CHL7001H S1 Applied Deep Learning

Lecture 1: Introduction II

What even is ML

A model is really just a function, *f(X) = y*

Supervised Machine learning is when you have some examples of *X* and *y* pairs, and you don't know the function *f* (and there probably isn't a perfect one), a training algorithm will modify *f(X)* in some way to approximate *y.*

e.g. I want a model to predict house prices

Features: lot size, neighbourhood, # bedrooms, age

Target: house price

e.g. I want a model to predict a house

Features: lot size, neighbourhood, # bedrooms, age

Target: house price

What if you also have a photo of the house? (An image can also be X or part of X)

What even is ML

Let $\hat{y} = f(X)$

Error or **loss** is some function of y and $\hat{y}$, such as $(y - \hat{y})^2$, where y is the real target, to capture how far away the predictions are from the actual labels.

Training a model is just finding f to minimize the error.

Vocabulary

- X/features/inputs/independent variables
- Y/targets/labels/dependent variables
- model/function/blackbox
- error/loss/cost

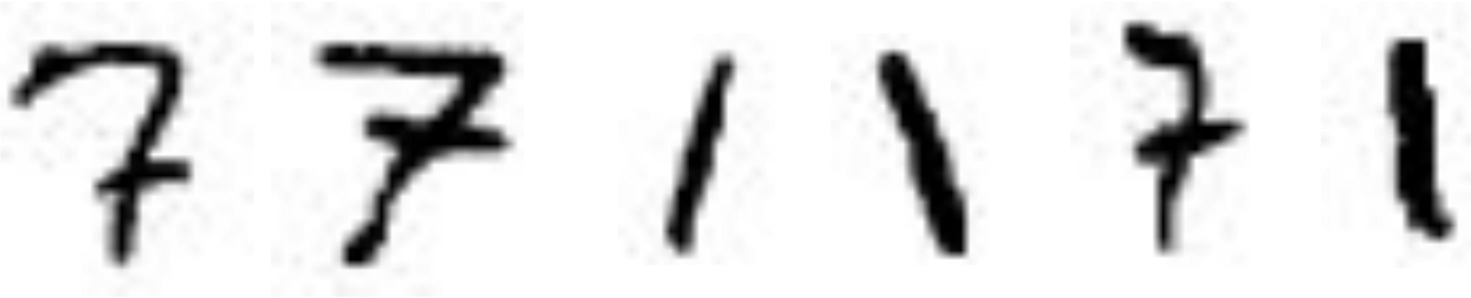
Testing a model

One of the most important concepts in supervised machine learning is **validation (test).**

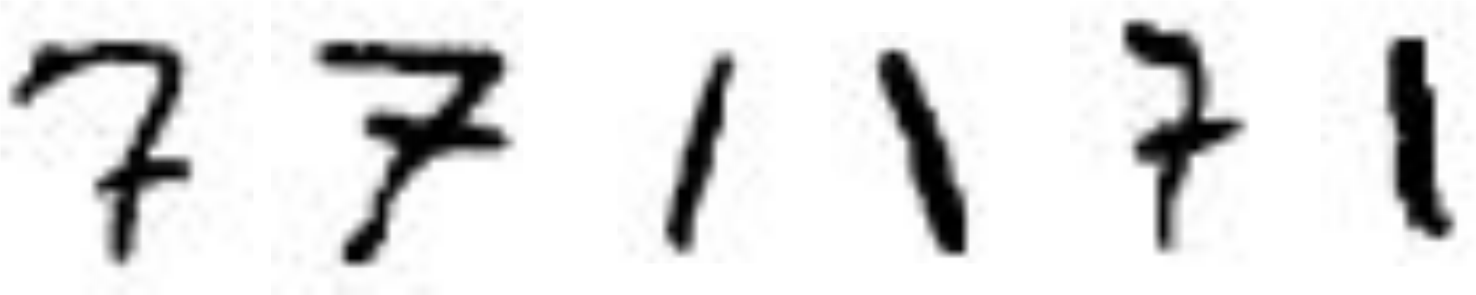
If I show my model some data, and then test out my model on that same data, and it gets 100%, we did great! ….or did we?

Analogy: you can memorize thousands of homework exercise answers, and still do poorly on an exam if you haven't learned how to solve an unseen problem.

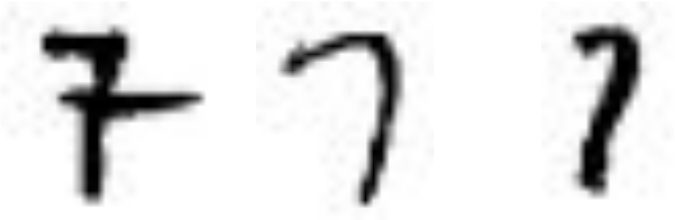
We’ve trained a handwritten digits model on these examples.



We’ve trained a handwritten digits model on these examples.



But how will the model perform on these examples?



✓ ? ?

The previous example is an instance of **overfitting**, more on this in the following lecture.

