Application of Artificial Intelligence

Opportunities and limitations through life & Earth sciences examples

Clovis Galiez

Grenoble
Statistiques pour les sciences du Vivant et de l'Homme

April 10, 2019

Goal

- Discover and practice machine learning (ML) techniques
  - Linear regression
  - Logistic regression
  - Neural networks
- Experiment some limitations
  - Curse of dimensionality
  - Hidden overfitting
  - Sampling bias
- Towards autonomy with ML techniques
  - Design experiments
  - Organize the data
  - Evaluate performances

Today's outline

- Short summary of the last lecture
- Logistic regression exercise correction
- Cross-validation
- Application to IBD prediction

Last lecture

Remember

What do you remember from last lecture?

Last lecture

Remember

What do you remember from last lecture?

- Curse of dimensionality

Last lecture

Remember

What do you remember from last lecture?

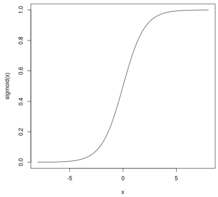
- Curse of dimensionality
  - Experimental evidence
  - Regularization helps to get the right parameters
- Logistic regression

Logistic regression

The best predictor is: $f(\vec{x}) = p(Z = 1|\vec{x})$. Problem: $p(Z = 1|\vec{x})$ is unknown.
Many situations[1] lead to the following form:

$$\exists \vec{w} \text{ such that } p(Z = 1|x) = \sigma(\vec{w}.\vec{x} + b)$$

where the function σ is the logistic sigmoid $\sigma : x \mapsto \frac{1}{1+e^{-x}}$



[1] For instance $\vec{x}|Z = i \sim \mathcal{N}(\vec{\mu_i}, \Sigma)$, or x_i's being discrete.

Conditional likelihood

Exercise

1. Show that it is not possible to find the parameters $\vec{w}$ by maximum likelihood if we don't know the distribution of $\vec{x}$.

2. Let $f(\vec{x}) = p(Z = 1|\vec{x}) = \sigma(\vec{w}.\vec{x} + b)$. Show that the *conditional* log-likelihood $LL = \log P(z_1, ..., z_N|\vec{x}_1, ..., \vec{x}_N, \vec{w}, b)$ writes:

$$LL(\vec{w}, b) = \sum_{i=1}^{N} [z_i . \log f(\vec{x}_i) + (1 - z_i). \log(1 - f(\vec{x}_i))]$$

3. To what well-known loss the optimization of this conditional likelihood corresponds?

4. Interpret geometrically the role of parameters $\vec{w}$ and b.

Choice of the regularization parameter

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2 + \lambda||\vec{\beta}||_1$$

Exercise

1. What happens if λ is small?
2. What happens if λ is huge?

Choice of the regularization parameter

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2 + \lambda ||\vec{\beta}||_1$$

Exercise

1. What happens if λ is small?
2. What happens if λ is huge?

How to choose the right value of the regularization parameter λ?

Cross-validation

λ should be chose to **generalize** as best as possible!

Cross-validation

λ should be chose to **generalize** as best as possible!

X_1	X_2	...	X_N	Y
-0.74	0.57	...	-0.82	0
0.26	0.07	...	0.49	1
-0.53	-0.07	...	0.71	1
0.69	0.27	...	0.45	1
-0.79	0.07	...	0.9	0
-0.18	-0.97	...	-0.25	0
-0.56	-0.21	...	0.24	1
-0.66	0.16	...	-0.96	1
-0.02	-0.18	...	-0.95	0
-0.44	0.46	...	-0.25	1

$\rightarrow$ Val. loss $= 0.5$

Training set

Validation set

Cross-validation

λ should be chose to **generalize** as best as possible!

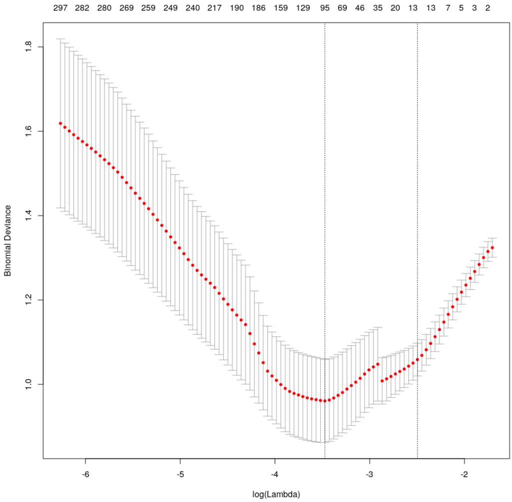
X_1	X_2	...	X_N	Y
-0.74	0.57	...	-0.82	0
0.26	0.07	...	0.49	1
-0.53	-0.07	...	0.71	1
0.69	0.27	...	0.45	1
-0.79	0.07	...	0.9	0
-0.18	-0.97	...	-0.25	0
-0.56	-0.21	...	0.24	1
-0.66	0.16	...	-0.96	1
-0.02	-0.18	...	-0.95	0
-0.44	0.46	...	-0.25	1

