Welcome to DATA1030: Hands-on data science

Instructor: Andras Zsom

HTA: Asher Labovich

TAs: Sarah Bao, Junhui Huang, John Farrell, Penggao Gu, Shiyu Liu,

Yicen Ye, Zhaocheng (Harry) Yang, Huaxing Zeng, Devraj Raghuvanshi

What is classification?

Let’s assume I move to a tropical island and I never had papayas before.

How could I figure out which papayas I’ll like?

- Step 1: Inspect the papaya and collect data (e.g., color, firmness, weight), try some papayas and write down whether I like them (1) or not (0).
- **Step 2: Train a machine learning model on the data.**
- Step 3: Deployment, act based on the model's predictions:
  - When I get a new papaya, run the model on it to get a prediction.
  - If the model predicts that I’ll like the papaya, eat. (I might still eat bad papayas but hopefully not too often)
  - Otherwise, discard. (I might discard tasty papayas hopefully not too often)

Feature matrix and target variable

- The papaya properties (column names) and the collected values for individual papayas is the feature matrix (X).
- Wheter I find a papaya tasty (1) or not (0) is the target variable (Y).
  - Our ML model predicts the target variable given the feature values.

X	color	firmness	weight (g)	Y
papaya 1	yellow	firm	275	**1**
papaya 2	brown	soft	290	**0**
...	...	...	...	**...**

X	color	firmness	weight (g)	Y
papaya i	yellow	soft	260	1
...	...	...	...	...
papaya n	green	hard	200	0

Quiz

- What other data/info could you collect on the papayas?

Notice all the decisions I made!

- I decided which features I collect and how I represent each feature.
- I decided how to represent my target variable.

These (conscious or uncoscious) decisions often make or break an ML project! - Think carefully about what data you collect and why! - Think carefully about how you represent data!

There are several ways I could solve this task which changes the properties of the feature matrix, target variable, and the ML problem!

Feature matrix: structured vs. unstructured datasets

The feautre matrix above is **structured data** because it is tabular so it can be stored in an excel/csv/SQL table.

- each data point / sample is described in an identical way
- the first value is always the color, the second value is the firmness, and the third value is the weight in grams.

Datasets are sometimes **unstructured**!

Some examples of unstructured data:

- images: image size can vary from sample to sample
- text: the length of documents vary from sample to sample
- videos
- voice recordings

Target variable: classification vs. regression

Binary classification:

- Task is expressed as a yes (1) or no (0) question.
- Is this papaya tasty? Yes or no?
- Will the customer click on my ad? Yes or no?
- Does the patient have cancer? Yes or no?
- Should this person be hired? Yes or no?
- Will it rain tomorrow? Yes or no?

Multiclass classification:

- Task is expressed as a multiple choice question. Possible answers are independent categories.
- What is the topic of this article? Politics, economy, science, fashion, pop culture?
- What animal is on this picture? Lion, tiger, elephant, giraffe, etc?
- What's the emotional state of the person who wrote this tweet? Sad, happy, confused, angry, neutral?

Ordinal regression/classification:

- Task is expressed as a multiple choice question. Possible answers are an ordered list of categories.
- On a scale of 1-5, how much do I like the papaya?
- How satisfied is the customer based on the recorded call? Not at all < somewhat < satisfied < very satisfied?
- How happy is the customer who left a review? very unhappy < somewhat unhappy < neutral < somewhat happy < very happy

Regression:

- Task is expressed such that the answer to the question is on a continuous scale. The scale could be bound on one or both ends but not required. The quantity we predict could have a physical unit. If it does, the unit needs to be clearly decribed!
- What will the temperature be tomorrow? The number depends on whether the temperature is measured in Fahrenheit, Celsius, or Kelvin. The unit must be disclosed!
- What will the bitcoin price be an hour from now? The number depends on whether price is in USD, EUR, HUF, etc. The unit must be disclosed!
- How much will it rain tomorrow? In what unit is this expressed?
- What will be the sale price of this house? The number depends on whether price is in USD, EUR, HUF, etc. The unit must be disclosed!

Quiz

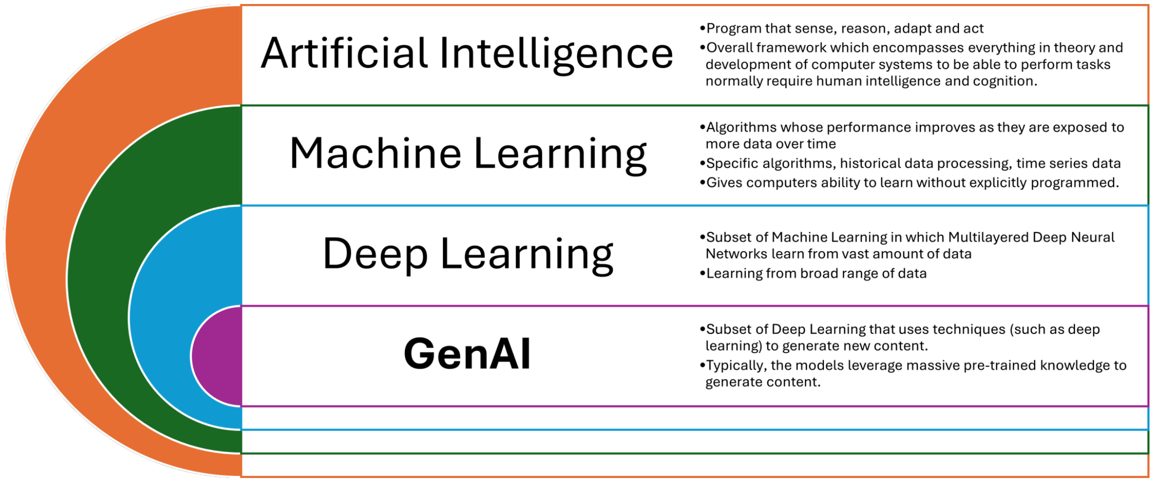
- What type of feature matrix and target variable do we work with in the problems below?

