Introduction to Programming

For Archaeologists

Part 6: Advanced Methods

2021-2022

Universiteit Leiden
The Netherlands

Discover the world at Leiden University

Topics of this lecture series

1. Introduction: Python, variables, comments
2. Lists & Loops
3. Loading and manipulating data
4. Graphs & Plots
5. SQL & Databases
6. **Advanced methods: Machine Learning, QGIS integration**

Assignment

Assignment deadlines

- Assignment 1: 22 April
- Assignment 2: 6 May
- Assignment 3: **20 May**

Assignment 3, due today at 23.59

Topics of this lecture

- Machine Learning
  - Train / test set
  - Accuracy metrics
  - Bias in ML
  - Features / Labels
- QGIS integration

After this session:

- You can conceptually explain what Machine Learning is
- You know what a test / train split is
- You know what features and labels are in ML
- You are aware there are different performance metrics
- You can explain how biased data can affect ML outcomes
- You can give an example of bias in archaeological data
- You are aware of the integration of Python with QGIS

Classification

Assigning labels (classes) to items

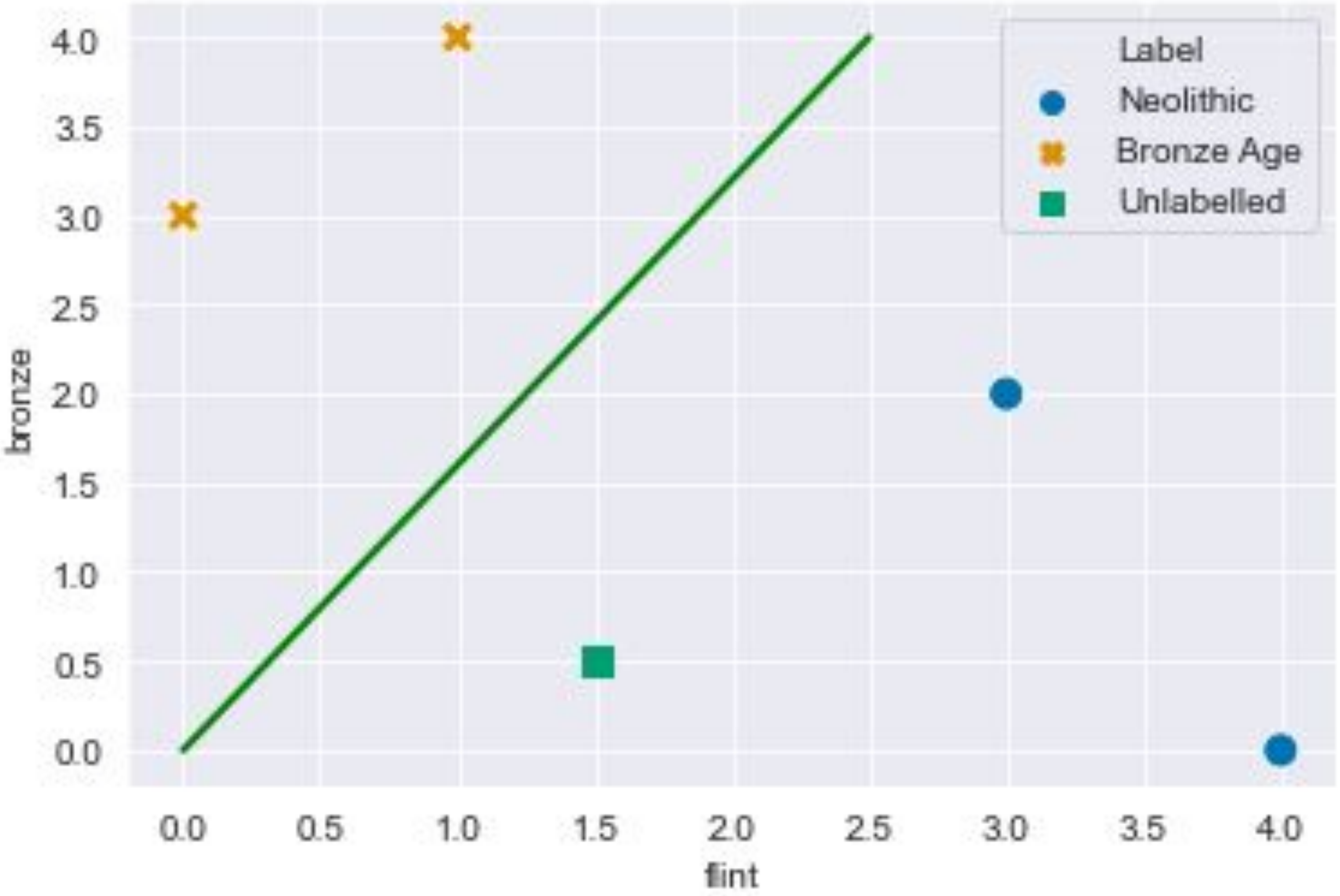
pot_ID	**height**	**width**	**…**	**label**
1	13.6	5.8	…	Pot type A
2	40.8	12.3	…	Pot type B
…	…	…	…	…
42	44.35	13.3	…	????

