

Responses	1	2	3	4	5	Total
	0	0	0	0	**2** (100%)	**2**

2 The lab session was well-paced and clear, from 1 (totally disagree) to 5 (totally agree)

	Average rank
1 2 3 4 5	4.5

Responses	1	2	3	4	5	Total
	0	0	0	**1** (50%)	**1** (50%)	**2**

3 Theory and hands-on activities were well balanced, from 1 (totally disagree) to 5 (totally agree)

	Average rank
1 2 3 4 5	4.5

Feedback last tutorial (n=1)

4 What do you like about this lab? What should be kept?

Respondent	Response
	I really like the fact that you take the time to walkthrough things and ask about our background. Also it's really cool that you provide us with useful and clear ressources like the cheatsheets.
Total responses to question	1/2

Other points from previous feedback forms:

- Keep SpeakUp and little quizzes.
- Detailed explanations and walkthrough guidance.
- Openness and time to answer questions.
- Well-documented notebooks.

Feedback last tutorial (n=1)

5 What do you dislike about this lab? What should be changed (and how)?

Respondent	Response
	Maybe sometimes you can go a little faster for certain things but it's ok

Other points from previous feedback forms:

- Hard to know what library or function to use and when. Provide supporting information.
- Provide a bit more time to complete exercise cells with nice results.
- Control the pace of the tutorial and provide more non-copy-and-pasting exercises.

Feedback last tutorial (n=2)

8 Any additional comment?

Respondent	Response
	I think maybe there should be more time for answering questions on the lab or homework. But I do like the tutorials and speakup so it's not easy...maybe balance more and have less homeworks/labs
	Would it be possible to record the lab sessions ? Also lecture 2 is still not uploaded and I couldn't attend it on Monday.
Total responses to question	2/2

Other points from previous feedback forms:

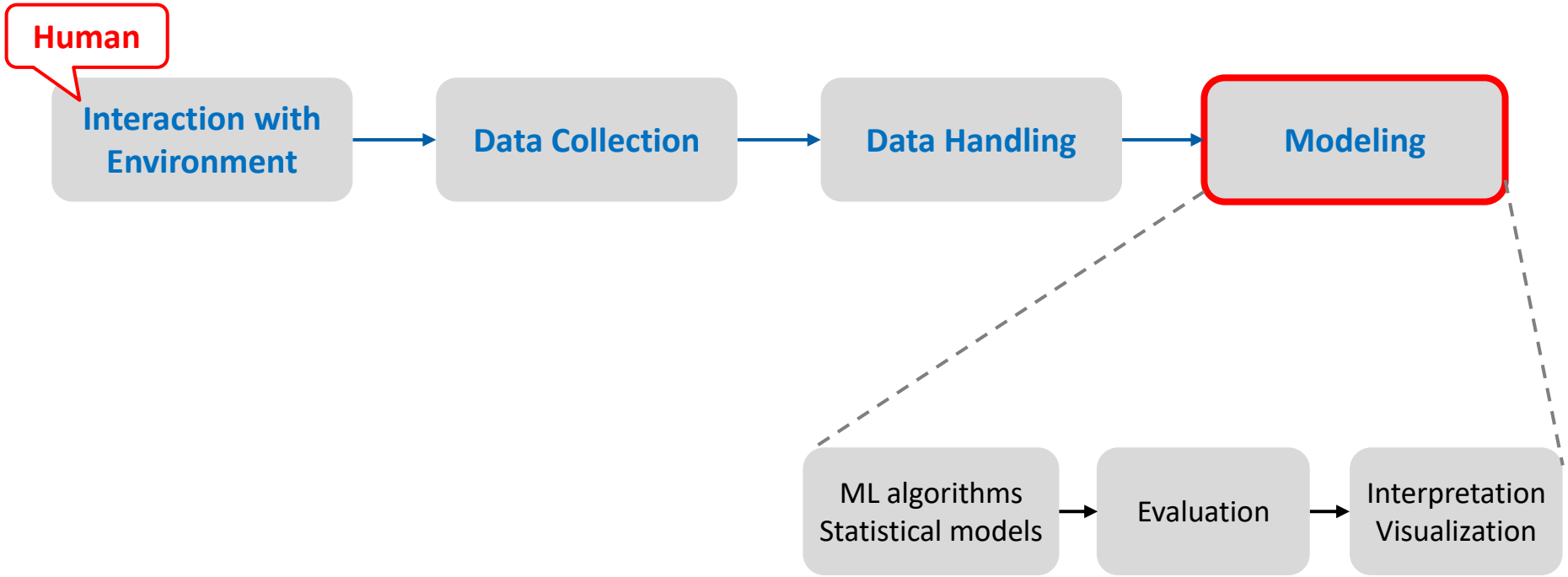
- Time of uploading for the material.
- Investigating the possibility of recording lab sessions.

