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 3, 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

Announcement

8th of April 2019. 18:30-19:30.
Building Ensimag, Amphi D.

Conference Yves Demazeau (UGA):

"Panorama de l'IA".

Announcement

18th of April 2019. 10:00-12:00.
Building ARSH, Bat. B1

Conference Alain Létourneau
(University of Sherbrook):

AI and ethics.

Today's outline

- Short summary of the last lecture
- Experiment the curse of dimensionality (tutorial)
- Logistic regression

Last lecture

Remember

What do you remember from last lecture?

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality
  - Overfit can stem from too many features (capacity of description increases exponentially)
  - More data helps

Last lecture

Remember

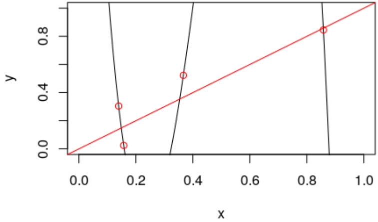
What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality
  - Overfit can stem from too many features (capacity of description increases exponentially)
  - More data helps
  - Restricting the parameter space: regularization
    - Ridge
    - Lasso

Ridge regularization example

Let's come back to the model $Y = \sum_{i=0}^{3} \beta_i x^i + \epsilon$.

The maximum likelihood with 4 points will give a $\vec{\beta}$ fitting perfectly the points:



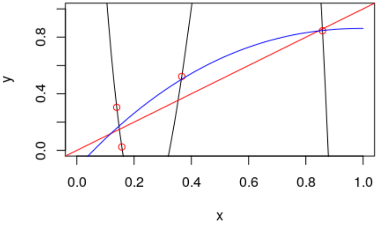
Maximum *likelihood* coefficients:

β_0	β_1	β_2	β_3
5.169	-54.388	155.755	-114.487

Ridge regularization example

Let's come back to the model $Y = \sum_{i=0}^{3} \beta_i x^i + \epsilon$.

With a prior $\mathcal{N}(0, \eta^2)$ the maximum a posteriori of the vector $\vec{\beta}$ corresponds to (blue curve):



Maximum *a posteriori* coefficients

β_0	β_1	β_2	β_3
-0.1279	2.2561	-1.5779	0.3180

Ridge regularization

Consider the linear model $Y = \sum \vec{\beta}.\vec{x_i} + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

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

2. Show that putting a Gaussian prior centered on zero on the parameters is the same as solving the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2 + \lambda||\vec{\beta}||_2^2$$

This is called **ridge regularization**. What is it enforcing?

Ridge regularization

Consider the linear model $Y = \sum \vec{\beta}.\vec{x_i} + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

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

2. Show that putting a Gaussian prior centered on zero on the parameters is the same as solving the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2 + \lambda ||\vec{\beta}||_2^2$$

This is called **ridge regularization**. What is it enforcing?
It tells the model **to avoid high values** for the parameters.

From ridge to lasso

Suppose you model a variable Y depending on some explanatory variables x with a linear model:

$$Y = \beta_0 + \sum_{i=1}^{N} \beta_i . x_i + \epsilon$$

Imagine now that you know that actually **only few** variables actually explain your target variable.

Question!

Gaussian priors on β_i centered on 0 avoid high values of β_i.
Will it push the non-explanatory variables down to 0?

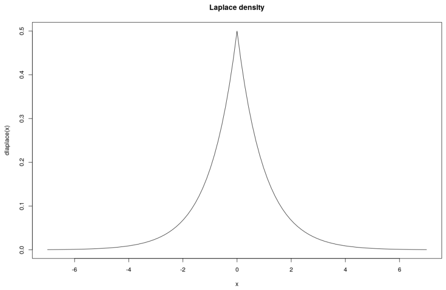
- Think individually (5')
- Vote

Lasso penalization

What should be the shape around 0 of the prior distribution if we want to use less parameters?

Lasso penalization

What should be the shape around 0 of the prior distribution if we want to use less parameters?
Something like:



$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$$

Exercise

Work out the formula to see what the model will minimize.

Show that curse of dimensionality happens!

Design a simple experiment showing the curse of dimensionality in the linear regression setting.

- Individual reflexion (5')
- Small group reflexion (5-10')
- Individual implementation in R (20')

Experimental plan

Simulate in R a dependence between a Gaussian vector $\vec{X}$ and an output variable.
Find the maximum likelihood of the parameters of a linear regression.
Add components to $\vec{X}$ that are not related to the output variable? Are the coefficients near to 0?

Regularization

Complete your experiment to show that regularization helps.

- Individual reflexion (2')
- Small group reflexion (3-5')
- Individual implementation in R (20')

Experimental plan

Use the R package `glmnet` to implement a ridge and lasso regularization.
Optimize the parameters of the regularized linear regression.
Are the non-explanatory coefficients near to 0? For which regularization?

Logistic regression (classification)

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary $(0/1)$ random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_{+}^{M} \to [0, 1]$ is a

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary $(0/1)$ random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \to [0, 1]$ is a function chosen to minimize some *loss* in order to have

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \rightarrow [0,1]$ is a function chosen to minimize some *loss* in order to have $f(\vec{x}) \approx z$ for realizations $\vec{x}, z$ of $\vec{X}, Z$.

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \rightarrow [0, 1]$ is a function chosen to minimize some *loss* in order to have $f(\vec{x}) \approx z$ for realizations $\vec{x}, z$ of $\vec{X}, Z$.

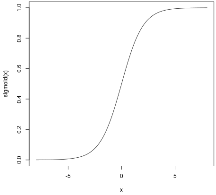
Which loss?

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:

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

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.

Curse of dimensionality in classification

From the previous exercise, if the k^{th} component of the feature vector $\vec{x}$ plays no role in the classification process, what should be the value of w_k?

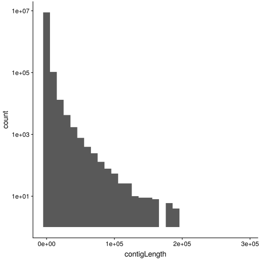
What can you expect in practice?

If you expect only few explanatory components in your vector of features $\vec{x}$, what shall you do?

See you next week to work with regressions!

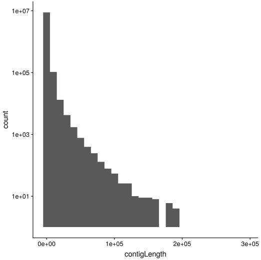
Noisy mixture: the metagenomic struggle!

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

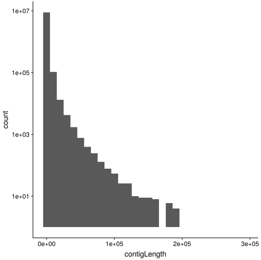
ATGATCAGTATTTACGTATACTACCTGAC

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

$\longrightarrow$