Rule-Based Approaches

- Opposite of machine learning
- Uses rules created by experts to assign labels
- E.g.: “if pot is higher than 30cm, assign label B”
- Depends on skill of rule maker
- Can get very complex, very fast
- No labelled data needed!

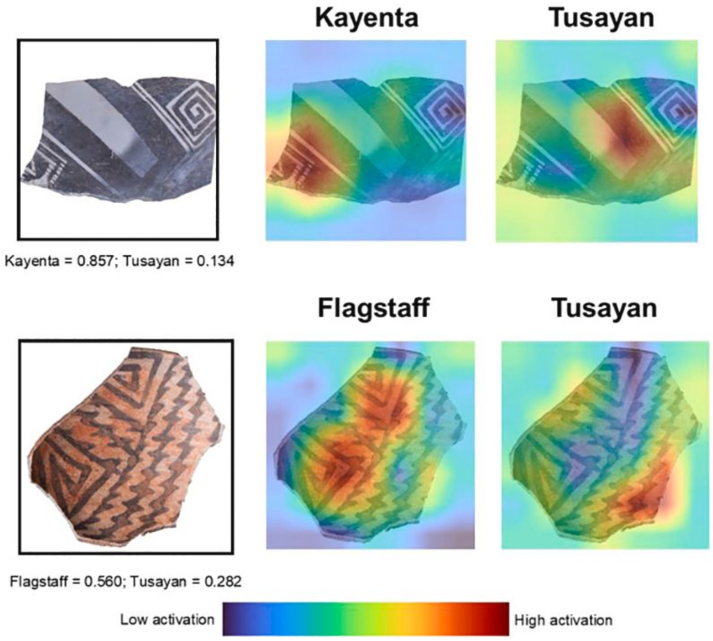
What is Machine Learning?

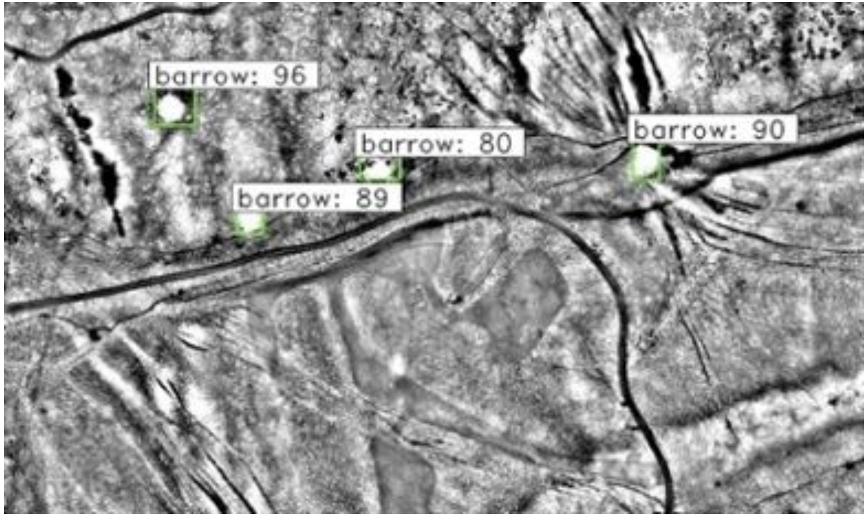
- Computer learns statistical relationships from a dataset labelled by humans
- No human intervention during learning: all based on examples
- Generally more effective (if enough data is available!)
- Labelling enough data can be time consuming…


Label
● Neolithic
✖ Bronze Age
■ Unlabelled
bronze
flint
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Deep Learning

Classifying pottery


Kayenta
Tusayan
Kayenta = 0.857; Tusayan = 0.134
Flagstaff
Tusayan
Flagstaff = 0.560; Tusayan = 0.282
Low activation
High activation



Classifying LiDAR data

Datasets and test/train

- Need labelled data to train algorithm (train set)
- Need to test trained model (test set)
- Often use 80% train, 20% test
- Evaluation on ‘unseen’ test set shows you how well the model works

Features and Labels

Features: the columns/attributes the algorithm learns from

Labels: the classes the algorithm should predict

Features

Labels

pot_ID	**height**	**width**	**…**	**label**
1	13.6	5.8	…	Pot type A
2	40.8	12.3	…	Pot type B

Train a model

- Using train set, with selected features and labels
- Select type of algorithm
  - Many exist
  - Support Vector Machines (SVM) often used
- With ‘small’ data and standard ML models this is really fast, under a second generally
- With ‘big’ data (GBs of data, images, LiDAR data) and deep learning, this can take days or even weeks!

Performance metrics

- Performance on test set expressed by certain metrics
  - Precision
  - Recall
  - F1 Score
- Often expressed as percentage (85.8%) or 0-1 (0.858)

- Example: classifying pottery, handformed or not?

Evaluation

Recall: out of all the hand formed pots, what percentage have been correctly labelled as hand formed?

$$\text{Recall} = \frac{tp}{tp+fn}$$

Precision: when a pot is marked as handformed, how often is the algorithm correct?

$$\text{Precision} = \frac{tp}{tp+fp}$$

Evaluation

F1 Score: the harmonic mean of recall and precision

Overall measure of algorithm performance

$$F^1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Bias in Machine Learning

- Algorithms are ‘objective’, but:
- ML models can only predict what they’ve been trained to predict
- Models only as reliable as the human(s) selecting / collecting / labelling the training data
- Training data should be a true representation of reality (as real as possible!)
- If not: human bias transferred to ML model -> predictions flawed
- “Garbage in = garbage out”

Bias in Machine Learning - examples

- Trial software: bias against black defendants
- Image recognition: ‘cooking’ always done by women
- Predictive policing: predicted only poor neighbourhoods
- Photo labelling: black people get label ‘gorilla’

Bias in Machine Learning - in archaeology

- Confirmation bias: only data from places we know have archaeology
- Sampling bias: e.g. only clear examples selected
- Preservation bias: computer might think only flint was used in stone age, no organic materials
- Depositional bias: e.g. bronze artefacts deposited in rivers
- Personal bias: people into flint will often find more flint than pottery in surveys
- Institutional bias: artefact image recognition learned to classify by looking at different scale bars in photos

Machine Learning in Python

The Four Horsemen of basic machine learning in Python

Python & QGIS integration

- Allows you to script geographical computations
- Particularly useful for steps you need to repeat many times

Example: site catchment

- What kind of soils, level of elevation or slope around site?
- A number of steps to be repeated for all sites -> script it!
- For each site, draw circle of 25km, get all land use polygons within circle, calculate %, assign to site point as attribute or export as CSV
- Then: other analysis in QGIS or Python

https://archaeoinformatics.net/python-for-site-catchment-qgis/

QGIS Plugins

- Similar to Python libraries
- Allows you to import code other people wrote
- You can make a plugin in Python and share it

- List of archaeology related plugins: https://plugins.qgis.org/plugins/tags/archaeology/

After this session:

- You can conceptually explain what Machine Learning is
- You know what a test / train split is
- You know what features and labels are in ML
- You are aware there are different performance metrics
- You can explain how biased data can affect ML outcomes
- You can give an example of bias in archaeological data
- You are aware of the integration of Python with QGIS

Questions?

- **Any questions about any of the subjects?**

- Contact me at
  - a.brandsen@arch.leidenuniv.nl

Slides are available on Brightspace

Follow up courses

Minors:

- AI & Society (more theoretical, with small practical)
- AI & Data Science (more practical)

In Archaeology:

- MA, Quantitative Methods (How to do stats in Python)
- MA, Data Analysis with Python (How to do advanced analysis with Python)

Online:

- https://www.learnpython.org/ (free)
- https://www.codecademy.com/catalog/language/python
- https://www.udemy.com/topic/python/

Exam

- 24th of May, 13.00, F1.01
- Paper exam
- Questions about slides, exercises, and literature

Exercises

github.com/alexbrandsen/Introduction-to-Programming-for-Archaeologists

- Go to github
- Click on ‘modules’
- Right click on the 6th module
- Select ‘save link as’ or ‘download as’
- Save the file in the ‘modules’ folder within your own Scripts folder
- Start Anaconda
- Start Jupyter Notebook
- Navigate to the notebook file and run it