Where we are

Week	Lecture	Lab Sessions	Project
1	Introduction	Tutorial	
2	Data Handling	Tutorial + **Homework**	
3	Regression & Classification	Tutorial + **Homework**	
4	Model Selection & Evaluation	Tutorial + **Homework**	Presentation of data sets and research questions
5	Latent Variable Models	Tutorial + **Homework**	M1: Preferences on team members and data sets
6	Unsupervised Learning	Tutorial + **Homework** + PO	
7	Spring Break	Spring Break	Spring Break

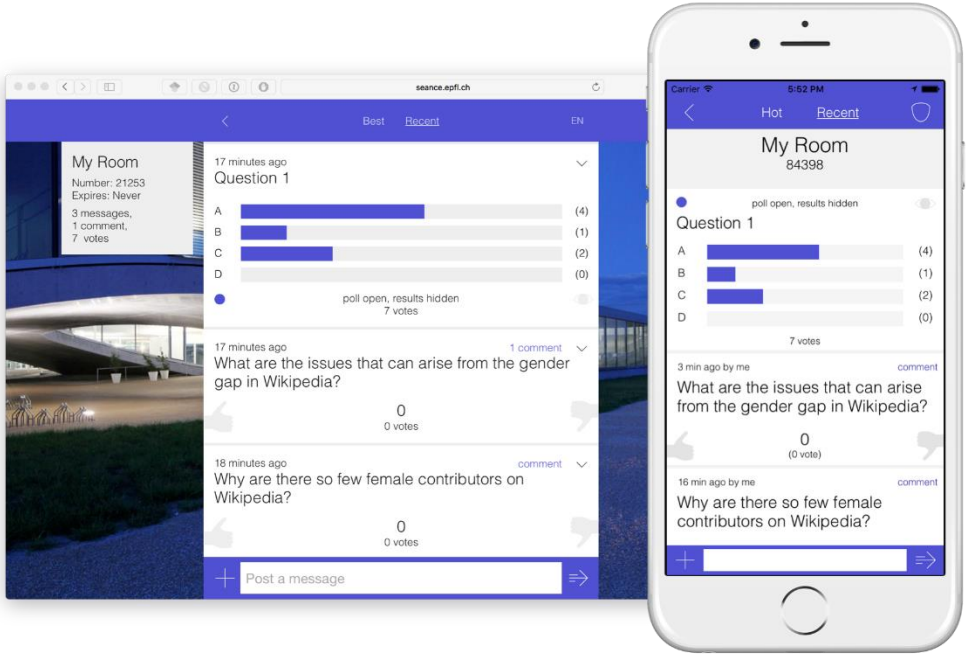
PO = project office hours

ML for Behavioral Data: Modeling



SpeakUp

- **Android / iOS:**
  http://speakup.info/
- **Web App:**
  https://web.speakup.info/
- **Room number:** 95305



Scikit-learn

SpeakUp #1: How do you feel about scikit-learn?

A: I've never heard of scikit-learn.

B: I'm **not confident at all** about using scikit-learn.

C: I'm **slightly confident** about using scikit-learn.

D: I'm **fairly confident** about using scikit-learn.

E: I'm **very confident** about using scikit-learn.

Quiz

SpeakUp #2:

What is the default score in scikit-learn, when using a classifier?

A: Recall

B: ROC-AUC

C: Accuracy

D: Precision

Quiz

Speakup #3:

Which classifier's method should you call to get the class predictions?

A: predict(X, y)

B: predict(X)

C: predict_proba(X, y)

D: predict_proba(X)

Quiz

Speakup #4:
Which command is used to train a model in scikit-learn?

A: train

B: fit_transform

C: fit

D: transform

Quiz

Speakup #5:
Which line of code returns X with values scaled between 0 and 1?

A: MinMaxScaler().fit(X)

B: MinMaxScaler().fit(X, y)

C: StandardScaler().fit_transform(X)

D: MinMaxScaler().fit_transform(X)

Quiz

Speakup #6: Given an array X with 10,000 observations and 50 features and an array y with the label for each observation, supposed that we run the command:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.25)

What would be the dimensions of X_train, y_train, X_test, and y_test?

A: X_train:(7500,50) y_train:(7500,) X_test:(2500,50) y_test(2500,)

B: X_train:(7500,50) y_train:(2500,2) X_test:(7500,50) y_test(2500,2)

C: X_train:(10000,50) y_train:(10000,) X_test:(10000,50) y_test(10000,)

D: X_train:(2500,) y_train:(2500,50) X_test:(7500,) y_test(7500,50)

Quiz

Speakup #7: Suppose that we have run the line of code: *data = load_iris().* Identify the line of code below where an error will occur.