$\rightarrow$ Val. loss $= 0.8$

Training set

Validation set

Cross-validation experimental results



[R package: `cv.glmnet`]

Classification of microbial communities.

Application to human health.

Microbiome importance in human health

The bright side:

Health status highly correlated with the diversity of the gut microbiome [Valdes et al. 2018]

The dark side:

Germany: Ten die from E.coli-infected cucumbers

28 May 2011

The death toll in Germany from an outbreak of E.coli caused by infected cucumbers has risen to at least 10.

The cucumbers, believed to have been imported from Spain, were contaminated with E.coli which left people ill with hemolytic-uremic syndrome (HUS).

Hundreds of people are said to have fallen sick.

AFP

It is unclear whether the cucumbers were infected at source or in transit

[Karch et al. EMBO Mol. Med. 2012]

Studying the microbiome: hard work!

How to study micro-organisms?

- Isolate the organism
- Grow in culture
- Observe, experiment 

Studying the microbiome: hard work!

How to study micro-organisms?

- Isolate the organism
- Grow in culture
- Observe, experiment 

Far from being always possible, often need symbiosis.
Only doable for tiny fraction of micro-organisms.

Studying the microbiome: hard work!

How to study micro-organisms?

- Isolate the organism
- Grow in culture
- Observe, experiment 

Far from being always possible, often need symbiosis.
Only doable for tiny fraction of micro-organisms.

A better way to study micro-organisms?

Accessing the DNA of the microbiome: shotgun metagenomics

Sample

→

Sequencing

→

ATGATCAGTATTACCTGACAGTAGCTTG

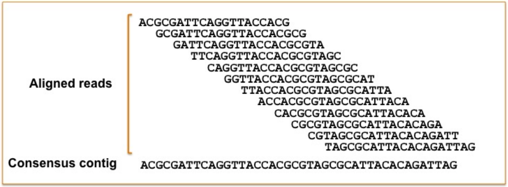
ATGATCAGTATTTACGTATACTACCTGAC

TTACTCAGTTATTACCTGACAGTAGCTT

ATGATCAGTATTACCTGACAGTATACAT

Fragmented sequences (reads $\sim 10^9 \times$ 250bp)

Assembly: from reads to **contigs**:

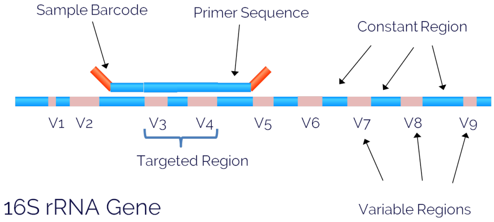


(Algorithmic and machine learning challenges here!)

Barcodes to identify species

Some parts of the genome of micro-organisms are specific to each species and allows to identify them.

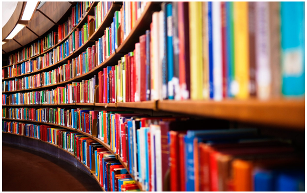
For example the 16S region in bacteria:



The big picture

sample

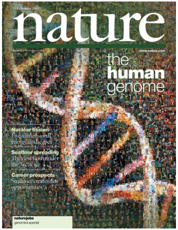
$\xrightarrow[\text{information}]{\text{DNA}}$