Fun reading parable:

https://blog.evjang.com/2018/02/teacup-story.html

This is a simple example. In the real world there are much more subtle versions of these failure modes.

Success in production machine learning requires non-stop sanity checking of all kinds.

If you’re fuzzy on the high level concepts of ML after this lecture, it’s a good idea to search around on the internet for a few introductions. It’s helpful to be exposed to different definitions and perspectives.

Machine Learning Workflow

Before Training

Training

1. Identify potential ML use cases.

Identify potential ML/DL use cases

Questions to ask

1. Do I need ML/DL to solve that? Or is there any simpler heuristic?
2. Is it possible to be tackled by ML/DL?
3. How well do you expect ML/DL to solve it?

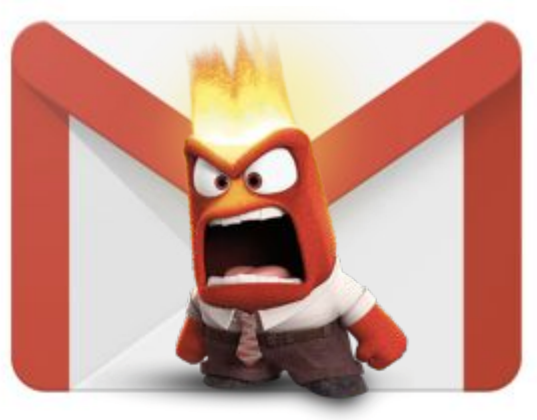
Not ML

Automatically send all your emails with "help" in the subject line to Steve?

Why don't we need ML?
This situation has a simple "if-then" rule which can be clearly defined and is easy to check if it's correct

ML

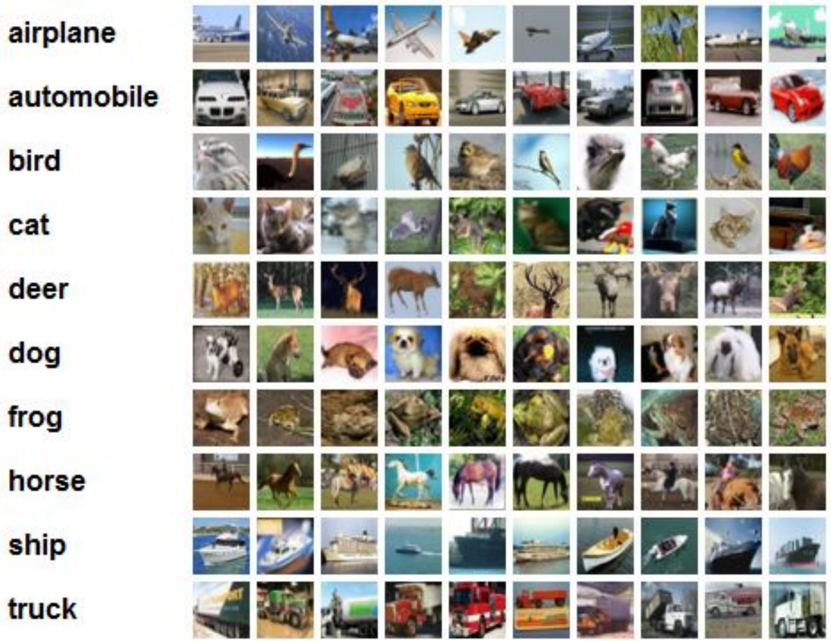
Automatically send all your angry emails to Steve?

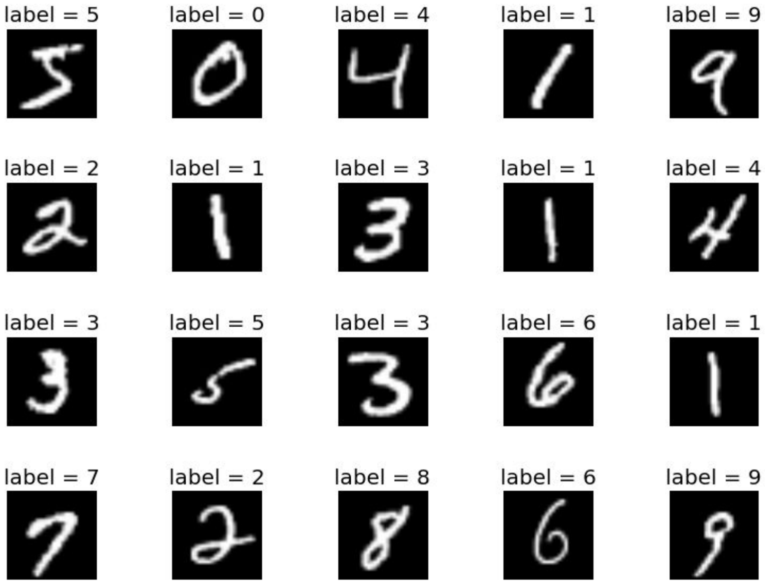


Why is ML needed here?
Language is complex, and figuring out if someone is angry is too complicated for hand written rules

What tends to be a good DL problem?

Complex problem that we don’t fully understand. E.g. image recognition.





