PROGRAMMING IN PYTHON II

Project Design and Outline

Andreas Schörgenhumer
Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

Contact

Andreas Schörgenhumer

Institute for Machine Learning
Johannes Kepler University
Altenberger Str. 69
A-4040 Linz

E-Mail: schoergenhumer@ml.jku.at

Write mails only for personal questions

Institute ML Homepage

Copyright Statement

MOTIVATION

Motivation (2)

- When designing or implementing an ML project, you have to consider and constantly reevaluate multiple aspects

Motivation (2)

- When designing or implementing an ML project, you have to consider and constantly reevaluate multiple aspects
- Spoiler alert: The choice of the ML method itself is only one aspect of many

Motivation (2)

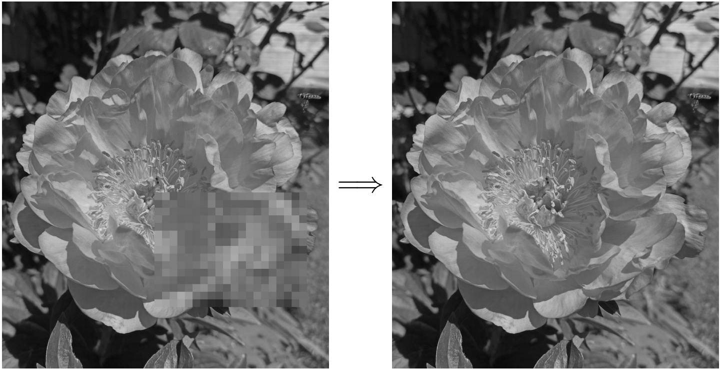
- We will go through the outline of the project design

Motivation (2)

- We will go through the outline of the project design
- We will cover the details during the semester

Motivation (2)

- We will go through the outline of the project design
- We will cover the details during the semester
- We will use our ML project as example



OVERVIEW

Project Design

- Common important aspects:

Project Design

- Common important aspects:
    1. What is the **project goal**?

Project Design

- Common important aspects:
    1. What is the **project goal**?
    2. What **data** do you have? What data do you need? What does the data look like?

Project Design

- Common important aspects:
    1. What is the **project goal**?
    2. What **data** do you have? What data do you need? What does the data look like?
    3. What **hardware/software** do you have? What hardware/software could you have?

Project Design

- Common important aspects:
  1. What is the **project goal**?
  2. What **data** do you have? What data do you need? What does the data look like?
  3. What **hardware/software** do you have? What hardware/software could you have?
  4. What **ML method(s)** should you use?

Project Design

- Common important aspects:
  1. What is the **project goal**?
  2. What **data** do you have? What data do you need? What does the data look like?
  3. What **hardware/software** do you have? What hardware/software could you have?
  4. What **ML method(s)** should you use?
  5. How to **evaluate** the methods/models?

Project Design

- Common important aspects:
  1. What is the **project goal**?
  2. What **data** do you have? What data do you need? What does the data look like?
  3. What **hardware/software** do you have? What hardware/software could you have?
  4. What **ML method(s)** should you use?
  5. How to **evaluate** the methods/models?
- **There is no one-fits-all solution!** Specific tasks require specific considerations!

PROJECT GOAL

Project Goal

What is the project Goal?

Project Goal

What is the project Goal?

- Very important aspect and often overlooked

Project Goal

What is the project Goal?

- Very important aspect and often overlooked
- Requires communication with people from different fields, including management

Project Goal

What is the project Goal?

- Very important aspect and often overlooked
- Requires communication with people from different fields, including management
- **Do not** make simplifications here! Make sure you are aware of the real (end) goal and communicate this!

Project Goal

What is the project Goal?

- Very important aspect and often overlooked
- Requires communication with people from different fields, including management
- **Do not** make simplifications here! Make sure you are aware of the real (end) goal and communicate this!
- Rinse and repeat to overcome language barriers

DATA

Data (1)

What data do you have? What data do you need? What does the data look like?

Data (1)

What data do you have? What data do you need? What does the data look like?

- Data is money. Big data is big money.

Data (1)

What data do you have? What data do you need? What does the data look like?

- Data is money. Big data is big money.
- Sometimes the goals will follow from sufficiently large existing data
  - Best case but rather rare (our hunger for data is only limited by computational restrictions!)

