Machine Learning 0xJ1-A Introduction

Structure of the Course

- Part one
  - Basic principles
  - How to ask a question
  - How to evaluate a response to the question
  - How to analyse data
  - Regression vs Classification
  - SVM
  - Neural networks
  - . . . all with codelabs
- Part two
  - Reinforcement of part 1
  - More codelabs with more “realistic” problems. (They’ll still be small.)
  - Special topics
- Slides and notes in English (because ML is)
- Lectures in French (because we’re in France)
- Short quiz at beginning of each half day
- Short evening projects
- Friday oral presentation (5 minutes)

Why ML?

- Playing with blocks vs doing maths
- Sometimes we know how to solve problems (e.g., sorting)
- Sometimes we don’t (e.g., recognise a cat, read handwriting on envelopes)
- Not magic

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 terms we used to use for ML
   - Artificial intelligence
   - Expert systems

Disclaimers

- The literature is overwhelmingly in English
- Time is short
- You should plan to spend three hours working on your own per hour in class (at least, if this were a more classically structured course)

Types of ML

1. $\boxed{\text{Supervised}}$
   (a) Training data: input and correct responses
   (b) Regression (continuous) (example: home prices)
   (c) Classification (discrete) (example: medical outcome (alive/dead))
2. $\boxed{\text{Unsupervised}}$
   (a) Clustering
   (b) Deep neural networks
   (c) Associative (example: human experience, e.g. from a career)

(d) Dimensionality reduction

3. $\boxed{\text{Reinforcement}}$

   (a) Make a choice, get feedback

   (b) Online

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

Talk about course structure

- In class: mostly theory, some code, some maths
- Group work, TD (also in class, but also outside)
- Between classes: coding assignments (python)
- Communication: email, github (ideally use similar names)
- Help each other via email, github issues, etc.
- Participative evaluation
- Don't copy. Learn.
- Final project (oral)

$\boxed{\text{Curse of Dimensionality}}$

1. *Fléau (ou : malédiction) de la dimension*
2. Volume of unit cube $\pm\epsilon$
3. Distance from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$
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. Marginal probability
6. Joint probability

Addition rule: independent events

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Addition rule: dependent events

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Multiplication rule: independent events

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

Multiplication rule: dependent events

$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

Conditional probability

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$\cup_i A_i = A \quad \wedge \quad A_i \cap A_j = \varnothing \implies$$

$$P(A_1 \mid B) = \frac{\Pr(B \mid A_1) \Pr(A_1)}{\sum_i \Pr(B \mid A_1) \Pr(A_1) + \cdots + \Pr(B \mid A_k) \Pr(A_k)}$$

Statistics

1. Goal for a bit: think like a statistician
2. Said differently: goal is to compare reality to a model
3. Or to find a model and then compare.
4. Good statistical models are often relatively simple.

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Sadly, sometimes people forget 1.

Statistics is about making 2–4 efficient, rigorous, and meaningful.

What is data science?

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

- What does the public perceive?
- What takes the most time?
- What is most often forgotten?

Is this the same as what statistics is?

Study design

1. Anecdote

Some properties of anecdote:

- is data
- haphazardly collected
- is generally not representative
- sometimes result of selective retention
- does not accumulate to be representative
- might be true (by chance)
- is ok to use as hypothesis, but be clear that hypothesis is anecdote

2. Study types

    - Observational
    - Experimental

    What can go wrong?

    - Forgetting that association $\neq$ causation
    - Not random
    - Confounding variables

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

Variable types

- all = numerical + categorical
- numerical = continuous + discrete
- categorical = regular + ordinal

bias vs variance

Illustrate with bullet holes on a round target.

Statistical concepts

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)

Population statistics

- **Deviation** is distance from mean

- **Variance** is mean square of deviations
- **Standard deviation** is square root of variance

$$s^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n - 1}$$

$$\sigma^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n}$$

$$\text{Var}(X) = \sigma^2 = (\overline{x} - x_1)^2 \Pr(X = x_1) + \cdots (\overline{x} - x_n)^2 \Pr(X = x_n)$$

`Mean`

- sample mean vs population
- Sample standard deviation and variance: divide by $n - 1$
- Illustrate with balance beam
- Illustrate with weights hanging off a balanced beam
- Illustrate with distribution and balanced on pivot at centre of mass

$$\mu = E(X) = \sum w_i x_i = \mathbf{w} \cdot \mathbf{x}$$

$$\mu = E(X) = \sum \Pr(X = x_i) x_i$$

$$\mu = E(X) = \int x f(x) \, \mathrm{d}x$$

boxplot-vs-pdf.png

1. Distributions
   - Important: pdf (pmf), cdf
     - pdf = densité de probabilité
     - pmf = fonction de masse
     - cdf = fonction de répartition
   - There are many others, we won't use them here, but they are often useful.

2. Normal distributions
    - 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
3. Sampling distributions
    - 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).
4. Confidence intervals
    - 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.

Sample n points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

$\boxed{\text{boxplot illustrations (.png) } \times 2}$

Linear Algebra

B is a basis for V iff any of these conditions are met:

- B is a minimal generating set of V
- B is a maximal set of linearly independent vectors
- Every vector $v \in V$ can be expressed in a unique way as a sum of $b_i \in B$

The conditions are equivalent.

Eigenvectors, eigenvalues

$$Av = \lambda v$$

$$Av = \lambda 1 v \iff (A - \lambda 1)v = 0$$

$\boxed{\text{Eigenvector video}}$