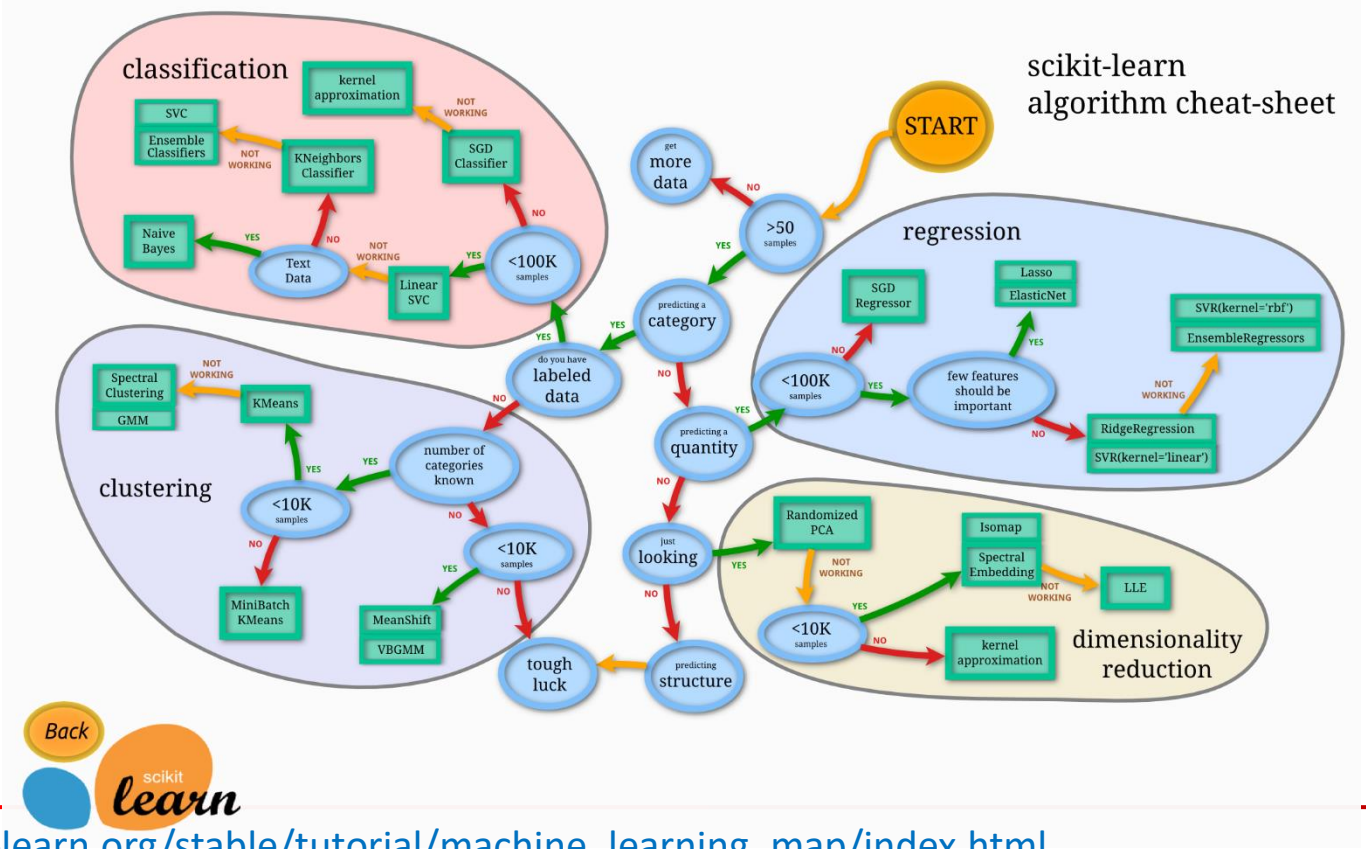
A: X_train, X_test, y_train, y_test = train_test_split(data['data'], data['target’])

B: knn = KNeighborsClassifier(n_neighbors=1)

C: knn.fit(X_train, y_train)

D: y_pred = knn.predict(X_test, y_test)

Scikit-learn in brief



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Scikit-learn resources

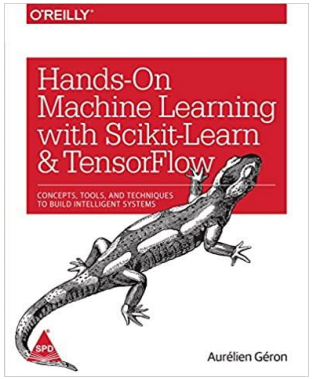
Choose from below according to your **level**:

- https://www.tutorialspoint.com/scikit_learn/index.htm
  - Walkthrough explanation (text + code)
  - From modelling to regressions, from classification to clustering

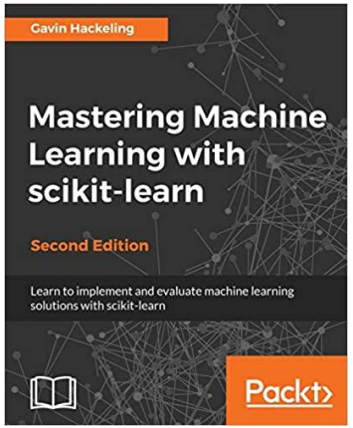
- https://scikit-learn.org/stable/tutorial/index.html
  - Extended collection of official tutorials
  - More examples with several datasets.