Good target

A good target captures the ***true information*** that would be **useful** for the application.

For example, suppose you are forecasting volume at a McDonalds restaurant given historical data, so that they could make more intelligent staffing choices.

Bad target:
Good target:

Good target

A good target captures the ***true information*** that would be **useful** for the application.

For example, suppose you are forecasting volume at a McDonalds restaurant given historical data, so that they could make more intelligent staffing choices.

Bad target: predicting the number of staff
Good target: predicting the number of orders, or the total number of items ordered, or something of that nature.

Machine Learning Workflow

Before Training

1. Identify potential ML use cases.
2. Design the project based on production requirements.

Training

Project Design

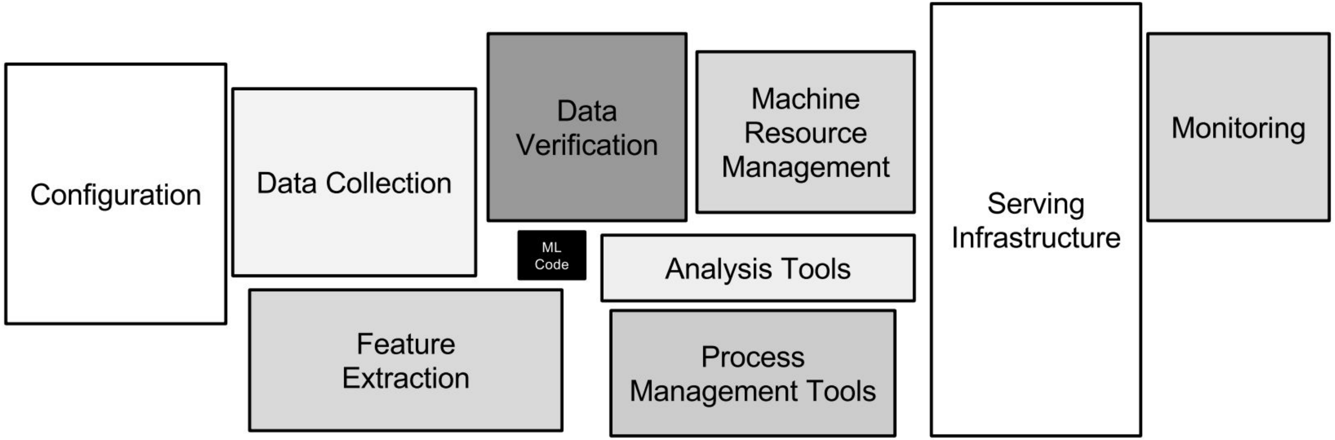


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

SOURCE: Hidden Technical Debt in Machine Learning Systems

Production considerations

- How does the production data and environment looks like?
- How do you measure if you’re solving the problem? (example McDs)
- Are there constraints such as inference speed?

In project planning, it’s important to make sure the output of a model is actually actionable!

Machine Learning Workflow

Before Training

1. Identify potential ML use cases.
2. Design the pipeline based on production requirements.
3. Prepare and organize data.

Training

Data Feasibility Assessment

Useful data

- Say you have population data: people, ages, family configurations, incomes, etc.
    - Can you build a model that predicts tomorrow's weather?
- Common sense evaluations on whether something is actually modelable given what you have.

Example: Predict crop yield

Non-exhaustive list of data problems

Data Quantity

1. # of training samples
2. Coverage
3. Granularity of data
4. Frequency
5. Availability of orthogonal data

Data Quality

1. Input documentation
2. Usefulness
3. Missing data
4. Noisy
5. Stability
6. Ease of preprocessing
7. Representation of training/production environment
8. Bias

Example: Predict crop yield

What if you only have data on 10 plants?

Example: Predict crop yield

What if you only have data from only one season?

Data deficiencies (non-exhaustive)

Data Quantity

1. # of training samples
2. Coverage
3. Granularity of data
4. Frequency
5. Availability of orthogonal data

Data Quality

1. Input documentation
2. Usefulness
3. Missing data
4. Noisy
5. Stability
6. Ease of preprocessing
7. Representation of training/production environment
8. Bias

Bias

Human biases in data:

- Reporting, selection bias
  - E.g. farmers not as careful or skips recording data for unsuccessful crops.

Human biases in data collection and annotation:

- Existing biases in the way the data is generated will be replicated by machine learning models
  - E.g. Amazon had to scrap a resume screening model that replicated human screener's bias against women.

Machine learning development is a staged but highly iterative process

Before Training

Training

1. Identify potential ML use cases.
2. Design the pipeline based on production requirements.
3. Prepare and organize data.
4. Preprocessing, cleaning, visualizing, splitting dataset into train and test.
5. Run a preliminary model to test feasibility and establish a baseline.
6. Spikes: try different models, loss, regularization and etc.
7. Optimization with experiments.
8. Analyze performance and mistakes, and **iterate**.

Notice

For Thursday’s workshop, please bring a laptop!

(If you’re using Windows, you might want to install Ubuntu in a virtual machine in advance!!)