Data (1)

What data do you have? What data do you need? What does the data look like?

- Data is money. Big data is big money.
- Sometimes the goals will follow from sufficiently large existing data
  - Best case but rather rare (our hunger for data is only limited by computational restrictions!)
- Sometimes the goals will follow from existing but insufficiently large data
  - Common case
  - Has influence on choice of ML method
  - Allows for educated guesses at sufficiently large data size
  - Can be a starting point for collecting more data

Data (2)

- Sometimes the goals are not backed up by any data
  - Very tricky and potentially dangerous!
  - You would have to make guesses about how much and which data would be needed
  - You would have to make guesses about the ML method performance in advance
  - You will need to interface with the data collection process (first get small dataset, then collect more)
  - You might waste a lot of time and money

Data (2)

- Sometimes the goals are not backed up by any data
  - Very tricky and potentially dangerous!
  - You would have to make guesses about how much and which data would be needed
  - You would have to make guesses about the ML method performance in advance
  - You will need to interface with the data collection process (first get small dataset, then collect more)
  - You might waste a lot of time and money
- A dataset might be unsuitable for your purposes
  - Biases, artifacts, labeling errors, . . .

Data (3)

- Make the most of your data

Data (3)

- Make the most of your data
  - Talk to experts in the field of application/read up on the topic

Data (3)

- Make the most of your data
  - Talk to experts in the field of application/read up on the topic
  - Perform analysis of the data (e.g., clustering) and look for possible issues (e.g., biases, batch-effects)

Data (3)

- Make the most of your data
  - Talk to experts in the field of application/read up on the topic
  - Perform analysis of the data (e.g., clustering) and look for possible issues (e.g., biases, batch-effects)
  - Check if there is auxiliary data available
    - Pre-training on similar data, unused sorted-out data, data that is not suitable for training but for evaluation, . . .

Data (3)

- Make the most of your data
  - Talk to experts in the field of application/read up on the topic
  - Perform analysis of the data (e.g., clustering) and look for possible issues (e.g., biases, batch-effects)
  - Check if there is auxiliary data available
    - Pre-training on similar data, unused sorted-out data, data that is not suitable for training but for evaluation, . . .
  - Perform data preprocessing and augmentation
    - Normalization, oversampling, cross-validation splits, data augmentation, . . .

HARDWARE/SOFTWARE

Hardware/Software

What hardware/software do you have? What hardware/software could you have?

Hardware/Software

What hardware/software do you have? What hardware/software could you have?

- CPU, GPU or TPU based?

Hardware/Software

What hardware/software do you have? What hardware/software could you have?

- CPU, GPU or TPU based?
- Size of RAM and disk storage?

Hardware/Software

What hardware/software do you have? What hardware/software could you have?

- CPU, GPU or TPU based?
- Size of RAM and disk storage?
- Hardware compatible with ML software? Software restrictions from company/collaborations?

Hardware/Software

What hardware/software do you have? What hardware/software could you have?

- CPU, GPU or TPU based?
- Size of RAM and disk storage?
- Hardware compatible with ML software? Software restrictions from company/collaborations?
- Short-term or long-term project?
  - Rent or own? Little compute over long time or lots of compute over short term?
  - Possible start: First design/implement/experiment on owned hardware, then perform final tuning on rented hardware if needed

ML METHODS

Methods

What ML method(s) should you use?

Methods

What ML method(s) should you use?

- Depends on goal, data and hardware

Methods

What ML method(s) should you use?

- Depends on goal, data and hardware
- You will need a theoretical understanding of the methods to judge which ones to consider
  - Literature research
  - Later semesters of AI study

Methods

What ML method(s) should you use?

- Depends on goal, data and hardware
- You will need a theoretical understanding of the methods to judge which ones to consider
  - Literature research
  - Later semesters of AI study
- Start with baselines/less complex methods and models
  - Statistics, logistic regression, SVM, Random Forest, ...
  - Check Supervised Learning before Reinforcement Learning and Unsupervised Learning

EVALUATION

Evaluation

How to evaluate the methods/models?

Evaluation

How to evaluate the methods/models?

- Which score/performance measure?

Evaluation

How to evaluate the methods/models?