The AI landscape

All examples described so far were supersived ML problems!

- we collect some data (feature matrix and one target variable)
- train a model to predict the target variable for previously unseen data
- we act based on the predictions.

This is but one area of AI! Let's zoom out!



Artificial Intelligence

- Program that sense, reason, adapt and act
- Overall framework which encompasses everything in theory and development of computer systems to be able to perform tasks normally require human intelligence and cognition.

Machine Learning

- Algorithms whose performance improves as they are exposed to more data over time
- Specific algorithms, historical data processing, time series data
- Gives computers ability to learn without explicitly programmed.

Deep Learning

- Subset of Machine Learning in which Multilayered Deep Neural Networks learn from vast amount of data
- Learning from broad range of data

GenAI

- Subset of Deep Learning that uses techniques (such as deep learning) to generate new content.
- Typically, the models leverage massive pre-trained knowledge to generate content.

Learning objectives

By the end of the semester, you will be able to

- explore and visualize the dataset,
- develop a ML pipeline from scratch to deployment,
- make data-driven decisions during the pipeline development,
- handle non-standard ML problems like missing data, non-iid data,
- provide explanations with your model,
- explain your findings to technical and non-technical audiences.

The supervised ML pipeline

We will follow these steps during the course!

0. Data collection/manipulation: you might have multiple data sources and/or you might have more data than you need

- you need to be able to read in datasets from various sources (like csv, excel, SQL, parquet, etc)
- you need to be able to filter the columns/rows you need for your ML model

- you need to be able to combine the datasets into one dataframe

1. Exploratory Data Analysis (EDA): you need to understand your data and verify that it doesn't contain errors

- do as much EDA as you can!

2. Split the data into different sets: most often the sets are train, validation, and test (or holdout)

- practitioners often make errors in this step!
- you can split the data randomly, based on groups, based on time, or any other non-standard way if necessary to answer your ML question

3. Preprocess the data: ML models only work if X and Y are numbers! Some ML models additionally require each feature to have 0 mean and 1 standard deviation (standardized features)

- often the original features you get contain strings (for example a gender feature would contain 'male', 'female', 'non-binary', 'unknown') which needs to be transformed into numbers
- often the features are not standardized (e.g., age is between 0 and 100) but it needs to be standardized

4. Choose an evaluation metric: depends on the priorities of the stakeholders

- often requires quite a bit of thinking and ethical considerations

5. Choose one or more ML techniques: it is highly recommended that you try multiple models

- start with simple models like linear or logistic regression
- try also more complex models like nearest neighbors, support vector machines, random forest, etc.

6. Tune the hyperparameters of your ML models (aka cross-validation or hyperparameter tuning)

- ML techniques have hyperparameters that you need to optimize to achieve best performance
- for each ML model, decide which parameters to tune and what values to try
- loop through each parameter combination
  - train one model for each parameter combination
  - evaluate how well the model performs on the validation set
- take the parameter combo that gives the best validation score
- evaluate that model on the test set to report how well the model is expected to perform on previously unseen data

7. Interpret your model: black boxes are often not useful

- check if your model uses features that make sense (excellent tool for debugging)
- often model predictions are not enough, you need to be able to explain how the model arrived to a particular prediction (e.g., in health care)

A few notes on course policies

- Please read and make sure you understan the syllabus on canvas!
- **Course structure and grading**
    - 45% weekly problem sets
    - 40% final project
    - 5% inclass quizzes graded for completion, not correctness!
    - 10% final exam
- **Course policies**
    - if you submit after the deadline, it is a late submission
    - late submission is possible for no later than 3 days after the deadline
    - you have 6 days (144 hours) of penalty-free late submissions
    - documentation (doctor's note, Dean's note, SAS letter) is required for all accommodations!
- **GenAI policy**
    - responsible use of GenAI:
        - you start to solve a coding problem but encounter an error message or a bug you don't know how to fix, you can GenAI for help
        - you answer an essay question and ask GenAI to fix the grammar or improve style or clarity
    - You can use GenAI more broadly if you
        - cite the tool used
        - include an explanation on how you used the tool (i.e., link to the chat)
        - document your own contributions vs. the tool's contribution
    - You are graded based on your own contributions!

Mud card

In [ ]: