

ML Week 0x01 Introduction

Some preliminaries

- English
- This is an introduction. There's so much more.
- None of this gives quick fixes. Some is quicker than others.
- The goal is to play with data.
- It's important to see examples. So we'll spend time in code for that reason, too.
- We have to learn scikit-learn, even if it's not what you use later.

What is ML?

1. Some algorithms we know how to write
 - (a) Sort numbers
 - (b) Fly a plane
2. Some algorithms we don't know how to write (example: drive a car)
 - (a) Drive a car
 - (b) Read addresses on envelopes
 - (c) Detect spam
3. Maybe we can write programs to write programs when we can't
4. Some things we used to say
 - Artificial intelligence
 - Expert systems

Types of ML

1.

 - (a) Training data: input and correct responses
 - (b) Regression (continuous) (example: home prices)

(c) Classification (discrete) (example: medical outcome (alive/dead))

2. Unsupervised

(a) Clustering

(b) Deep neural networks

(c) Associative (example: human experience, e.g. from a career)

(d) Dimensionality reduction

3. Reinforcement

(a) Make a choice, get feedback

(b) Online

(c) Can be stochastic (example: predicting weather from local clues)

Curse of Dimensionality

1. *Fléau (ou : malédiction) de la dimension*

2. Volume of unit cube $\pm \epsilon$

3. Distance from $(0, 0, \dots, 0)$ to $(1, 1, \dots, 1)$

4. Physics: $1/r^{d-1}$

5. It's easy to get lost...

6. Richard Ernest Bellman, Dynamic programming, Princeton University Press, 1957.

Probability

1. Event

2. Complement of an event

3. Disjoint (mutually exclusive)

4. Independent events — knowing one outcome gives no information about other

5. addition ($\times 2$), multiplication ($\times 2$)

6. Conditional probability

7. Marginal probability

8. Joint probability

Statistics

1. Goal for a bit: think like a statistician
2. What is statistics? ($\times 3$)
3. Said differently: goal is to compare reality to a model
4. Or to find a model and then compare.
5. Good statistical models are often relatively simple.
6. What is data science ($\times 5$)

Study design

1. Anecdote
2. Study types ($\times 2$)
3. Observational studies can't conclude causality
4. Observational studies can be
 - prospective: identify individuals, collect information
 - retrospective
 - we can combine them
5. Experimental studies
 - We do stuff
 - Can conclude causation if properly designed
 - controlling: hold other variables constant (e.g., drink pill with full glass of water even if we don't care)
 - randomization: cancel out effects we can't control
 - replication: enough participants
6. Study types example
 - Sunscreen use correlated to skin cancer rates.
 - Confounding variable

7. Random sampling hazards

- Not actually random
- Convenience sample
- Non-response bias

Statistical concepts

1. Variable types

- Input: Features
- Input variables measure: Explanatory variable
- Output: Response variable
- Training set
- Test set (tune parameters) (compare model parameters)
- Validation set (tune hyperparameters) (measure performance of model)
- Cross validation
- Bias - same errors regardless of input (inflexible)
- Variance - different errors with same input (too flexible)

2. Population statistics ($\times 6$)

- sample mean vs population mean ($\times 7$)
- Sample standard deviation and variance: divide by $n - 1$

3. Distributions

- Important: pdf (pmf), cdf, ppf ($\times 5$)
 - pdf = densité de probabilité
 - pmf = fonction de masse
 - cdf = fonction de répartition
 - ppf = ?
- The rest: just so you've heard of them
- Boxplot ($\times 2$)

4. Normal distributions ($\times 2$) — TODO: compare in matplotlib, quantile-quantile plot

- Sample mean vs population mean
- How close are they?

- Point estimate: if you have to guess, this is it
- Correction: if I want to be on average weighted right as much possible

5. Sampling distributions ($\times 3$)

- Sampling mean is unimodal and approximately symmetric
- It is centred at population mean.
- The standard deviation of the sample mean tells us how far a point sample's mean is likely to be from the population mean. In other words, how much error we are likely to have in the point estimate's mean. **Standard error.**
- TODO: Generate uniform population, sample, and plot sampling distribution
- TODO Generate highly skewed population, sample, and plot sampling distribution
- In real life, we don't have access to the population parameters. We have to *estimate* them from samples. So we can't *know* the standard error (erreur type).

6. Confidence intervals ($\times 3$)

- Sampling is usually expensive.
- Reminder: Independent random samples!
- Correct language: “We are 95% confident that the population parameter is between. . .”
- Incorrect language: describe the confidence interval as capturing the population parameter with a certain probability.
- This is one of the most common errors: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.
- Another especially important consideration of confidence intervals is that they only try to capture the population parameter. Our intervals say *nothing* about the confidence of
 - capturing individual observations
 - a proportion of the observations
 - about capturing point estimates

Confidence intervals only attempt to capture population parameters.