# Final Projects FDS

## A.Y. 2024/2025

11/11/2024
Profs. Indro Spinelli, Fabio Galasso
TAs Matteo Migliarini, Matteo Rampolla

# Calendar

Assignment 1: 26 sept - 28 oct (4 weeks)

Assignment 2: 28 oct - 29 nov (4 weeks)

Final Project: 29 nov– 29 dec (4 weeks)

    - Project announcement on 11 Nov

    - First presentations of ideas on 25 Nov

    - Final project presentations on 16 Dec

# Final Project

Start date: 25 Nov

    • Now, collect ideas!

 First presentation: 27 Nov

    • 3 mins incl Q&A, 3 slides: task / setup definition

Final presentation: 16 Dec

    • 3 mins + 1 min Q&A: Progress and presentation evaluation

Written report submitted by 27 Dec

    • Report evaluation

    • Reports to indicate role of each member in the work

# Suggestion for First-Presentation

 Slide 1 – Task and motivation

- Task statement and definitions

- Motivation & Related work

Slide 2 – Models, tools

- Tentative material and methods (research papers, github repos)

- Investigation: Feature, model, etc.

Slide 3 – Analysis

- Benchmark: Evaluation dataset and metric

# Project types

**Application project**

Pick an application that is of interest to you

Explore how to apply learning algorithms to solve it implementing your own solution

Focus is not on the best results but on the deep understanding of how to set up and solve a machine learning problem.

# Project types

**Analytical project**

Choose one or more existing projects/algorithms on a topic which you like

Reproduce their results

Using their code is fine, but cite the source -- see my note on plagiarism

Analyze the approach and the results

If you leveraged existing source code, then conduct ablation studies, propose

modifications and evaluate how these affect the results

# Plagiarism

**Watch out for plagiarism**

Plagiarism is severely prohibited and would invalidate your project

Leveraging resources is fine, BUT acknowledge the source

Specify your contribution in detail

# Goal

Choose a machine learning and/or computer vision application

E.g. from existing Kaggle competition, or from literature

Choose an existing dataset (acquiring one is not recommended): Kaggle, https://www.visualdata.io/ https://github.com/caesar0301/awesome-public-datasets https://datasetsearch.research.google.com/

Choose a task: Regression, classification, generation, dimensionality reduction

# Goal

Choose a task: Regression, classification, generation, dimensionality reduction

Apply methods to the task, present the analysis of your results

Analyze an existing project

Modify an existing project

Implement from scratch your solutions

# Goal

Choose a task: Regression, classification, generation, dimensionality reduction

Apply methods to the task, present the analysis of your results

Analyze an existing project

Modify an existing project

Implement from scratch your solutions

# Suggestion

Have each team member sketch 20 ideas before meeting

Filter out list by doing quick Google searches

There may be an existing GitHub for your idea (ok to leverage it, but cite it)

Pay attention to how long the training takes and how much data you need

Ask yourselves: are there little tweaks/experiments that haven't been done yet?

Can you extend the idea e.g. to a new application?

Which of your initial ideas makes the best story to tell?

Which of those lets you obtain best illustrative pictures?

# How to read

You can find information on blogs, papers, journals, Github repos, websites that summarize or explain papers!

If you consider papers:

- Don't read all of them (at least at the beginning)
- Look at the figures and captions before anything
- First pass reading order: Abstract, Methods, Conclusion, Results

You need to find something interesting about the chosen topic, not to review the entire literature!

# Avoid this

Reproduce a source without your contribution (nor comments, nor analysis)

Team starts late:

- Just instance and draft of code up by milestone
- Didn't hyperparameter search much

A few standard graphs: loss curves, accuracy chart, simple architecture graphics

Your report is not clear. As a data scientist, illustrating your ideas, solutions and analysis is part of project

Conclusion doesn't have much to say about the task besides that it didn't work

# Aim for this

Workflow set-up configured ASAP!!

- Creative hypothesis is being tested
- Have running code
- Have a benchmark to compare your results to

• Mixing knowledge from different aspects in ML

• Have a meaningful graphic

Conclusion and Results should teach me something

- ++interactive demo
- ++novel / impressive engineering feat
- ++good results

# Choose a method

A method discussed in class:

- Digital image processing algorithm, linear regression, optimization via gradient descent or Newton's method, logistic regression, ML analysis via bias/variance, Neural Networks (supervised and unsupervised)
- Explore other methods
- Ask us ;)

# Choose a task

Classification

- https://www.kaggle.com/c/titanic
- https://www.kaggle.com/c/digit-recognizer
- https://www.kaggle.com/c/nlp-getting-started
- https://www.kaggle.com/zalando-research/fashionmnist
- https://www.kaggle.com/c/kobe-bryant-shot-selection/data
- https://www.kaggle.com/kazanova/sentiment140

Regression

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- https://www.kaggle.com/c/how-much-did-it-rain-ii/data
- https://www.kaggle.com/mirichoi0218/insurance

# Choose a dataset

Choose a dataset! Pick one from the online lists of datasets

- https://github.com/caesar0301/awesome-public-datasets
- https://www.visualdata.io/
- http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase
- https://www.ind-dataset.com/
- https://github.com/renmengye/few-shot-ssl-public
- http://www.cvpapers.com/datasets.html
- http://riemenschneider.hayko.at/vision/dataset/
- http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase
- https://www.kaggle.com/datasets

Or find one of your choice!

# EEEASY examples

Titanic survivors, a binary classification with ML

- [https://www.kaggle.com/c/titanic](https://www.kaggle.com/c/titanic)
- Predict whether a person survived or not looking at some personal information
- 12 features representing the age, gender, ticket cost, ...

Digit Recognition (MNIST)

- Recognize the digit from handwritten digit images
- Multi-class classification problem (balanced)
- 28x28 images

# Final project presentation goals

Have code up and running

Data source explained correctly

- Give the true train/test/val split
- Number training examples
- Where you got the data

What Github repo, or other code you're considering

- Ran baseline model have results. Points off for no model running, no results

Data pipeline is in place and explained clearly

Discussion of results, including surprising findings

Reasonable literature review (3+ sources)

# Final project presentation goals

Have code up and running

Data source explained correctly

- Give the true train/test/val split
- Number training examples
- Where you got the data

What Github repo, or other code you're considering

- Ran baseline model have results. Points off for no model running, no results

Data pipeline is in place and explained clearly

Discussion of results, including surprising findings

Reasonable literature review (3+ sources)

# Project Report

• 1-2 page progress report. Not super formal. Notebooks are ok.

The holy structure!

Title, Abstract, Introduction, Related work, Proposed method explained, Dataset and Benchmark, Experimental results, Conclusions and Future work,References

# For a good project report

5 W's

- What? (a problem)
- Why? (motivation)
- How? (proposed strategy)
- Where? (dataset and benchmark)
- Who? (team assignments)

It is desired.. your considerations on:

- Influence of parameter and method choice
- Results: what is expected and what is surprising.. not just numbers!

Observations must be substantiated by results or references

# What I cannot create I do not understand Feynman



# Questions?