- https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Scikit_Learn_Cheat_Sheet_Python.pdf
  - Scikit-learn cheat sheet on one A4 page

Further readings



Géron, A. (2019). **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.** O'Reilly Media.



Hackeling, G. (2017). **Mastering Machine Learning with scikit-learn.** Packt Publishing.

Questions?

Tutorial 3.1 Agenda

- **10 mins** Fundamentals on Regression
  - Data generation, data split, data scaling, model development, model predictions, model effectiveness
- **20 mins** Hands-on Use Case on Regression
- **10 mins** Final Discussion

Tutorial 3.1 Hands on

We will use Noto, but feel free to use your own environment:

- Go to https://noto.epfl.ch/
- Login with your GASPAR
- If you have **NOT** already cloned the repository:
  - Go to Git → Clone → https://github.com/d-vet-ml4ed/mlbd
- Go to Git → Pull
- Go through Tutorials/Tutorial03/Notebooks/Regression.ipynb

Discussion and questions

5-MIN BREAK

We resume at 9:15am

Tutorial 3.2 Agenda

- **5 mins** Fundamentals on Classification
  - Data generation, data split, data scaling, model development, model predictions, model effectiveness
- **20 mins** Hands-on Use Case on Classification
- **10 mins** Final Discussion

Tutorial 3.2 Hands on

We will use Noto, but feel free to use your own environment:

- Go to https://noto.epfl.ch/
- Login with your GASPAR
- If you have **NOT** already cloned the repository:
  - Go to Git → Clone → https://github.com/d-vet-ml4ed/mlbd
- Go to Git → Pull
- Go through Tutorials/Tutorial03/Notebooks/Classification.ipynb

Discussion and questions

Debrief last homework

Homework 1: Data Handling

Introduction

In this homework, you will apply different data exploration, cleaning, and visualization techniques. It is very important to take some time to understand the data.

The homework is due **Mar 09, 2021 23:59 CET**. The fully-run notebook must be uploaded to your private GitHub private repository. We will only grade the cells with the heading

```
#### GRADED CELL ####
```

If you have any questions, feel free to use the Q&A forum in Moodle.

7 required exercises will be graded:

Section	Part	Required Function	Points
1	Data Exploration	get_feature_stats	20
2	Data Cleaning	handle_missing_values	20
2	Data Cleaning	handle_inconsistent_data	20
2	Data Cleaning	handle_skewness	20
3	Visualization	plot_correlation	10
3	Visualization	plot_grades	5
3	Visualization	Written interpretation	5
	Total Points		100

The solution of this homework will be uploaded on the public GitHub repository at:

https://github.com/d-vet-ml4ed/mlbd/tree/main/Homeworks/Homework01/Homework01-DataHandling_sol.ipynb

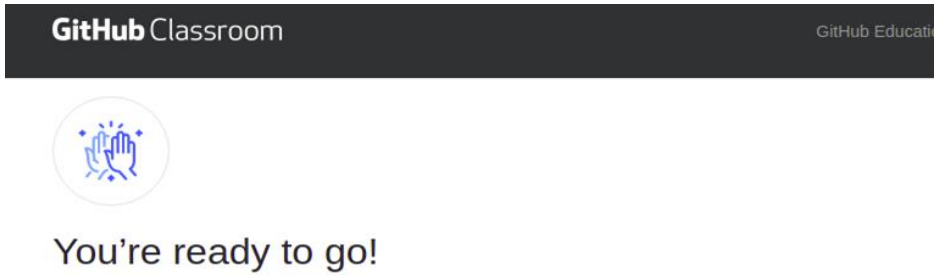
Questions?

Introduction to homework 2

The link to the homework will be available slightly after this lab session.
We will make an announcement in the forum, when the homework will be published.

Complete the exercises in the .ipynb notebook, commit your work, and **push the changes to GitHub**. TAs will pull the completed assignment after the deadline Mar 16, 2021 23:59 CET (any changes after the deadline will not be considered)

If you have any questions, feel free to use the Q&A forum in Moodle!



Important upcoming dates

- Mar 15, 2021 15:15 17:00
  - Lecture #4
- By Mar 16, 2021 23:59 CET
  - Submission Deadline for Homework #2
- Mar 17, 2021 08:15 10:00
  - Lab Session #4

Questions?