- Which score/performance measure?
- Do you need to correct for biases?

Evaluation

How to evaluate the methods/models?

- Which score/performance measure?
- Do you need to correct for biases?
- Which aspects of the goal are more important?

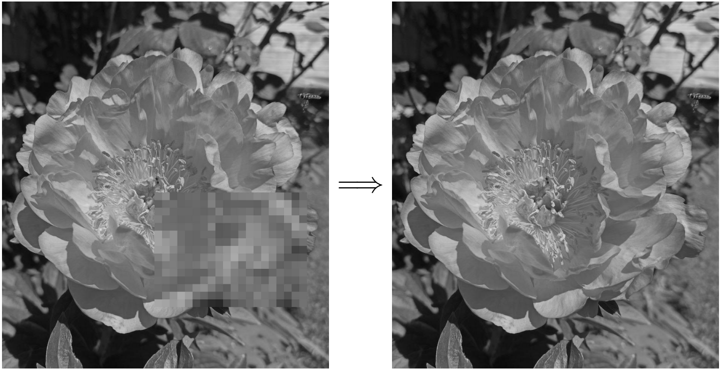
Evaluation

How to evaluate the methods/models?

- Which score/performance measure?
- Do you need to correct for biases?
- Which aspects of the goal are more important?
- What do you want to generalize to?

PYTHON II PROJECT

Python II Project: Goal (1)



Python II Project: Goal (2)

- Image Depixelation (recreation of pixelated area within an image)
- End goal: Best score on challenge server leaderboard
  - Pixelated image is fed into model, and model predicts plausible values for the pixelated area
  - Size of the pixelated area and pixelation block size should be freely selectable
  - Images should be grayscale images

Python II Project: Goal (2)

- Image Depixelation (recreation of pixelated area within an image)
- End goal: Best score on challenge server leaderboard
  - Pixelated image is fed into model, and model predicts plausible values for the pixelated area
  - Size of the pixelated area and pixelation block size should be freely selectable
  - Images should be grayscale images
- What is “plausible”?
  - Luckily, the challenge server decides for us: “Plausibility” is measured by the root-mean-squared error (RMSE) of the predicted pixels

Python II Project: Data (1)

- We will create our own dataset

Python II Project: Data (1)

- We will create our own dataset
- We will have the following data:
  - JPEG images up to 250kB
  - 100 images per student
  - Assumption: We collect roughly 30k valid images

Python II Project: Data (2)

- We will crop out parts of the original images, so we know the ground truth (no need to collect labels)

Python II Project: Data (2)

- We will crop out parts of the original images, so we know the ground truth (no need to collect labels)
- This is probably a case with sufficient data for training our methods, but
  - We can use data augmentation to increase the dataset size
  - We could use additional data from the Internet (but it will not be necessary)

Python II Project: Data (2)

- We will crop out parts of the original images, so we know the ground truth (no need to collect labels)
- This is probably a case with sufficient data for training our methods, but
  - We can use data augmentation to increase the dataset size
  - We could use additional data from the Internet (but it will not be necessary)
- We will have to
  - Clean up the raw data (exclude invalid files)
  - Perform analysis and preprocessing
  - (Possibly) perform data augmentation

Python II Project: Hardware, Software and Methods

- Hardware:
  - Notebook with CPU and 4GB of RAM should suffice
  - No need to rent/buy expensive hardware to speed up computations (you can, if you want)

Python II Project: Hardware, Software and Methods

- Hardware:
  - Notebook with CPU and 4GB of RAM should suffice
  - No need to rent/buy expensive hardware to speed up computations (you can, if you want)
- Main software:
  - Python $\geq$ 3.9
  - PyTorch

Python II Project: Hardware, Software and Methods

- Hardware:
  - Notebook with CPU and 4GB of RAM should suffice
  - No need to rent/buy expensive hardware to speed up computations (you can, if you want)
- Main software:
  - Python $\geq$ 3.9
  - PyTorch
- Methods:
  - Simple **Convolutional Neural Network (CNN)**
  - You may also use other NN types/more complex settings
  - Design and fine-tuning is up to you

Python II Project: Evaluation

- Challenge server score determines the evaluation method
- Will use the root-mean-squared error (RMSE) of the predicted pixels