COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Machine Learning Workflow

Dr SHI Lei

Durham University

learning lab

Learning Objectives

- Understand what is Machine Learning Workflow
- Familiar with some Machine Learning terminologies

What is Machine Learning Workflow?

learn from data		perform tasks
train a model	to	infer / predict

train a model : use the data to create a model and fine-tune the model to make better predictions.

we can then use this predictive model to make predictions on previously unseen data in order to perform specific tasks.

How does machine learning work in real life?

EXAMPLE. to build a tea detection application

1. Problem Framing

Green tea / Oolong tea

model

1. Problem Framing

Green tea / Oolong tea

1. Problem Framing

Green tea / Oolong tea

Measure of success
(prediction accuracy)

Green tea / Oolong tea

2. Data Preparation

features

Caffeine (mg)

Acidity (PH level)

→ A model to split two types of tea (green tea / oolong tea)

Data Preparation

Construct dataset	Transform data

Model is only as good as our data, i.e. how good the model can be is directly determined by how good the data is.

Green tea / Oolong tea

2. Data Preparation

training data

feature feature label

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	Green tea
65	6.2	Oolong tea
40	7.9	Green tea
60	6.5	Green tea
70	7.8	Oolong tea

instance

Green tea / Oolong tea

2. Data Preparation

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	Green tea
65	6.2	Oolong tea
40	7.9	Green tea
60	6.5	Green tea
70	7.8	Oolong tea

Green tea / Oolong tea

2. Data Preparation

Issues from dataset

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	Green tea
0.065 too small	6.2	Oolong tea
40	7.9	missing
60	6.5	Green tea
70	missing	Oolong tea

Green tea / Oolong tea

2. Data Preparation

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?	
50	5.8	Green tea	label
65	6.2	Oolong tea	label
40	7.9	Green tea	label
60	6.5	Green tea	label
70	7.8	Oolong tea	label

Green tea / Oolong tea

2. Data Preparation

transform label: (oolong -> 0; green -> 1)

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	1
65	6.2	0
40	7.9	1
60	6.5	1
70	7.8	0

Green tea / Oolong tea

2. Data Preparation

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	1
65	6.2	0
40	7.9	1
60	6.5	1
70	7.8	0

Green tea / Oolong tea

2. Data Preparation

training set

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
50	5.8	1
65	6.2	0
40	7.9	1

test set

Caffeine (mg)	Acidity (PH level)	Green tea or oolong tea?
60	6.5	1
70	7.8	0

Green tea / Oolong tea

3. Model Selection

image

video

text

voice

Green tea / Oolong tea

3. Model Selection

Linear model

Green tea / Oolong tea

4. Model Training

Data preparation and model selection are more important.

A machine learning engineer normally spent much more time in data preparation than in model training.

How good our model can be is directly determined by how good our data is – "garbage in garbage out".

Green tea / Oolong tea

4. Model Training

The training data is used to incrementally improve the model's performance, i.e. the ability to tell whether a given cup of tea is oolong tea or green tea.

Green tea / Oolong tea

4. Model Training

determine the position of the line

parameter parameter

output slope input y-intercept

$$y = m \bullet x + b$$

dependent variable independent variable

Green tea / Oolong tea

4. Model Training

- Training aims to find a <u>specific pair of m and b values</u> so that the line can be positioned well to be able to divide tea instances into 2 groups – oolong tea and green tea.
- The training process goes <u>iteratively</u> – in each iteration, m and b values are updated; the performance is calculated and compared with the performance in the last iteration in order to determine how to update m and b values for the next iteration.
- This process <u>stops</u> when new m and b values do not improve the model's performance, or changes is very little.

Green tea / Oolong tea

4. Model Training

First iteration: a random pair of m and b values is used to attempt to predict the output with those values.

It may perform very poorly ☹

It will get better over iterations 😅

Green tea / Oolong tea

4. Model Training

Green tea / Oolong tea

5. Model Testing

Once the training process has stopped, the next step is to evaluate the model, using the <u>test set</u>.

training set

test set

Green tea / Oolong tea

5. Model Testing

Aims to see if the trained model is any good when performing a specified task on previously unseen data, i.e. data from the test set.

"accuracy"

In the tea example, the aim is to test "how well" the trained model can separate oolong tea and green tea.

Green tea / Oolong tea

6. Hyperparameter Tuning

Aims to see if there is any way to improve the training process thus further improving the model.

$$y = \boxed{m} \bullet x + \boxed{b}$$

parameter (of the model) VS **hyper**parameter (of the training process)

A hyperparameter could be Learning Rate.

Green tea / Oolong tea

6. Hyperparameter Tuning

Learning Rate: how big the change to make in parameter values.

↓

- how accurate a model can become
- how long the training process takes

- Hyperparameter tuning is an experimental procedure
- Sometimes a bit more of an art than science or engineering

Green tea / Oolong tea

7. Inference / prediction

A cup of mystery tea → model → Oolong tea

What is Machine Learning Workflow?

Machine Learning Workflow

1. Problem Framing
2. Data Preparation
3. Model Selection
4. Model Training
5. Model Testing
6. Hyperparameter Tuning
7. Inference / Prediction

Some key Machine Learning Terms

label	transforming label	training / learning
feature	instance	training set
parameter	hyperparameter	test set

✓ Takeaway Points

- There are 7 steps in the machine learning workflow.
- Data preparation is extremely important and takes long time.
- Training/learning/modelling is to find parameters for the model.
- The training process goes iteratively; and stops when updating parameter does not improve the model's performance; hyperparameters need to be tuned during the process.