catalog of species

Metagenomics insights on the human gut microbiome

2000's
Human genome

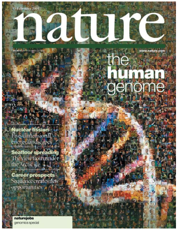


2010's
Gut metagenomes

$\approx$ 20k protein-coding genes

Metagenomics insights on the human gut microbiome

2000's
Human genome



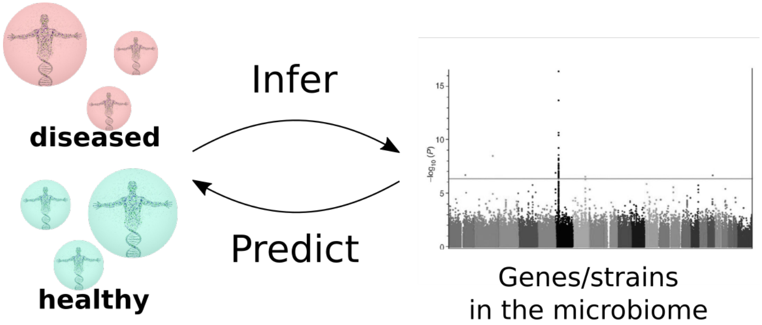
2010's
Gut metagenomes

$\approx$ 20k protein-coding genes $\xrightarrow{\times 100}$ $\approx$ 2M protein-coding genes

Human gut microbiome is rich!

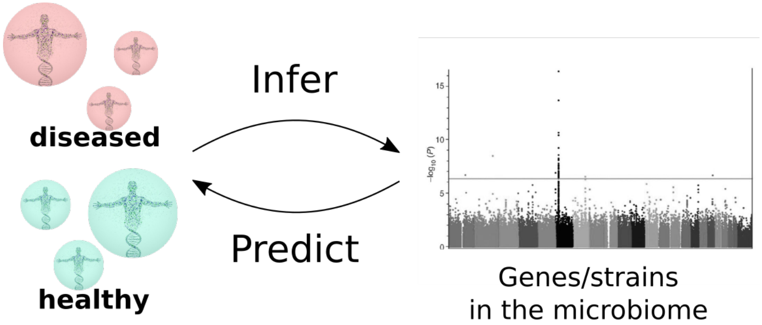
MWAS: metagenome-wide association studies

Relates the variation of the microbiome to the phenotype.



MWAS: metagenome-wide association studies

Relates the variation of the microbiome to the phenotype.



Today

You will diagnosis Inflammatory Bowel Disease through the structure of the gut microbial community.

MWAS in an ideal world

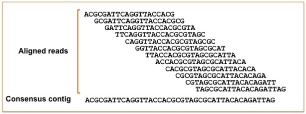
sampling

→

sequencing

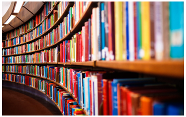
→

assembly



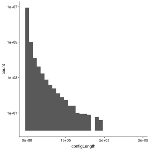
---

species catalog



→

species abundances



→

predictive model

$$\sigma(\sum w_i s_i)$$

It's a classification problem!

Predict IBD!

Fetch:

- the R script at
  `clovisg.github.io/teaching/asdia/ctd3/ibd.zip`
- the data at
  `clovisg.github.io/teaching/asdia/ctd3/ibdStart.zip`

Microbial species abundances have been computed for 396 individuals (148 with IBD, 248 healthy).

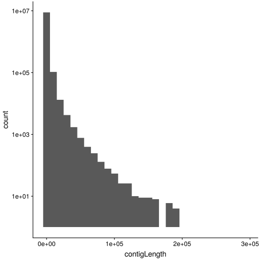
Your mission

Build a model that predicts IBD status based on the microbial composition of their gut.

See you next week to work with regressions!

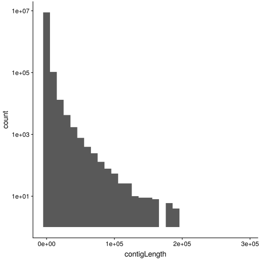
Noisy mixture: the metagenomic struggle!

Assembly process breaks with intra-population variations.



Noisy mixture: the metagenomic struggle!

Assembly process breaks with intra-population variations.



Millions of small contigs coming from thousands of species...

ATGATCAGTATTACCTGACAGTAGCTTG

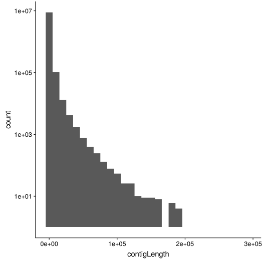
ATGATCAGTATTTACGTATACTACCTGAC

TTACTCAGTTATTACCTGACAGTAGCTT

ATGATCAGTATTACCTGACAGTATACAT

Noisy mixture: the metagenomic struggle!

Assembly process breaks with intra-population variations.



Millions of small contigs coming from thousands of species...

ATGATCAGTATTACCTGACAGTAGCTTG

ATGATCAGTATTTACGTATACTACCTGAC

TTACTCAGTTATTACCTGACAGTAGCTT

ATGATCAGTATTACCTGACAGTATACAT

$\longrightarrow$