Introduction to Final Projects

Fundamental of Data Science
25 November 2020
Prof. Fabio Galasso,
Guido D'Amely, Alessandro Flaborea, Luca Franco,
Muhammad Rameez Ur Rahman, Alessio Sampieri

SAPIENZA
UNIVERSITÀ DI ROMA

Assignments and final project

- Calendar
  - Assignment 1: 23 sept - 20 oct (4 weeks)
  - Assignment 2: 21 oct – 1 dec (4 weeks)
  - Final Project: 25 nov – 26 dec (4 weeks)

Logistics

- Start date: 25 Nov
  - ***Now***, collect ideas!

- First presentation: 6 Dec
  - 3 mins incl Q&A, 3 slides: task / setup definition
- Second presentations: 16 Dec
  - 3 mins + 2 min Q&A: Progress and presentation evaluation
- Written report submitted by 26 Dec
  - Report evaluation

- *Reports to indicate role of each member in the work*

Suggestion for First-Presentation Slides

- Slide 1 – Task and motivation
  - Task statement and definitions
  - Motivation
  - Related work
- Slide 2 – Models, tools
  - Tentative material and methods
    - From coursework, open source and research code
  - Investigation
    - Feature, model, etc.
- Slide 3 – Analysis
  - Benchmark
    - Evaluation dataset and metric

Final Projects
Topic and development

SAPIENZA
UNIVERSITÀ DI ROMA

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
- 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

Project goal

- Choose a machine learning and/or computer vision application
  - For example well known problems from kaggle, literature, ...
- Choose a dataset (i.e. an existing.. acquiring one is lots of efforts)
  - From Kaggle
  - From some other online resource
  - E.g.
    - https://www.visualdata.io/
    - https://github.com/caesar0301/awesome-public-datasets
    - https://datasetsearch.research.google.com/
- Choose a task
  - Regression, classification, density estimation, clustering, …
- Apply methods to the task, present the analysis of your results
  - Analyze an existing project
  - Modify an existing project
  - Implement from scratch your solutions

Project goal suggestions

- 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 the models require
- Ask yourselves: are there little tweaks and/or 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?

Gather information on your idea

- 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
    - Results
    - Conclusion
- You need to find something interesting about the chosen topic, not to review the entire literature

Try to avoid this scenario

- 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
- Have running code
- Have a benchmark to compare your results to
- Creative hypothesis is being tested
- Mixing knowledge from different aspects in ML
- Have a meaningful graphic (pretty or info rich)
- Conclusion and Results teach me something
- ++interactive demo
- ++novel / impressive engineering feat
- ++good results

Choose a method

- A method discussed in class
  - E.g. digital image processing algorithm, linear regression, optimization via gradient descent or Newton's method, logistic regression, Gaussian discriminant analysis, Naïve Bayes, SVM, ML analysis via bias/variance, Neural Networks, clustering, dimensionality reduction
- 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/c/word2vec-nlp-tutorial/overview/part-1-for-beginners-bag-of-words
  - https://www.kaggle.com/zalando-research/fashionmnist
  - https://www.kaggle.com/c/kobe-bryant-shot-selection/data
  - https://www.kaggle.com/kazanova/sentiment140

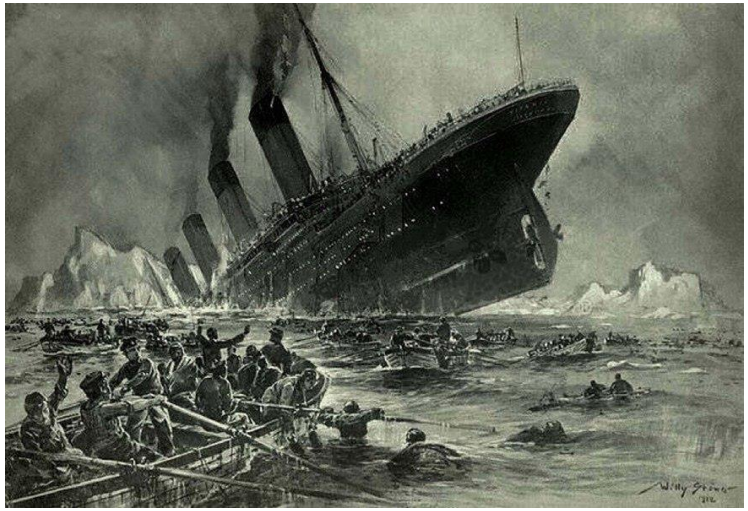
- Regression
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques
  - https://www.kaggle.com/c/restaurant-revenue-prediction/overview/evaluation
  - https://www.kaggle.com/c/how-much-did-it-rain-ii/data
  - https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather/overview
  - https://www.kaggle.com/mirichoi0218/insurance

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.htm
  - 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.htm
  - https://www.kaggle.com/datasets
- Or find one of your choice

Example tasks

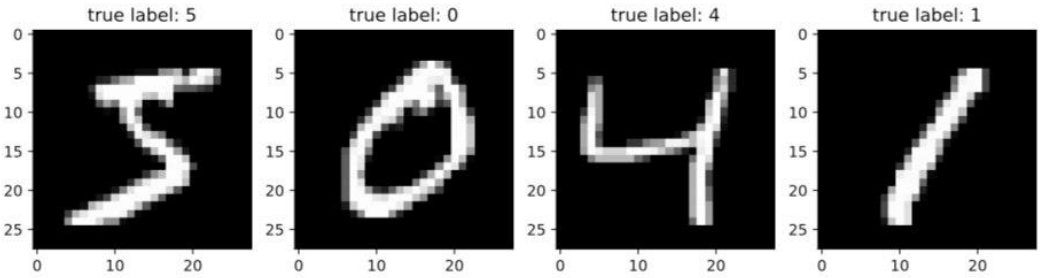
- Titanic survivors, a binary classification with ML
  - https://www.kaggle.com/c/titanic
- Predict whether a person survived or not looking at some personal information
  - 12 features representing the age, gender, ticket cost, …



PassengerId	Survived	Pclass	Name	Sex
1	0	3	Braund, Mr. Owen Harris	male
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
3	1	3	Heikkinen, Miss. Laina	female
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
5	0	3	Allen, Mr. William Henry	male
6	0	3	Moran, Mr. James	male
7	0	1	McCarthy, Mr.	male

Example datasets

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



Example datasets

- House Pricing
  - Predict the cost of a house from information about its architecture, type of suburbs, square meters, ...
  - Regression problem
  - Prepare the data for your model, tackle missing values and categorical features

Id	MSSubClass	MSZoning	LotFrontage
1 – 1460	20 – 190	RL 79% RM 15% Other (91) 6%	NA 18% 60 10% Other (1058) 72%
1	60	RL	65
2	20	RL	80
3	60	RL	68
4	70	RL	60
5	60	RL	84
6	50	RL	85

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
- Suggested sections
  - 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

Questions?

Thank you

Acknowledgements: some slides and material from Bernt Schiele, Mario Fritz, Misha Andriluka, Fei-Fei, Justin Johnson, Serena Yeung, Andrew Ng, Kian Katanforoosh, Pedro Pablo Garzon

SAPIENZA
UNIVERSITÀ DI ROMA